Université Rennes 2

Thèse d'Habilitation à Diriger des Recherches

Spécialité : Mathématiques appliquées

Contributions aux tests d'hypothèses non paramétriques et à l'apprentissage statistique

Contributions to nonparametric hypotheses testing and statistical learning

par

Magalie FROMONT

Soutenue à l'Université Rennes 2, le 2 décembre 2015, devant le jury composé de :

M.	Patrice	BERTAIL	Université Paris-Ouest Nanterre La Défense
Mme	Cristina	BUTUCEA	Université Paris-Est Marne-la-Vallée
Mme	Béatrice	LAURENT	INSA Toulouse
M.	Oleg	LEPSKI	Aix Marseille Université
M.	Pascal	MASSART	Université Paris-Sud
Mr.	Richard	NICKL	University of Cambridge
Mme	Anne	PHILIPPE	Université de Nantes
M.	Alexandre	TSYBAKOV	ENSAE ParisTech

Au vu des rapports de :

Patrice Bertail, Cristina Butucea et Richard Nickl.

Sous l'hypothèse que vous devriez, lectrice, lecteur, trouver votre nom dans les lignes qui suivent, la probabilité qu'il en soit rejeté n'est malheureusement pas nulle, mais je l'espère, plus faible qu'un niveau de risque acceptable. Si cet événement se réalise néanmoins, je vous présente mes plus sincères excuses.

Merci à Patrice Bertail, Cristina Butucea et Richard Nickl d'avoir accepté de rapporter ce manuscrit. Patrice, merci pour tes développements (pas seulement d'Edgeworth) sur le bootstrap, merci d'être là, toujours, avec simplicité et modestie, pour ceux qui rééchantillonnent, qui permutent, qui manient le jackknife, ou qui coupent les cheveux en... dix pour faire de la validation croisée. Cristina, merci pour tes contributions fondamentales à la théorie asymptotique des tests minimax : tu en es sans doute à ce jour la première spécialiste, et quel bonheur pour moi de pouvoir écrire ces mots au féminin ! Richard, thank you for your interest in my work, and for the relevant suggestions you made, that - really - helped to improve the paper I am probably most proud of. Thank you for your kind invitation to Evarist's birthday conference.

Merci à Béatrice Laurent, Oleg Lepski, Pascal Massart, Anne Philippe, et Alexandre Tsybakov, d'avoir bien voulu faire partie de mon jury. Béatrice, merci d'avoir été une directrice de recherche si exceptionnelle, d'être une enseignante et chercheuse si exemplaire. Merci pour l'amitié, la confiance, l'espoir que tu donnes, bref, merci pour l'« effet Béatrice » (Matthieu se reconnaîtra). Oleg, merci pour l'immensité et la finesse de tes connaissances sur le minimax. Merci de toujours les partager avec autant de passion et de générosité. Pascal, merci de m'avoir donné une famille scientifique, merci pour l'élégance de tes résultats sur la concentration de la mesure et la sélection de modèles, pour l'incroyable intuition statistique que tu nous as transmise, à notre mesure. Merci de croire encore à l'égalité des chances et d'avoir le courage d'agir pour... Merci pour ta belle âme d'universitaire. Anne, merci pour ta probité, jamais démentie dans tous ces comités de sélection que nous avons partagés, pour ton parcours scientifique, un modèle du genre. Sacha, merci pour ta bienveillance depuis ta présence dans mon jury de thèse comme rapporteur, pour tes encouragements. Merci pour ces cours inoubliables à l'IHP, pour l'indispensable ouvrage qui en est le fruit : sans eux, les méthodes à noyaux manqueraient fort vraisemblablement à mes travaux.

J'ai ici une pensée pour Evarist Giné et Yuri Ingster, que j'aurais tant aimé voir dans ce jury également.

Merci à mes compagnons d'aventures statistiques passées, présentes et futures. Bien sûr, je pense à Patricia en premier. Merci pour tout ce que nous partageons, pour tes milliers d'idées qui partent dans tous les sens, pour cette extraordinaire curiosité scientifique qui m'a permis de te rallier à la cause des tests, du bootstrap, et bientôt de l'apprentissage. Mélisande dit que nous voir travailler ensemble, c'est magique. Génie, dans mes trois vœux, il y a... : que cette magie dure encore ! Enfin, merci pour ton amitié, merci d'avoir été là, juste ce qu'il fallait, dans les moments difficiles que j'ai dû traverser. Christine, avec qui il n'est jamais de problème sans solution, merci pour ta si précieuse perspicacité, et ta sincérité à toute épreuve. Matthieu, merci pour ton inégalable et inébranlable enthousiasme, pour ton talent mathématique. Yann, virtuose de la parallélisation en C++, merci pour ton appui informatique. Merci pour ta patience quand Patricia et moi ne tenons presque pas deux minutes sans parler maths. François, merci pour la confiance que tu m'as toujours accordée. Eh oui, ça y est, je la soutiens cette habilitation ! Céline, merci pour notre collaboration, qui reste un excellent souvenir. Merci à Myriam, Nicolas, Ronan, Gwénaëlle, Arthur, Zoltán, Kacper, Victor-Emmanuel, Ania, Dominique, David, Sylvain, Gilles, pour nos divagations à venir...

Mélisande, merci pour ces trois années passées à travailler avec toi, pour ces rapports de lecture d'articles et d'ouvrages impressionnants de précision, pour ces airs de Chopin que tu magnifies. Merci d'être aussi pétillante, merci d'être une « niçoise pas comme les autres ».

Merci à tous mes collègues de Rennes 2, qui partagent avec moi ce quotidien souvent rock'n'roll. Merci pour les dîners improvisés, pour les échappées vers l'Italie, pour mon initiation aux danses bretonnes d'un soir, pour les douceurs littéraires et poétiques, échangées parfois bien tard, pour les critiques de séries (même si celle de The Good Wife me laisse perplexe), pour les débats (sans double barre de fin évidemment) piano/trompette...

Merci aux collègues de l'Ensai, qui ont partagé avec moi ce quotidien devenu souvent hard rock. Merci pour les rêveries écologiques, pour les réflexions didactiques et philosophiques (Michel Onfray n'a qu'à bien se tenir), pour les discussions autour du bootstrap, pour ma découverte d'un autre monde où tout est possible...

Merci aux autres collègues de Rennes, que j'ai plaisir à croiser lors de séminaires, conférences, réunions d'équipe ou séances de groupes de travail.

Merci à ceux que j'ai eu la chance de rencontrer à Orsay ou alentour : mes enseignants d'alors, mais aussi, par désordre alphabétique, Antoine, Frédérique, Gilles, Lucien, Peggy, Servane, Vincent, Yannick, et une légendaire Framboise.

Merci à mes compagnons d'aventures non statistiques : Anthony, Ariane, Catherine, Clarisse, Clément, David, Emmanuelle, François-Xavier, Guillaume, Hélène, Jean-Baptiste, Mélanie, Mohamed, Nathalie, Séverine, Yannick, la dream team *Toute l'APAEP court*, et à ma famille, Laurence, Damien, Camille et Carla, mes parents.

Merci à Léonor, Marceau, et Alexis...
Désormais, j'ai une petite princesse et deux petits princes dans ma vie.
Je dois tant à cet astéroïde B 612 !

Contents

Introduction

The present dissertation summarizes my research work in nonparametric Statistics, since my first steps as a PhD student under the supervision of Béatrice Laurent and Pascal Massart.

I fell in the amazing world of the bootstrap at the end of my Master's years when Pascal Massart encouraged me to read Evarist Giné's course for the St. Flour Summer School of Probability. Although I had, at this time, real bootstraps on my Dr. Martens shoes, which might have enabled me to pull up myself over it, I chose not to leave this world where *almost* anything is possible.
I particularly emphasize the word *almost* here, as one central guideline of my work was, and is still, the quantification of this *almost* through nonasymptotic studies of bootstrap approaches.

Nonasymptotic nonparametric Statistics is thus at the core of this thesis, with a main subject: minimax adaptive tests based on aggregation or multiple procedures, which is developed in four chapters, each devoted to a different generic testing problem. Statistical learning issues are also tackled as such in a chapter exclusively focusing on the binary classification problem, but also in two-sample and independence testing problems through kernel or nearest neighbors methods, always with general bootstrap approaches in background.
The present document is therefore organized in five chapters, whose content is briefly described below.

The first chapter is devoted to minimax adaptive goodness-of-fit tests in three different models: a density model, a periodic regression model, and a Poisson process model.
In the density model, where the observed data are modeled by an i.i.d. sample from a probability distribution with density f, we propose to test that f belongs to a set of densities $\mathcal{F}_0$. We first consider the case where $\mathcal{F}_0$ is reduced to a single density f_0, for which we construct minimax adaptive tests, based on an aggregation principle. The main contribution of this work, as compared with anterior ones, lies in the nonasymptotic nature of the developed tests, that can thus be implemented on real data sets with moderate size, without any loss of significance. Then, these tests are extended to the case where $\mathcal{F}_0$ is a parametric translation/scale family. Up to our knowledge, the obtained tests were the first minimax adaptive tests for such parametric null hypotheses.
In the fixed design regression model with a Gaussian noise, we consider the classical signal detection problem, focusing on the case where the signal is assumed to be periodic. We first measure the exact impact of such an assumption, by evaluating the minimax separation rates over periodic Sobolev balls. Then, motivated by a real application in target detection via laser vibrometry, we construct minimax adaptive tests that take advantage of the periodicity of the signal, thus leading to better theoretical and experimental performance than more general signal detection tests.
In the Poisson process model, where the observed data are modeled by a - possibly inhomogeneous - Poisson process, we are interested in the problem of testing that the process is homogeneous, that is, has a constant intensity. We particularly focus on alternatives with very irregular intensities that can be especially difficult to distinguish from constant intensities in practice. The determination of minimax separation rates over classes of such very irregular intensities, here chosen as weak Besov bodies, constitutes the most important contribution of this work. We indeed prove that these minimax separation rates are so large that there is, on the one hand, no additional price to pay for adaptivity and, on the other hand, no difference with the minimax risk in estimation problems, which is completely

unusual in the minimax testing scene. The corresponding minimax adaptive tests are constructed from an aggregation approach inspired by already existing ones in other settings, but with a new choice of critical values for the involved single tests, which is more general and which allows to detect more alternatives than usual.

The topics which are the most representative of my research concerns are certainly the two-sample problems where bootstrap approaches have to be considered to define nonasymptotic testing procedures. As most of the ideas developed in these topics are inspired by anterior works in statistical learning, and more precisely in binary classification, their description is postponed to the third chapter of the manuscript, the second one being dedicated to binary classification issues.
Thus, the second chapter presents the construction of classification rules as minimizers of the penalized empirical risk, with general weighted bootstrap penalties based on symmetrization arguments or other tricks from the empirical processes theory. This work, which extends papers on Rademacher penalties or complexities, may be viewed as a step towards routine nonasymptotic studies of general bootstrap approaches. The developed concentration tools are rather general and may indeed be used in other contexts. But due to the particular assumptions of the binary classification framework and the global minimax point of view that is adopted here, they have to be improved to fit more sophisticated statistical problems: the interested reader may have a look at the posterior developments of Arlot on extended bootstrap approaches, including cross validation ones.
A short work on the functional binary classification problem then follows, based on nearest neighbors and hold-out validation approaches.

The third chapter therefore deals, as announced, with two-sample problems. Three different models are considered as in the first chapter: a density model, a regression model (here with a random design and a heteroscedastic noise), and a Poisson process model. In all these models, we construct new testing procedures from aggregation schemes which involve specific nonasymptotic bootstrap or permutation approaches. The proposed test statistics are U-statistics based on general kernels, which are closely linked to minimax adaptive estimation by model selection, thresholding, and approximation kernel methods. The main novelty is that the kernels can also be chosen as reproducing characteristic kernels, which are now well-known in the statistical learning community, or as nearest neighbors kernels. In the Poisson process model, we prove that our tests satisfy nonasymptotic minimax adaptivity properties, and that the reproducing characteristic kernel choice in particular leads to a nice interpretation of some results in terms of uniform separation rates in the corresponding Reproducing Kernel Hilbert Space.

In the fourth chapter, we address the problem of testing independence of two point processes, motivated by the neuroscience problem of synchrony detection in spike train analysis. We introduce new testing procedures, whose test statistics are based on U-statistics, and whose critical values are constructed from bootstrap and permutation approaches. This work is rather atypical in the present thesis since the tests are here single tests (and not based on aggregation), mainly studied from an asymptotic point of view. Some foundations of a forthcoming nonasymptotic study are proposed in the last parts of the PhD thesis of Mélisande Albert.

The fifth and last chapter focuses on two very different multiple testing issues. The first one follows the theoretical work presented in the fourth chapter and deals with the neuroscience application which has motivated it. A multiple testing procedure is proposed and studied, in order to answer the problem of detecting precise locations of dependence periods between two spike trains, modeled by point processes. The second one is a purely theoretical issue about the definition of counterparts for uniform and minimax separation rates for multiple tests, with the ambition to lay the bases of a minimax theory for multiple testing.

As they represent the main concerns of the present dissertation, let us now give a common framework for nonparametric statistical hypothesis testing questions, essentially intended to set the notation and recall the usual basic vocabulary. Note that the notation and the vocabulary for binary classification

and multiple testing problems are set in the corresponding chapters, that is the second and fifth ones. The bootstrap and permutation formalism is introduced in the second, third and fourth chapters.

Considering an observed random variable $\mathbf{X}$, defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and given a possible set $\mathcal{P}$ of distributions P for $\mathbf{X}$, a hypothesis is defined through a subset of $\mathcal{P}$. In a classical single testing problem, two hypotheses are considered: the null hypothesis (H_0), which is viewed as the favorite one, and expressed from a subset $\mathcal{P}_0$ as

$$(H_0) \quad P \in \mathcal{P}_0,$$

and an alternative (H_1), expressed from a subset $\mathcal{P}_1 \subset \mathcal{P} \setminus \mathcal{P}_0$ as

$$(H_1) \quad P \in \mathcal{P}_1.$$

In a multiple testing problem, more hypotheses are considered, and then they are often confused with their associated subset of $\mathcal{P}$. Such convention is used in Chapter 5.
In an abstract way, a (single) test is a rule which allows to decide, from the observation of $\mathbf{X}$, to reject, or not, the null hypothesis (H_0) in favor of the alternative (H_1). Mathematically, it is represented as a statistic ϕ depending on $\mathbf{X}$,

- with value 1 when $\mathbf{X}$ leads to a rejection of (H_0) in favor of (H_1),
- with value 0 when $\mathbf{X}$ does not lead to a rejection of (H_0) in favor of (H_1),
- otherwise, with a value in $(0,1)$, equal to the conditional probability of rejection of (H_0) given $\mathbf{X}$.

In the present thesis, only nonrandomized tests are considered, that is tests with values in $\{0,1\}$. Such tests are evaluated through their first and second kind error rates

$$\mathrm{ER}_1(\phi) = \sup_{P \in \mathcal{P}_0} P(\phi = 1) \quad \text{and} \quad \mathrm{ER}_2(\phi) = \sup_{P \in \mathcal{P}_1} P(\phi = 0). \tag{1}$$

Note that the first kind error rate $\mathrm{ER}_1(\phi)$ of ϕ is also sometimes named the size of the test ϕ.
Given some prescribed error rates levels α and β in $(0,1)$, following the classical Neyman-Pearson principle, it is required in priority that

$$\mathrm{ER}_1(\phi) \leq \alpha, \tag{2}$$

and ϕ is then said to be level α, or of level α. It is next required that

$$\mathrm{ER}_2(\phi) \leq \beta, \tag{3}$$

with $\mathcal{P}_1$ as large as possible, ideally equal to $\mathcal{P} \setminus \mathcal{P}_0$.
However, in general, the set $\mathcal{P}$ of possible distributions for $\mathbf{X}$ is so large that constructing such "ideal" tests is impossible when $\alpha + \beta < 1$. Smaller sets of possible distributions are therefore considered. Then, a distance between the null hypothesis and these sets, from which condition (3) is guaranteed, is evaluated, leading to the notions of uniform separation rates and minimax separation rates, precisely defined in each concerned chapter of this dissertation.

Nonparametric tests of level α, that is satisfying (2), can thus be evaluated through their uniform separation rates over some sets of possible distributions of particular interest for the considered study, which are said to be optimal when they are of the same order as the corresponding minimax separation rates. These criteria are of course closely linked to the more traditional power function $P \in \mathcal{P}_1 \mapsto P(\phi = 1)$, which is used in other optimality criteria such as the consistency of the tests.
A test ϕ_α of level α is said to be less conservative than another test ϕ'_α if $\mathrm{ER}_1(\phi'_\alpha) \leq \mathrm{ER}_1(\phi_\alpha) \leq \alpha$. In particular, if basically $\phi'_\alpha = 1 \Rightarrow \phi_\alpha = 1$, then ϕ_α is clearly less conservative, but also more powerful than ϕ'_α.

A nonrandomized test ϕ_α of level α is often defined from a real valued test statistic, that is, a real valued statistic T, depending on $\mathbf{X}$, associated with a critical value c_α depending on α (and possibly $\mathbf{X}$), as $\phi_\alpha = \mathbb{1}_{\{T>c_\alpha\}}$ or $\phi_\alpha = \mathbb{1}_{\{T\geq c_\alpha\}}$. When $\alpha \mapsto \phi_\alpha$ is increasing in the sense that $0 \leq \alpha \leq \alpha' \leq 1 \Rightarrow \phi_\alpha \leq \phi_{\alpha'}$, this test can also be defined through its corresponding p-value given by

$$\inf\{\alpha, \phi_\alpha = 1\}.$$

Combining Lemma 1 and Corollary 14 in [14], we prove the following useful lemma, which allows to conveniently go back and forth test statistics and critical values on the one hand, p-values on the other hand. Before stating this result, let us recall some basic definitions.
Let T be a real valued statistic T depending on $\mathbf{X}$. It is well-known that the cumulative distribution function (c.d.f.) F of the distribution of T, defined by $F(t) = P(T \leq t)$ for all t in $\mathbb{R}$, is a càdlàg function, and that its generalized inverse function or quantile function F^{-1} defined by $F^{-1}(u) = \inf\{t,\ F(t) \geq u\}$ for all u in $[0,1]$ is a càglàd function. On the contrary, the function F_- defined by

$$\forall t \in \mathbb{R}, \quad F_-(t) = P(T < t), \tag{4}$$

is a càglàd function, and is thus named in the following the càglàd c.d.f. of the distribution of T. Its generalized inverse function F_-^{-1} defined by

$$\forall u \in (0,1), \quad F_-^{-1}(u) = \sup\{t,\ F_-(t) \leq u\}, \tag{5}$$

is then a càdlàg function. More properties about all these functions, and useful links between them can be found in [14, Lemma 13].

Lemma 1. *Let T be a real-valued statistic depending on $\mathbf{X}$, whose distribution does not depend on P provided that P belongs to $\mathcal{P}_0$. Denote by F and F_- the (càdlàg) c.d.f. and the càglàd c.d.f. of this distribution under (H_0), and by F^{-1} and F_-^{-1} their respective generalized inverse functions as defined above. Let $p(T) = 1 - F_-(T)$, and α be some fixed level in $(0,1)$. Then the four tests $\mathbb{1}_{\{T>F^{-1}(1-\alpha)\}}$, $\mathbb{1}_{\{T>F_-^{-1}(1-\alpha)\}}$, $\mathbb{1}_{\{p(T)\leq\alpha\}}$, $\mathbb{1}_{\{p(T)<\alpha\}}$ are of level α, and their associated p-value is $p(T)$, which satisfies*

$$\sup_{P\in\mathcal{P}_0} P\left(p(T) \leq \alpha\right) \leq \alpha.$$

Moreover,

$$\begin{cases} \mathbb{1}_{\{p(T)<\alpha\}} \leq \mathbb{1}_{\{p(T)\leq\alpha\}}, \\ \mathbb{1}_{\{p(T)<\alpha\}} = \mathbb{1}_{\{T>F_-^{-1}(1-\alpha)\}} \leq \mathbb{1}_{\{T>F^{-1}(1-\alpha)\}}, \\ \mathbb{1}_{\{p(T)\leq\alpha\}} = \mathbb{1}_{\{T>F_-^{-1}(1-\alpha)\}} = \mathbb{1}_{\{T>F^{-1}(1-\alpha)\}} \text{ a.s. if } F_-(F_-^{-1}(1-\alpha)) < 1-\alpha, \\ \mathbb{1}_{\{p(T)\leq\alpha\}} = \mathbb{1}_{\{T\geq F_-^{-1}(1-\alpha)\}} \text{ a.s. if } F_-(F_-^{-1}(1-\alpha)) = 1-\alpha. \end{cases}$$

Note that when the c.d.f. F is continuous, the four tests considered in the above lemma are almost surely equal. As this lemma also applies if F is a conditional c.d.f., it can be used for bootstrap or permutation based tests. In this case, the considered conditional distributions are naturally noncontinuous: the less conservative above test is of course $\mathbb{1}_{\{p(T)\leq\alpha\}}$.

Let us finally specify a few conventions and notations that are used in the present document.
All the collections of hypotheses, of tests, of classes, considered here are at most countable.
For every x, y in $\mathbb{R}$, we set $x \wedge y = \inf\{x, y\}$ and $x \vee y = \sup\{x, y\}$.
The symbols $\lfloor . \rfloor$ and $\lceil . \rceil$ respectively denote the floor and ceiling functions as usual, and for every x in $\mathbb{R}$, $(x)_+ = x\mathbb{1}_{x\geq 0}$.
For any finite set $\mathcal{C}$, $\#\mathcal{C}$ denotes the cardinality of $\mathcal{C}$ that is the number of elements in $\mathcal{C}$.
The Lebesgue measure on $\mathbb{R}$ or on $\mathbb{R}^d$ is denoted by λ.
All along this dissertation, C denotes a positive constant that may vary from line to line. The dependency of C with respect to various parameters is specified by the notation $C(.)$. Universal positive constants are more likely denoted by κ.

Chapter 1

Goodness-of-fit tests

1.1 Introduction

Goodness-of-fit testing problems are largely encountered in the statistical literature, being part of the oldest and most fundamental points in the hypothesis testing theory. From the historical Pearson's chi-square test to more modern tests based on kernel methods in the statistical learning spirit, many nonparametric goodness-of-fit tests have been developed in various models, in order to apply to more and more precise questions in numerous fields.
In this chapter, we deal with nonparametric goodness-of-fit tests in three different models: a classical density model, a periodic Gaussian regression model, and a Poisson process model, each being dedicated to specific applications.
$\mathbf{X}$ generally denotes a set of random variables which are defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, observed on an interval $\mathbb{X}$ of $\mathbb{R}$, whose distribution P_f depends on an unknown function f, typically assumed to belong to some subspace $\mathcal{F}$ of $\mathbb{L}_2(\mathbb{X}, \mu)$, for some σ-finite measure μ.

A goodness-of-fit testing problem can be expressed as the problem of testing

$$(H_0) \quad f \in \mathcal{F}_0 \quad \text{against} \quad (H_1) \quad f \notin \mathcal{F}_0,$$

for a given subset $\mathcal{F}_0$ of $\mathbb{L}_2(\mathbb{X}, \mu)$.

For any event $\mathcal{E}$ based on $\mathbf{X}$, $\mathbb{P}_{(H_0)}(\mathcal{E})$ then denotes as usual $\sup_{f \in \mathcal{F}_0} P_f(\mathcal{E})$, and $\mathbb{E}_f$ denotes the expectation with respect to P_f.

1.1.1 Nonasymptotic minimax adaptivity

The point of view that is adopted here to evaluate the considered tests is nonasymptotic, and based on minimax adaptivity criteria. So, given a first kind error level α in $(0, 1)$, any of our tests ϕ is primarily required to be of level α, that is to satisfy the property (2), that can also be expressed with the present notation as

$$(\mathcal{P}_{\text{level},\alpha}) \qquad \big| \ \mathbb{P}_{(H_0)}(\phi = 1) \leq \alpha.$$

Then, given a second kind error level β in $(0, 1)$, any of our tests ϕ is secondarily required to achieve, over several classes of alternatives simultaneously, the minimax separation rates defined for α and β as follows.

Definition 1 (Uniform and minimax separation rate)**.** Let d be a metric over the space $\mathcal{F}$, and a class of functions $\mathcal{F}_1 \subset \mathcal{F}$. Let α and β be fixed error rates levels in $(0, 1)$, and a test ϕ_α of (H_0) against (H_1) satisfying $(\mathcal{P}_{\text{level},\alpha})$.

The uniform separation rate of ϕ_α over $\mathcal{F}_1$ with prescribed second kind error rate β, for the metric d, is defined by

$$\mathrm{SR}_d^\beta(\phi_\alpha, \mathcal{F}_1) = \inf\left\{ r > 0, \sup_{f \in \mathcal{F}_1,\ d(f,\mathcal{F}_0) \geq r} P_f(\phi_\alpha = 0) \leq \beta \right\}.$$

The corresponding minimax separation rate over $\mathcal{F}_1$ with prescribed error rates α and β, for the metric d, is defined by

$$m\mathrm{SR}_d^{\alpha,\beta}(\mathcal{F}_1) = \inf_{\phi_\alpha \text{ satisfying } (\mathcal{P}_{\mathrm{level},\alpha})} \mathrm{SR}_d^\beta(\phi_\alpha, \mathcal{F}_1),$$

where the infimum is taken over all possible level α tests.

Definition 2 (Minimax (adaptive) test)**.** Let d be a metric over the space $\mathcal{F}$, and a collection $\overline{\mathcal{F}_1}$ of classes of functions $\mathcal{F}_1 \subset \mathcal{F}$. A level α test ϕ_α is said to be minimax over a class $\mathcal{F}_1$ of the collection $\overline{\mathcal{F}_1}$ for the metric d if $\mathrm{SR}_d^\beta(\phi_\alpha, \mathcal{F}_1)$ achieves $m\mathrm{SR}_d^{\alpha,\beta}(\mathcal{F}_1)$, possibly up to a multiplicative constant depending on α and β. It is said to be minimax adaptive over $\overline{\mathcal{F}_1}$ if $\mathrm{SR}_d^\beta(\phi_\alpha, \mathcal{F}_1)$ achieves, or nearly achieves, $m\mathrm{SR}_d^{\alpha,\beta}(\mathcal{F}_1)$, for every $\mathcal{F}_1$ in $\overline{\mathcal{F}_1}$ simultaneously, without knowing in advance to which class of the collection the function f may belong. This property is formalized in the following as

$$\left(\mathcal{P}_{\mathrm{adaptive},\alpha,\beta,\overline{\mathcal{F}_1},d}\right) \quad \Big| \ \mathrm{SR}_d^\beta(\phi_\alpha, \mathcal{F}_1) \text{ achieves or nearly achieves } m\mathrm{SR}_d^{\alpha,\beta}(\mathcal{F}_1), \text{ for every } \mathcal{F}_1 \text{ in } \overline{\mathcal{F}_1}.$$

These definitions due to Baraud [Bar02] translate, in a nonasymptotic framework, asymptotic criteria that in fact originate in Ingster's work [Ing82, Ing84, Ing93]. A precise definition of the asymptotic minimax testing setup with exact separation constants (that is not tackled in the present dissertation) can be found in [LT00].

We give below a condensed description of the main contributions to minimax testing and adaptive minimax testing, including the works presented here and references that are anterior, as well as posterior, to them. This bibliography is of course not exhaustive, as the literature about (adaptive) minimax testing is too huge to be presented in such a way, and I wish to apologize for authors that I omit here.

In particular, are only cited the minimax separation rates for $\mathbb{L}_2$-metrics as they are the only ones considered in this thesis. The investigated classes are Besov spaces $\mathcal{B}_{s,p,q}(R)$ or Besov bodies (see the books [Tri06, DL93, GN15] for definitions for instance, and articles on minimax estimation or testing, like [DJ98]). Some particular Besov spaces are distinguished, such as Hölder spaces $\mathcal{H}_s(R) = \mathcal{B}_{s,\infty,\infty}(C(s)R)$ and Sobolev spaces $\mathcal{S}_{s,2}(R) = \mathcal{B}_{s,2,2}(C(s)R)$, as they were the first ones to be studied in the fundamental papers by Ingster.

The references on multivariate Nikol'skii-Besov spaces are postponed to Section 3.2, since such anisotropic spaces are only considered in this section. Weak Besov spaces $w\mathcal{B}_{s'}(R')$ are defined in Section 1.4.

Moreover, testing problems in more complex models such as convolution models (see [But07, BMP09] for instance), high-dimensional models (see [VV10, ITV10, BI13]), or random graphs or networks models (see [ACV14] for instance), are not cited in the following summary: we only focus on the models considered in this chapter or on models close to them. In the following, $s'' = s - (1/(2p) - 1/4)_+$. The parameter ε stands for the standard deviation of the white noise model, while n denotes the sample size in the regression and density models, and the parameter which defines the asymptotics in each cited reference in the Poisson process model.

	Models			
	White noise	**Gaussian regression**	**Density**	**Poisson process**
	Minimax separation rates			
$\mathcal{H}_s(R)$	$\varepsilon^{\frac{4s}{4s+1}}$ [Ing93]	$n^{-\frac{2s}{4s+1}} (s > 1/4)$ [GP01]	$n^{-\frac{2s}{4s+1}}$ [Ing93, Pou02]	
$\mathcal{S}_{s,2}(R)$	$\varepsilon^{\frac{4s}{4s+1}}$ [Ing82]	$n^{-\frac{2s}{4s+1}} \wedge n^{-1/4}$ [HK99] [4]	$n^{-\frac{2s}{4s+1}}$ [Ing84]	$n^{-\frac{2s}{4s+1}}$ [IK07]
$\mathcal{B}_{s,p,q}(R)$	$\varepsilon^{\frac{4s''}{4s''+1}}$ [Ing93, $(p \geq 2)$] [LS99, $(p<2, sp>1)$]	$n^{-\frac{2s''}{4s''+1}} \wedge n^{-1/4}$ [Bar02, $(p=q\leq 2)$] [LLM12, $(p=q\leq 2)$]		$n^{-\frac{2s}{4s+1}}$ [IK07, $(p=2)$] [7, $(p=2, q=\infty)$]
$\mathcal{B}_{s,2,\infty}(R)$ $\cap w\mathcal{B}_{s'}(R')$				$n^{-\frac{2s}{1+4s}} (s' \leq 2s)$ $\left(\frac{\ln n}{n}\right)^{\frac{s'}{2s'+1}} (s'>(2s) \vee (1/2))$ [7]
	Adaptive minimax separation rates			
$\mathcal{H}_s(R)$		$\left(\frac{\sqrt{\ln\ln n}}{n}\right)^{\frac{2s}{4s+1}} (s>1/4)$ [GP05, BHL03]	$\left(\frac{\sqrt{\ln\ln n}}{n}\right)^{\frac{2s}{4s+1}}$ [Ing00] [5]	
$\mathcal{S}_{s,2}(R)$	$\left(\varepsilon\left(\ln\ln\left(\varepsilon^{-2}\right)\right)^{\frac{1}{4}}\right)^{\frac{4s}{4s+1}}$ [Spo96]	$\left(\frac{\sqrt{\ln\ln n}}{n}\right)^{\frac{2s}{4s+1}} (s>1/4)$ $n^{-1/4}$ $(s\leq 1/4)$ [Bar02, BHL03] [4]	$\left(\frac{\sqrt{\ln\ln n}}{n}\right)^{\frac{2s}{4s+1}}$ [Ing00]	
$\mathcal{B}_{s,p,q}(R)$	$\left(\varepsilon\left(\ln\ln\left(\varepsilon^{-2}\right)\right)^{\frac{1}{4}}\right)^{\frac{4s''}{4s''+1}}$ [Spo96, $(sp>1)$]		$\left(\frac{\sqrt{\ln\ln n}}{n}\right)^{\frac{2s}{4s+1}}$ [Ing00,$(p\geq 2, q>1)$] [5, $(p=2, q=\infty)$]	$\left(\frac{\sqrt{\ln\ln n}}{n}\right)^{\frac{2s}{4s+1}}$ [7, $(p=2, q=\infty)$]
$\mathcal{B}_{s,2,\infty}(R)$ $\cap w\mathcal{B}_{s'}(R')$				$\left(\frac{\sqrt{\ln\ln n}}{n}\right)^{\frac{2s}{4s+1}} (s' \leq 2s)$ $\left(\frac{\ln n}{n}\right)^{\frac{s'}{2s'+1}} (s'>(2s) \vee (1/2))$ [7]

1.1.2 Aggregated tests

Most of minimax adaptive tests have been constructed on the following general principle of aggregation. First, are considered a collection $\{\mathcal{F}_{0,m},\ m \in \mathcal{M}\}$ of subsets of $\mathcal{F}$ such that $\mathcal{F}_0 \subset \cap_{m\in\mathcal{M}}\mathcal{F}_{0,m}$, and the collection of associated hypotheses $\{(H_{0,m}),\ m \in \mathcal{M}\}$, such that

$$(H_{0,m})\ f \in \mathcal{F}_{0,m}.$$

Then, for every m in $\mathcal{M}$ and every level u in $(0,1)$, a single test $\phi_{m,u}$ of

$$(H_{0,m})\ f \in \mathcal{F}_{0,m} \quad \text{against} \quad (H_{1,m})\ f \notin \mathcal{F}_{0,m},$$

satisfying $(\mathcal{P}_{\text{level},u})$, is constructed.
Given α in $(0,1)$, a collection $\{u_{m,\alpha},\ m\in\mathcal{M}\}$ of individual levels in $(0,1)$ is chosen, and the aggregated test corresponding to the collection of single tests $\Phi_\alpha=\{\phi_{m,u_{m,\alpha}},\ m\in\mathcal{M}\}$ is defined as

$$\bar{\Phi}_\alpha=\sup_{m\in\mathcal{M}}\phi_{m,u_{m,\alpha}}. \tag{1.1}$$

The aggregated test $\bar{\Phi}_\alpha$ is thus defined as the test rejecting (H_0) when at least one $(H_{0,m})$ is rejected with $\phi_{m,u_{m,\alpha}}$.
Note that in most of the papers dealing with such aggregated tests, the $\phi_{m,u_{m,\alpha}}$'s are not considered as individual tests of $(H_{0,m})$ against $(H_{1,m})$, but as some tests of the original single null hypothesis (H_0) against the alternative (H_1). This is only in this case that one can say that each $\phi_{m,u_{m,\alpha}}$ is of level $u_{m,\alpha}$, since $\phi_{m,u_{m,\alpha}}$ only satisfies $\left(\mathcal{P}_{\text{level},u_{m,\alpha}}\right)$, which does not necessarily guarantee that $\sup_{f\in\mathcal{F}_{0,m}}P_f\left(\phi_{m,u_{m,\alpha}}=1\right)\leq u_{m,\alpha}$.

One of the main concerns here is to choose the collection $\{u_{m,\alpha},\ m\in\mathcal{M}\}$ of individual levels in $(0,1)$, so that the aggregated test $\bar{\Phi}_\alpha$ finally satisfies $(\mathcal{P}_{\text{level},\alpha})$.
Assume that for every m in $\mathcal{M}$, a test statistic T_m, whose distribution does not depend on the unknown function f provided that f belongs to $\mathcal{F}_0$, is available, which is typically the case when $\mathcal{F}_0$ is reduced to a singleton. Then, $\phi_{m,u}$ may be defined as

$$\phi_{m,u}=\mathbb{1}_{\left\{T_m>F_{m,-}^{-1}(1-u)\right\}},$$

where $F_{m,-}^{-1}$ is the càdlàg quantile function (see (5)) of the known distribution of T_m under (H_0).

Bonferroni-type aggregated tests. In this setup, the most obvious choice for $u_{m,\alpha}$ is a Bonferroni-type choice with $u_{m,\alpha}=\alpha/\#\mathcal{M}$. This leads to the Bonferroni-type aggregated test $\bar{\Phi}_\alpha^{Bonf}$ defined by (1.1), based on the collection

$$\Phi_\alpha^{Bonf}=\left\{\phi_{m,\alpha/\#\mathcal{M}},\ m\in\mathcal{M}\right\}=\left\{\mathbb{1}_{\left\{T_m>F_{m,-}^{-1}(1-\alpha/\#\mathcal{M})\right\}},\ m\in\mathcal{M}\right\}. \tag{1.2}$$

Given a family of positive weights $(w_m)_{m\in\mathcal{M}}$ such that $\sum_{m\in\mathcal{M}}w_m\leq 1$, a weighted Bonferroni-type choice with $u_{m,\alpha}=w_m\alpha$, can also be considered. This leads to the weighted Bonferroni-type aggregated test $\bar{\Phi}_\alpha^{wBonf}$ defined by (1.1), based on the collection

$$\Phi_\alpha^{wBonf}=\left\{\phi_{m,w_m\alpha},\ m\in\mathcal{M}\right\}=\left\{\mathbb{1}_{\left\{T_m>F_{m,-}^{-1}(1-w_m\alpha)\right\}},\ m\in\mathcal{M}\right\}. \tag{1.3}$$

Note that $\bar{\Phi}_\alpha^{Bonf}$ is a particular case of weighted Bonferroni-type aggregated test with $w_m=1/\#\mathcal{M}$, for every m in $\mathcal{M}$.
From the properties of the càglàd c.d.f. $F_{m,-}$ and the càdlàg quantile function $F_{m,-}^{-1}$ of the distribution of T_m under (H_0) (see for instance [14, Lemma 13]), we deduce the following result, whose proof is straightforward.

Lemma 2. *Given α in $(0,1)$, let Φ_α^{wBonf} be the collection of weighted Bonferroni-type tests defined by (1.3), and $\bar{\Phi}_\alpha^{wBonf}$ its corresponding aggregated test, as given by (1.1). Then $\bar{\Phi}_\alpha^{wBonf}$ satisfies $(\mathcal{P}_{level,\alpha})$, and*

$$\begin{cases}\Phi_\alpha^{wBonf} &= \left\{\mathbb{1}_{\left\{w_m^{-1}(1-F_{m,-}(T_m))<\alpha\right\}},\ m\in\mathcal{M}\right\},\\ \bar{\Phi}_\alpha^{wBonf} &= \mathbb{1}_{\left\{\min_{m\in\mathcal{M}}w_m^{-1}(1-F_{m,-}(T_m))<\alpha\right\}}.\end{cases}$$

Less conservative aggregated tests. A less conservative choice than the Bonferroni-type one was proposed by Baraud, Huet and Laurent [BHL03]. It consists in the aggregated test $\bar{\Phi}_\alpha^{BHL}$ defined by (1.1), based on the collection

$$\Phi_\alpha^{BHL} = \Big\{\phi_{m,u_{m,\alpha}},\ m \in \mathcal{M}\Big\} = \Big\{\mathbb{1}_{\{T_m > F_{m,-}^{-1}(1-u_{m,\alpha})\}},\ m \in \mathcal{M}\Big\}, \tag{1.4}$$

with

$$u_{m,\alpha} = \sup\left\{u,\ \mathbb{P}_{(H_0)}\left(\exists m \in \mathcal{M},\ T_m > F_{m,-}^{-1}(1-u)\right) \le \alpha\right\}.$$

We further propose in [7] (see Section 1.4) to consider a more general weighted version of this procedure. The test $\bar{\Phi}_\alpha^{FLR}$ is the aggregated test defined by (1.1), based on the collection

$$\Phi_\alpha^{FLR} = \Big\{\phi_{m,u_{m,\alpha}},\ m \in \mathcal{M}\Big\} = \Big\{\mathbb{1}_{\{T_m > F_{m,-}^{-1}(1-u_{m,\alpha})\}},\ m \in \mathcal{M}\Big\}, \tag{1.5}$$

with

$$u_{m,\alpha} = w_m \sup\left\{u,\ \mathbb{P}_{(H_0)}\left(\exists m \in \mathcal{M},\ T_m > F_{m,-}^{-1}(1-w_m u)\right) \le \alpha\right\}.$$

Note that the choice $w_m = 1/\#\mathcal{M}$, for every m in $\mathcal{M}$, allows to recover $\bar{\Phi}_\alpha^{BHL}$.

Lemma 3. *Given α in $(0,1)$, let Φ_α^{FLR} be the collection of tests defined by (1.5) and $\bar{\Phi}_\alpha^{FLR}$ be its corresponding aggregated test, as given by (1.1). Then, $\bar{\Phi}_\alpha^{FLR}$ satisfies $(\mathcal{P}_{level,\alpha})$ and if F_- denotes the càglàd c.d.f. (see (4)) of the distribution of $\min_{m\in\mathcal{M}} w_m^{-1}(1-F_{m,-}(T_m))$ under (H_0),*

$$\begin{cases} \Phi_\alpha^{FLR} &= \Big\{\mathbb{1}_{\{w_m^{-1}(1-F_{m,-}(T_m))<F_-^{-1}(\alpha)\}},\ m \in \mathcal{M}\Big\}, \\ \bar{\Phi}_\alpha^{FLR} &= \mathbb{1}_{\{\min_{m\in\mathcal{M}} w_m^{-1}(1-F_{m,-}(T_m))<F_-^{-1}(\alpha)\}}. \end{cases}$$

Furthermore, $\bar{\Phi}_\alpha^{wBonf} \le \bar{\Phi}_\alpha^{FLR}$, which means that the test $\bar{\Phi}_\alpha^{FLR}$ is less conservative than $\bar{\Phi}_\alpha^{wBonf}$. As a particular case, with $w_m = 1/\#\mathcal{M}$, $\bar{\Phi}_\alpha^{Bonf} \le \bar{\Phi}_\alpha^{BHL}$, which means that $\bar{\Phi}_\alpha^{BHL}$ is less conservative than $\bar{\Phi}_\alpha^{Bonf}$.

Proof. Let us consider Φ_α^{FLR} defined by (1.5). Then,

$$\begin{aligned} u_{m,\alpha} &= w_m \sup\left\{u,\ \mathbb{P}_{(H_0)}\left(\min_{m\in\mathcal{M}} w_m^{-1}(1-F_{m,-}(T_m)) < u\right) \le \alpha\right\}, \\ &= w_m F_-^{-1}(\alpha). \end{aligned}$$

Therefore,

$$\Phi_\alpha^{FLR} = \Big\{\mathbb{1}_{\{w_m^{-1}(1-F_{m,-}(T_m))<F_-^{-1}(\alpha)\}},\ m \in \mathcal{M}\Big\},$$

and the aggregated test $\bar{\Phi}_\alpha^{FLR}$ can be written as

$$\bar{\Phi}_\alpha^{FLR} = \mathbb{1}_{\{\min_{m\in\mathcal{M}} w_m^{-1}(1-F_{m,-}(T_m))<F_-^{-1}(\alpha)\}}.$$

The fact that $\bar{\Phi}_\alpha^{wBonf}$ satisfies $(\mathcal{P}_{\text{level},\alpha})$ leads to $F_-(\alpha) \le \alpha$, so $\alpha \le F_-^{-1}(\alpha)$ and $\bar{\Phi}_\alpha^{wBonf} \le \bar{\Phi}_\alpha^{FLR}$. □

Use of instrumental conditional distributions. In cases where $\mathcal{F}_0$ is not reduced to a singleton, the distribution of the test statistics under (H_0) may still depend on the unknown function f. Assume that given some random variable Z, the conditional distribution of T_m becomes free from f, provided that f belongs to $\mathcal{F}_0$. Denoting by $q_m^{(Z)}$ the càdlàg quantile function of this conditional distribution under (H_0), for every u in $(0,1)$, let $\phi_{m,u}$ be the test given by

$$\phi_{m,u} = \mathbb{1}_{\left\{T_m > q_m^{(Z)}(1-u)\right\}}.$$

Then notice that for every u in $(0,1)$, if f is in $\mathcal{F}_0$, by definition of $q_m^{(Z)}$,

$$P_f\left(\phi_{m,u}=1\right)=\mathbb{E}_f\left[P_f\left(T_m>q_m^{(Z)}(1-u)\Big|Z\right)\right]\leq u.$$

Hence, $\phi_{m,u}$ is a test of (H_0) against (H_1) of level u, and $q_m^{(Z)}(1-u)$ is a reasonable random critical value associated with T_m.
Now define a collection $\left\{u_{m,\alpha}^{(Z)},\ m\in\mathcal{M}\right\}$ of random variables in $(0,1)$, depending on Z, such that

$$u_{m,\alpha}^{(z)}=w_m\sup\left\{u,\ \mathbb{P}_{(H_0)}\left(\exists m\in\mathcal{M},\ T_m>q_m^{(z)}(1-w_m u)\Big|Z=z\right)\leq\alpha\right\}. \tag{1.6}$$

This leads to the collection of tests

$$\Phi_\alpha^{condFLR}=\left\{\phi_{m,u_{m,\alpha}^{(Z)}},\ m\in\mathcal{M}\right\}=\left\{\mathbb{1}_{\left\{T_m>q_m^{(Z)}\left(1-u_{m,\alpha}^{(Z)}\right)\right\}},\ m\in\mathcal{M}\right\}, \tag{1.7}$$

and the corresponding aggregated test $\bar{\Phi}_\alpha^{condFLR}$, defined by (1.1), which satisfies $(\mathcal{P}_{\text{level},\alpha})$.

Links with multiple tests. Some precise links between the above aggregated tests and classical Bonferroni and min-p multiple tests are established under particular conditions in [14, Proposition 2] (see Chapter 5, Section 5.3.1). Assume that for every m in $\mathcal{M}$, the distribution of T_m under $(H_{0,m})$ does not depend on f provided that f belongs to $\mathcal{F}_{0,m}$, and let $p(T_m)=1-F_{m,-}(T_m)$ be the p-value associated with the test $\phi_{m,u}=\mathbb{1}_{\left\{T_m>F_{m,-}^{-1}(1-u)\right\}}$ as in Lemma 1. Then, we have seen in Lemma 2 that

$$\Phi_\alpha^{wBonf}=\left\{\mathbb{1}_{\left\{w_m^{-1}p(T_m)<\alpha\right\}},\ m\in\mathcal{M}\right\}.$$

This collection of test is clearly related to the classical weighted Bonferroni multiple test based on the set of p-values $\{p(T_m),\ m\in\mathcal{M}\}$ for the collection of hypotheses $\{(H_{0,m}),\ m\in\mathcal{M}\}$.

Now, if the distribution of $\min_{m\in\mathcal{M}}w_m^{-1}p(T_m)$ is free from f when f belongs to $\cap_{m\in\mathcal{M}}\mathcal{F}_{0,m}$, and if F_- denotes here the càglàd c.d.f. of this distribution, one has also seen in Lemma 3 that

$$\Phi_\alpha^{FLR}=\left\{\mathbb{1}_{\left\{w_m^{-1}p(T_m)<F_-^{-1}(\alpha)\right\}},\ m\in\mathcal{M}\right\}.$$

This collection of test is precisely linked to the first step of a classical weighted min-p multiple test based on the set of p-values $\{p(T_m),\ m\in\mathcal{M}\}$ for the collection of hypotheses $\{(H_{0,m}),\ m\in\mathcal{M}\}$.

Finally, the aggregated test $\Phi_\alpha^{condFLR}$ can be related to multiple tests based on randomization such as the ones constructed by [RW05].

Notice however that the assumptions needed to establish the exact present parallel between aggregated tests and multiple tests, though apparently simple, are quite strong, and that among the testing problems investigated in this chapter, the only signal detection framework satisfies them with a collection of hypotheses $\{(H_{0,m}),\ m\in\mathcal{M}\}$ rich enough.

In the present chapter, we only consider continuous distributions, so $F_{m,-}$ and $F_{m,-}^{-1}$ are respectively equal to F_m and F_m^{-1}, which are the classical c.d.f. and quantile function of the known distribution of T_m under (H_0).

In the following, for any bounded real function g defined on $\mathbb{X}$, we set $\|g\|_\infty=\sup_{x\in\mathbb{X}}|g(x)|$, and for all $M>0$, $\mathbb{L}_\infty(M)$ denotes the $\mathbb{L}_\infty$-ball with radius M: $\mathbb{L}_\infty(M)=\{g:\mathbb{X}\to\mathbb{R},\ \|g\|_\infty\leq M\}$.

1.2 Goodness-of-fit tests in a density model

This section, mainly based on a joint paper with Béatrice Laurent [5], is devoted to goodness-of-fit tests in the following density model.

$\mathcal{M}^{(1)}_{\text{density}}$	$\mathbf{X} = \mathbf{X}_n = (X_1, \dots, X_n)$ is a sample of n i.i.d. random variables with distribution P_f of density f with respect to the Lebesgue measure λ on $\mathbb{X} = \mathbb{R}$.

We assume that f belongs to $\mathbb{L}_2(\mathbb{R}, \lambda)$, endowed with the classical $\mathbb{L}_2$-norm $\|.\|_2$, $\mathbb{L}_2$-metric d_2, and their corresponding scalar product $\langle ., . \rangle_2$.

Given a specified probability density f_0 with respect to λ, belonging to $\mathbb{L}_2(\mathbb{R}, \lambda)$, we consider the problems of testing

$$(H_0) \quad f \in \mathcal{F}_0 \quad \text{against} \quad (H_1) \quad f \notin \mathcal{F}_0,$$

where either $\mathcal{F}_0 = \{ f_0 \}$ or

$$\mathcal{F}_0 = \left\{ \frac{1}{\sigma} f_0 \left(\frac{\cdot - \mu}{\sigma} \right), (\mu, \sigma) \in K \right\}, \tag{1.8}$$

for some subset K of $\mathbb{R} \times (0, +\infty)$.

Many anterior works to [5] were devoted to the present goodness-of-fit testing problem in the density model from the minimax point of view, as seen in Section 1.1.1, the closest one being [Ing00]. However, none of them were based on nonasymptotic approaches, and even if considering the asymptotic point of view, very few of them dealt with the case where $\mathcal{F}_0$ is a parametric family. To our knowledge, the only one was in fact [Pou02], yet, in this article, the issue of adaptivity is not tackled. Hence the exact adaptive minimax separation rates were not known.
Our paper [5] fills a part of these gaps: here are constructed nonasymptotic goodness-of-fit tests, that are minimax adaptive at least when $\mathcal{F}_0 = \{f_0\}$, for which the classical minimax adaptive rates are achieved. In the case where $\mathcal{F}_0$ is a parametric family, a loss of efficiency of the order of a $(\ln n)^{1/2}$ factor (instead of the usual $(\ln \ln n)^{1/2}$ one) is highlighted. Although similar phenomena have already been observed in [4], or [CD12], the obtained results do not allow to be sure that such an extra $(\ln n)^{1/2}$ factor is unavoidable.
As explained in the following, the proposed tests are closely linked with model selection approaches. Other tests were developed in this spirit, such as the data-driven versions of Neyman's test [Ney37] proposed in [BR92, KL95, IKL97]. But all those tests were constructed with a penalized criterion usually dedicated to the estimation of the density f, such as Schwarz's BIC, so that the tests can not be minimax adaptive. We prove in Section 1.2.2 that a more appropriate choice for the penalty term would be the one used in the estimation of quadratic functionals of f like in [Lau05]. Our tests are based on an even sharper choice.

1.2.1 Minimax adaptive goodness-of-fit tests for a noncomposite null hypothesis

We here consider the case where $\mathcal{F}_0$ is the singleton $\{f_0\}$.
The tests we propose are based on the aggregation principle presented in Section 1.1.2. Considering a collection of subspaces $\{ S_m,\ m \in \mathcal{M} \}$ of $\mathbb{L}_2(\mathbb{R}, \lambda)$, we introduce for every m in $\mathcal{M}$,

$$\mathcal{F}_{0,m} = \{ f \in \mathbb{L}_2(\mathbb{R}, \lambda),\ \Pi_{S_m}(f - f_0) = 0 \},$$

where Π_{S_m} denotes the orthogonal projection onto S_m with respect to $\langle ., . \rangle_2$.
Following the aggregation principle, a first step is then to construct single tests of

$$(H_{0,m}) \quad \Pi_{S_m}(f - f_0) = 0 \quad \text{against} \quad (H_{1,m}) \quad \Pi_{S_m}(f - f_0) \neq 0.$$

These tests are here based on the estimation of $\|\Pi_{S_m}(f - f_0)\|_2^2$.

For all m in $\mathcal{M}$, let $\{b_l, l \in \mathcal{L}_m\}$ be some orthonormal basis of S_m with respect to $\langle ., . \rangle_2$.
The statistic

$$T_m = T_m(X_1, \dots, X_n) = \frac{1}{n(n-1)} \sum_{l \in \mathcal{L}_m} \sum_{i \neq j = 1}^{n} b_l(X_i) b_l(X_j) + \|f_0\|_2^2 - \frac{2}{n} \sum_{i=1}^{n} f_0(X_i) \tag{1.9}$$

is then an unbiased estimator of $\|\Pi_{S_m}(f)\|_2^2 + \|f_0\|_2^2 - 2\langle f, f_0 \rangle_2$, and thus a reasonable estimator of $\|\Pi_{S_m}(f - f_0)\|_2^2$. A particularly interesting case is when f_0 belongs to S_m, so that T_m is an unbiased estimator of $\|\Pi_{S_m}(f - f_0)\|_2^2$.
Let us denote by $F_{m,-}^{-1}$ the càdlàg quantile function of the distribution of T_m under the null hypothesis (H_0), which is, of course, completely free from the unknown density f.
We can now consider the single test

$$\phi_{m,u} = \mathbb{1}_{\left\{T_m > F_{m,-}^{-1}(1-u)\right\}},$$

and given a fixed level α in $(0,1)$, the collections of tests Φ_α^{Bonf} and Φ_α^{BHL}, as well as the corresponding aggregated tests $\bar{\Phi}_\alpha^{Bonf}$ and $\bar{\Phi}_\alpha^{BHL}$, as defined in Section 1.1.2 by (1.2), (1.4), and (1.1).

Notice that, except in very particular cases (see Lemma 1), the càdlàg quantile function can be replaced by the usual quantile function F_m^{-1}, hence, for simplicity, we only use F_m^{-1} in the following.

We introduce a collection of linear subspaces $\{ S_m,\ m \in \mathcal{M} \}$ generated by constant piecewise functions and scaling functions from a wavelet basis.

For all D in $\mathbb{N} \setminus \{0\}$ and k in $\mathbb{Z}$, let $b_{(1,D,k)} = \sqrt{D} \mathbb{1}_{[k/D,(k+1)/D)}$.
For all $D = 2^J$, J in $\mathbb{N}$, and k in $\mathbb{Z}$, let $b_{(2,D,k)} = 2^{J/2} \varphi(2^J . - k)$, where φ is a compactly supported scaling function or father wavelet such that, associated with a mother wavelet ψ, for all J in $\mathbb{N}$, $\left\{ 2^{J/2} \varphi(2^J . - k),\ k \in \mathbb{Z} \right\} \cup \left\{ \psi_{j,k} = 2^{j/2} \psi(2^j . - k),\ j \in \mathbb{N},\ j \geq J,\ k \in \mathbb{Z} \right\}$ form an orthonormal wavelet basis of $\mathbb{L}_2(\mathbb{R}, \lambda)$.

Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be some subsets of $\mathbb{D}_1 = \mathbb{N} \setminus \{0\}$ and $\mathbb{D}_2 = \{2^J,\ J \in \mathbb{N}\}$ respectively, such that $\mathcal{D}_1 \cup \mathcal{D}_2 \neq \emptyset$. Then, the considered collection $\{ S_m,\ m \in \mathcal{M} \}$ is defined by:

- $\mathcal{M} = \{ (\delta, D),\ \delta \in \{1,2\},\ D \in \mathcal{D}_\delta \}$,
- For $m = (\delta, D)$ in $\mathcal{M}$,
  - $\mathcal{L}_{(1,D)} = \{ (1, D, k),\ k \in \mathbb{Z} \}$,
  - $\mathcal{L}_{(2,D)} = \{ (2, D, k),\ k \in \mathbb{Z} \}$.
  - S_m is generated by $\{ b_l,\ l \in \mathcal{L}_m \}$ with the b_l's defined above.

Theorem 1 (Fromont, Laurent, 2006)**.** *Let $\mathbf{X} = \mathbf{X}_n$ be an i.i.d. sample distributed according to $\mathcal{M}_{\text{density}}^{(1)}$, and let f_0 be a specified density with respect to λ. Assume that f and f_0 are bounded functions. Fix some levels α and β in $(0,1)$, and let $\bar{\Phi}_\alpha$ be one of the tests $\bar{\Phi}_\alpha^{Bonf}$ and $\bar{\Phi}_\alpha^{BHL}$ defined above, with the corresponding individual levels $u_{m,\alpha}$'s. For any ε in $(0,2)$, there exist some positive constants $C_1(\beta)$ and $C_2(\beta, \varepsilon, \|f\|_\infty, \|f_0\|_\infty)$ such that $\mathbb{P}_f\left(\bar{\Phi}_\alpha = 0\right) \leq \beta$, as soon as*

$$\|f - f_0\|_2^2 > (1+\varepsilon) \inf_{m \in \mathcal{M}} \left\{ \|f - \Pi_{S_m}(f)\|_2^2 + F_m^{-1}(1 - u_{m,\alpha}) \right.$$
$$\left. + \frac{C_1(\beta)}{n} \left(\left(\sqrt{\|f\|_\infty} + \|f\|_\infty \right) \sqrt{D} + \frac{D}{n} \right) + \frac{C_2(\beta, \varepsilon, \|f\|_\infty, \|f_0\|_\infty)}{n} \right\}.$$

Let us notice that taking a single test $\phi_{m,\alpha}$ instead of $\bar{\Phi}_\alpha$, the same result holds but with $F_m^{-1}(1 - u_{m,\alpha})$ replaced by $F_m^{-1}(1 - \alpha)$. For a well-chosen collection $\{S_m,\ m \in \mathcal{M}\}$, these quantiles in fact only differ

by a logarithmic factor, which means that the aggregated test behaves almost as well as the best test among the considered collection of single tests. In this sense, the result of Theorem 1 can be viewed as an oracle type result. Such an oracle type result, with appropriate bias and variance terms, is known to lead to minimax adaptivity results, as proved below.

Let us consider the following classes of alternatives:

$$\mathcal{F}_{s,\delta}(R) = \Big\{ f \in \mathbb{L}_2(\mathbb{R},\lambda),\ \forall D \in \mathbb{D}_\delta,\ \Big\| f - \Pi_{S_{(\delta,D)}}(f) \Big\|_2^2 \leq R^2 D^{-2s} \Big\},$$

for $s > 0$, δ in $\{1,2\}$, and $R > 0$, which are linked with more classical Hölder balls or Besov bodies. Indeed, let for all $s = s_1 + s_2 > 0$ ($s_1 \in \mathbb{N}$, $s_2 \in (0,1]$), and $R > 0$,

$$\mathcal{H}_s(R) = \Big\{ f : [0,1] \to \mathbb{R},\ \forall x, y \in [0,1],\ |f^{(s_1)}(x) - f^{(s_1)}(y)| \leq R|x-y|^{s_2} \Big\},$$

and for all $s > 0$ and $R > 0$,

$$\mathcal{B}_{s,2,\infty}(R) = \left\{ f \in \mathbb{L}_2(\mathbb{R},\lambda),\ \forall j \in \mathbb{N},\ \sum_{k\in\mathbb{Z}} \langle f, \psi_{j,k}\rangle_2^2 \leq R^2 2^{-2js} \right\}.$$

Then, for s in $(0,1]$, $R > 0$, $\mathcal{H}_s(R) \subset \mathcal{F}_{s,1}(R)$, and for $s > 0$, $R > 0$, $\mathcal{B}_{s,2,\infty}(R) \subset \mathcal{F}_{s,2}\left((1-4^{-s})^{-1/2}R\right)$.

Corollary 1 (Fromont, Laurent, 2006)**.** *Take the same notations as in Theorem 1, and assume that* $n \geq 16$. *Choose* $\mathcal{D}_1 = \mathcal{D}_2 = \{2^J, 0 \leq J \leq \log_2\left(n^2/(\ln\ln n)^3\right)\}$. *For all* δ *in* $\{1,2\}$, $R' > 0$, *and* $s > 0$, $R > 0$ *such that*

$$(\ln\ln n)^{s+1/2}(\ln n)^{2s+1/2} n^{-1/2} \leq R \leq n^{2s}\left(\ln\ln n\right)^{-(3s+1/2)}, \tag{1.10}$$

there exists $C\left(s,\alpha,\beta,R',\|f_0\|_\infty\right)$ *such that*

$$\mathrm{SR}_{d_2}^{\beta}\left(\bar{\Phi}_\alpha, \mathcal{F}_{s,\delta}(R) \cap \mathbb{L}_\infty(R')\right) \leq C(s,\alpha,\beta,R',\|f_0\|_\infty) R^{\frac{1}{4s+1}} \left(\frac{\sqrt{\ln\ln n}}{n}\right)^{\frac{2s}{4s+1}}.$$

Note that when n is large enough the range (1.10) is not constraining, and all positive value for R and s can be taken into account.

It is known from the results of [Ing93] and [Ing00] that, in the particular case where $f_0 = \mathbb{1}_{[0,1]}$, the asymptotic minimax rate of testing over a Hölder, Sobolev or Besov ball with smoothness parameter s in $\mathbb{L}_2([0,1],\lambda)$ is of order $n^{-2s/(4s+1)}$ and that the loss of efficiency of order $(\ln\ln n)^{s/(4s+1)}$ is the price to pay for adaptivity. Hence, one can say in this case that considering the collection of classes of alternatives

$$\overline{\mathcal{F}}_1 = \left\{ \mathcal{F}_{s,\delta}(R) \cap \mathbb{L}_\infty(R'),\ \delta \in \{1,2\},\ R' > 0, \text{ and } s > 0,\ R > 0, \text{ satisfying (1.10)} \right\},$$

in the notation and under conditions of Corollary 1, the aggregated tests $\bar{\Phi}_\alpha^{Bonf}$ and $\bar{\Phi}_\alpha^{BHL}$ both satisfy $\left(\mathcal{P}_{\text{adaptive},\alpha,\beta,\overline{\mathcal{F}_1},d_2}\right)$, for every fixed levels α and β in $(0,1)$.

We even prove (see [5, Corollary 2]) that for n large enough, the collection $\overline{\mathcal{F}}_1$ can be extended to the classes $\mathcal{H}_s(R) \cap \mathbb{L}_\infty(R')$ $(s, R, R' > 0)$ though when $s > 1$, functions f in $\mathcal{H}_s(R)$ are not well approximated by their projections onto the considered linear spaces.

Moreover, in this case again, adding some spaces defined from the Fourier basis to the collection $\{S_m,\ m \in \mathcal{M}\}$ allows to obtain even more general theoretical results and to significantly improve the estimated powers of the tests in practice.

We are aware that the above results may be somewhat disappointing for readers used to multiple testing as the Bonferroni-type aggregated tests is proved to be minimax adaptive, as well as the BHL-type tests involving more sophisticated individual levels for the single tests. We have however seen in Lemma 3 that $\bar{\Phi}_\alpha^{BHL}$ is less conservative than $\bar{\Phi}_\alpha^{Bonf}$, which also clearly appears in our simulation study, so we guess that this could be seen here if a special attention was given to the constants appearing in the upper bounds for the uniform separation rates.

1.2.2 Links with model selection

In the present density model, the works on the estimation of the density f itself via model selection by penalization are numerous (see [BM97, BBM99, Cas00, Cas03], or [Sau, Ler12] for more recent references). It is thus now well-known that considering a collection of linear subspaces $\{S_m,\ m \in \mathcal{M}\}$ of $\mathbb{L}_2(\mathbb{R}, \lambda)$, where each S_m is generated by the basis $\{b_l,\ l \in \mathcal{L}_m\}$, f may be estimated by $\hat{f}_{\hat{m}} = \sum_{l\in\mathcal{L}_{\hat{m}}} \left(n^{-1}\sum_{i=1}^n b_l(X_i)\right) b_l$, with

$$\hat{m} = \operatorname{argmin}_{m\in\mathcal{M}} \left(-\sum_{l\in\mathcal{L}_m} \left(n^{-1}\sum_{i=1}^n b_l(X_i)\right)^2 + \operatorname{pen}(m)\right).$$

The penalty term $\operatorname{pen}(m)$ is chosen so that $\hat{f}_{\hat{m}}$ satisfies an oracle inequality and minimax adaptivity properties. Laurent [Lau05] proved that roughly estimating the quadratic functional $\theta(f) = \|f\|_2^2$ by $\|\hat{f}_{\hat{m}}\|^2$ does not however lead to satisfactory oracle and minimax adaptivity results. She proposes instead the corrected estimator:

$$\hat{\theta} = \sup_{m\in\mathcal{M}} \left(\frac{1}{n(n-1)} \sum_{l\in\mathcal{L}_m} \sum_{i\neq j=1}^n b_l(X_i)b_l(X_j) - \operatorname{pen}'(m)\right),$$

where $\operatorname{pen}'(m)$ is a penalty term which is not of the same order as $\operatorname{pen}(m)$. This penalty depends on the observed random variable $\mathbf{X}$ and a constant that remains to be precisely determined in practice. In particular, Laurent proved that when $f = \mathbb{1}_{[0,1]}$, for some constant $C > 0$, $\mathbb{E}[(\hat{\theta} - 1)^2] \leq C/n^2$. When the X_i's have their values in $[0, 1]$, a level α aggregated test of uniformity can then be obtained with

$$\mathbb{1}_{\left\{\hat{\theta}-1>\sqrt{C}/(n\sqrt{\alpha})\right\}} = \sup_{m\in\mathcal{M}} \mathbb{1}_{\left\{T_m>\operatorname{pen}'(m)+\sqrt{C}/(n\sqrt{\alpha})\right\}}.$$

But recall that the penalty term $\operatorname{pen}'(m)$ depends on a constant that has to be calibrated in practice in the estimation problem. In a testing context, calibrating the penalty amounts to calibrating some individual critical values $c_{m,\alpha}$'s so that the aggregated test $\sup_{m\in\mathcal{M}} \mathbb{1}_{\{T_m>c_{m,\alpha}\}}$, is of level α. Here, we take deterministic critical values $c_{m,\alpha} = F_m^{-1}(1 - u_{m,\alpha})$, but thinking that the penalty $\operatorname{pen}'(m)$ chosen in [Lau05] depends on $\mathbf{X}$, we could choose data-dependent critical values, as explained in Section 1.1.2, or in the spirit of bootstrap based tests (see Chapter 3 and Chapter 4).

1.2.3 Extension to a composite null hypothesis based on a translation/scale family

Let us now focus on the parametric translation/scale family $\mathcal{F}_0$ defined by (1.8), of which the families of Gaussian, uniform or exponential densities and translation models are typical examples.

The present testing procedures are essentially based on the idea that f belongs to $\mathcal{F}_0$ if and only if there exists (μ, σ) in K such that the variables $(X_i - \mu)/\sigma$ have P_{f_0} as distribution.
Considering a collection $\{S_m, m \in \mathcal{M}\}$ of linear subspaces of $\mathbb{L}_2(\mathbb{R}, \lambda)$ as in the above section, but with a right-continuous and Lipschitz scaling function φ, let for every m in $\mathcal{M}$,

$$\tilde{T}_m(X_1, \ldots, X_n) = \inf_{(\mu,\sigma)\in K} T_m\left(\frac{X_1 - \mu}{\sigma}, \ldots, \frac{X_n - \mu}{\sigma}\right).$$

Since the b_l's (l in $\mathcal{L}_m$), generating S_m, are right-continuous, the infimum over (μ, σ) in K can be replaced by the infimum over (μ, σ) in $K \cap \mathbb{Q}^2$ so that $\tilde{T}_m(X_1, \ldots, X_n)$ is indeed a random variable.

Note that the density of the variables $(X_i - \mu)/\sigma$ is $\sigma f(\sigma\ .\ + \mu)$ so $\tilde{T}_m(X_1, \ldots, X_n)$ is a reasonable estimator of

$$\inf_{(\mu,\sigma)\in K} \sigma \left\| f - \frac{1}{\sigma} f_0\left(\frac{\cdot - \mu}{\sigma}\right)\right\|_2^2,$$

which is closely related to the distance $d_2(f, \mathcal{F}_0)$.
Given a fixed level α in $(0,1)$, we define the collection of tests

$$\tilde{\Phi}_\alpha^{BHL} = \Big\{ \mathbb{1}_{\{\tilde{T}_m(X_1,\dots,X_n) > F_m^{-1}(1-u_{m,\alpha})\}},\ m \in \mathcal{M} \Big\},$$

where F_m^{-1} still denotes the quantile function of the distribution of T_m when $f = f_0$, and where $u_{m,\alpha}$ is chosen as in Φ_α^{BHL} above that is

$$u_{m,\alpha} = \sup\left\{ u,\ P_{f_0}\left(\exists m \in \mathcal{M},\ T_m > F_m^{-1}(1-u) \right) \le \alpha \right\}.$$

This enables to define according to (1.1) the corresponding aggregated test:

$$\bar{\tilde{\Phi}}_\alpha^{BHL} = \sup_{m \in \mathcal{M}} \mathbb{1}_{\{\tilde{T}_m(X_1,\dots,X_n) > F_m^{-1}(1-u_{m,\alpha})\}}.$$

This test satisfies $(\mathcal{P}_{\text{level},\alpha})$, based on the fact that if f belongs to $\mathcal{F}_0$, there exists (μ,σ) in K such that $f = f_0((.-\mu)/\sigma)/\sigma$, and

$$\tilde{T}_m(X_1,\dots,X_n) \le T_m\left(\frac{X_1-\mu}{\sigma}, \dots, \frac{X_n-\mu}{\sigma} \right).$$

Let us distinguish a particular case, based on the following invariance property:

$$\forall\ (\mu,\sigma) \in K,\ \tilde{T}_m(X_1,\dots,X_n) = \tilde{T}_m\left(\frac{X_1-\mu}{\sigma}, \dots, \frac{X_n-\mu}{\sigma} \right), \tag{1.11}$$

satisfied for instance when $K = \mathbb{R} \times (0,+\infty)$, $K = \{0\} \times (0,+\infty)$, or $K = \mathbb{R} \times \{1\}$.
When (1.11) holds, $F_m^{-1}(1-u_{m,\alpha})$ in $\bar{\tilde{\Phi}}_\alpha^{BHL}$ should be replaced by $\tilde{F}_m^{-1}(1-\tilde{u}_{m,\alpha})$, where $\tilde{F}_m^{-1}$ denotes the quantile function of the distribution of $\tilde{T}_m(X_1,\dots,X_n)$ when $f = f_0$, and

$$\tilde{u}_{m,\alpha} = \sup\left\{ u,\ P_{f_0}\left(\exists m \in \mathcal{M},\ \tilde{T}_m(X_1,\dots,X_n) > \tilde{F}_m^{-1}(1-u) \right) \le \alpha \right\}.$$

Such a choice leads to a less conservative test, but still guarantees that the test satisfies $(\mathcal{P}_{\text{level},\alpha})$ as soon as (1.11) holds.

Minimax adaptivity properties of our testing procedures are investigated in the case where $K = [\underline{\mu},\overline{\mu}] \times [\underline{\sigma},\overline{\sigma}]$, $\underline{\mu},\overline{\mu},\underline{\sigma},\overline{\sigma}$ being real numbers such that $\underline{\sigma} > 0$, f_0 is a given bounded density, satisfying some particular Lipschitz condition, and $\mathcal{F}_0$ satisfies a $\mathbb{L}_2$-entropy with bracketing condition, which holds for instance if $\mathcal{F}_0$ is a family of Gaussian densities, exponential densities, and uniform densities. We assume that there exist $v > 0$ and $c > 0$ such that for all $k \ge 2$,

$$\mathbb{E}\left[|X_i|^k \right] \le \frac{k!}{2} v c^{k-2},$$

and that f is bounded. Let n be large enough so that $n \ge 3$ and $n^2(\overline{\mu}-\underline{\mu}) \wedge (\overline{\sigma}-\underline{\sigma}) \ge 2$.
Consider now for all $s > 0$, and $R > 0$ the class:

$$\tilde{\mathcal{F}}_{s,2}(R) = \Big\{ f \in \mathbb{L}_2(\mathbb{R},\lambda),\ \forall D \in \mathbb{D}_2,\ \forall (\mu,\sigma) \in K, \\ \|\sigma f(\sigma.+\mu) - \Pi_{S_{(2,D)}}(\sigma f(\sigma.+\mu))\|_2^2 \le R^2 \sigma^{1+2s} D^{-2s} \Big\}.$$

Note that there exists some constant $\kappa > 0$ such that $\mathcal{B}_{s,2,\infty}(R) \subset \tilde{\mathcal{F}}_{s,2}\big(\kappa^{1/2}(1-4^{-s})^{-1/2}R\big)$.
We prove in [1] and [5], that when $\mathcal{D}_1 \subset \{2^J, 0 \le J \le \log_2(n^2)\}$ and $\mathcal{D}_2 = \{2^J, 0 \le J \le \log_2(n^2/\log^3 n)\}$, for all $R' > 0$, and $s > 0$, $R > 0$ satisfying

$$(\ln n)^{s+1/2} n^{-1/2} \le R \le n^{2s} (\ln n)^{-(3s+1/2)},$$

there exists $C = C(\underline{\mu}, \overline{\mu}, \underline{\sigma}, \overline{\sigma}, \|f_0\|_\infty, v, c, \alpha, \beta, s, R')$ such that

$$\mathrm{SR}_{d_2}^{\beta}\left(\tilde{\bar{\Phi}}_{\alpha}^{BHL}, \tilde{\mathcal{F}}_{s,2}(R) \cap \mathbb{L}_\infty(R')\right) \leq C R^{\frac{1}{4s+1}} \left(\frac{\sqrt{\ln n}}{n}\right)^{\frac{2s}{4s+1}}.$$

The above upper bound corresponds, up to a logarithmic factor, to the upper bound for the goodness-of-fit tests when $\mathcal{F}_0 = \{ f_0 \}$ obtained in Corollary 1. As explained at the beginning of this section, we do not know whether this logarithmic factor could be avoided.
Considering the present composite null hypothesis may amount to considering, in addition to the classical adaptivity with respect to the smoothness of the density, which typically entails a loss of efficiency of the order of a $(\ln \ln n)^{1/2}$ factor, another kind of adaptivity, this one, with respect to the unknown parameters (μ, σ) under (H_0). This could explain that the usual extra $(\ln \ln n)^{1/2}$ factor is here replaced by an extra $(\ln n)^{1/2}$ factor. Similar phenomena are observed in [4] (see Section 1.3) and [CD12] for instance, where the issue of the adaptivity with respect to the unknown period, and the unknown translation parameter are respectively treated in regression setups. Such a logarithmic factor also appears in minimax separation rates evaluated with the $\mathbb{L}_\infty$-metric (see [LT00] for instance).

1.2.4 Concentration inequalities: basic tools for nonasymptotic properties

Although asymptotic results on the evaluation of minimax separation rates are here cited, and used to justify the minimax adaptivity properties of our tests, the obtained results and in particular our upper bounds for the uniform separation rates are of nonasymptotic nature. They indeed hold for a fixed sample size n (as large as it is), highlighting the dependence of these uniform separation rates with respect to the radius R of the considered classes of alternatives. The fundamental tools to obtain such nonasymptotic results are concentration inequalities.
As the test statistics we introduce are constructed from the estimation of quadratic and linear functionals of f, they are naturally based on U-statistics of order 2, and linear statistics. More precisely, denoting by P_n the empirical process associated with the sample $\mathbf{X}_n$, the test statistic T_m can be decomposed as

$$T_m = \frac{1}{n(n-1)} \sum_{i \neq j = 1}^{n} \sum_{l \in \mathcal{L}_m} \left(b_l(X_i) - \langle f, b_l \rangle_2\right) \left(b_l(X_j) - \langle f, b_l \rangle_2\right) + 2 \int_{\mathbb{R}} \left(\Pi_{S_m}(f) - f_0\right) \left(dP_n - dP_f\right) + \|f - f_0\|_2^2 - \|f - \Pi_{S_m}(f)\|_2^2,$$

where the first term is a U-statistic of order 2 and the second term a linear statistic.
In order to precisely control T_m as well as its quantile function, we therefore use the concentration inequalities for U-statistics and linear statistics of [HRB03] and [BM98] respectively.
For the problem of testing that f belongs to a translation/scale parametric family, we have to control

$$\tilde{T}_m(X_1, \ldots, X_n) = - \sup_{(\mu,\sigma) \in K} \left(-T_m \left(\frac{X_1 - \mu}{\sigma}, \ldots, \frac{X_n - \mu}{\sigma} \right) \right).$$

Assuming that K is compact enables to consider a finite and bounded grid of points in K, and to reduce the problem to a control of a supremum over this finite grid on the one hand, to a control of the supremum of the difference

$$T_m \left(\frac{X_1 - \mu_1}{\sigma_1}, \ldots, \frac{X_n - \mu_1}{\sigma_1} \right) - T_m \left(\frac{X_1 - \mu}{\sigma}, \ldots, \frac{X_n - \mu}{\sigma} \right),$$

on a compact set $[\mu_1, \mu_2] \times [\sigma_1, \sigma_2]$ of K as small as possible. The first supremum is evaluated by using concentration inequalities for U-statistics and linear statistics again, while the second one is evaluated by using a Dvoretsky-Kiefer-Wolfowitz type exponential inequality in [BLM99], involving entropy with bracketing arguments.

1.3 Periodic signal detection

This work, in collaboration with Céline Lévy-Leduc [4], was initially motivated by an application in optronics, more precisely in laser vibrometry, where the model that has to be considered is the following Gaussian periodic fixed design regression model.

$\mathcal{M}^{(1)}_{\text{regression}}$	$\mathbf{X} = \mathbf{X}_n = (X_1, \ldots, X_n)'$ has a distribution P_f defined by: $X_i = f(i/n) + \sigma\varepsilon_i$ for i in $\{1, \ldots, n\}$, f being an unknown real valued periodic function (called the signal), the ε_i's being independent standard Gaussian random variables, and σ being a positive real number.

We are interested in the classical signal detection problem, that is the problem of testing

$$(H_0) \quad f = 0 \quad \text{against} \quad (H_1) \quad f \neq 0.$$

Whereas this problem had already been investigated from the general nonasymptotic minimax adaptivity point of view in [Bar02, BHL03], we here focus on the best possible use of the periodicity properties of the signal. Measuring accurately the impact of such properties is therefore the first purpose of the present work. The second purpose is to construct a minimax adaptive test, which takes the periodicity of the signal into account, just keeping in mind that in applications, the period of the signal is unknown, as well as the variance σ^2 of the Gaussian noise.

The interval $[0,1]$ is endowed with the measure μ_n given by $\mu_n = n^{-1}\sum_{j=1}^n \delta_{j/n}$, $\delta_{j/n}$ being the Dirac measure at j/n, and $\mathbb{L}_2([0,1], \mu_n)$ with its usual norm denoted by $\|.\|_2$ and given by

$$\forall g \in \mathbb{L}_2([0,1], \mu_n), \quad \|g\|_2^2 = \frac{1}{n}\sum_{i=1}^n g^2(i/n).$$

The corresponding metric is denoted by d_2.

1.3.1 Minimax separation rates over periodic Sobolev balls

Throughout this section, as a preliminary step, the variance σ^2 is assumed to be known.

Given α in $(0,1)$ and β in $(0, 1-\alpha)$, we investigate the minimax separation rates with prescribed error rates α and β for the metric d_2, over periodic Sobolev balls defined as follows. For k in $\{1, \ldots, n\}$, s in $\mathbb{N} \setminus \{0\}$ and $R > 0$, let:

$$\mathcal{S}_{s,2,k}(R) = \left\{ f \in \mathcal{C}^s([0,1]),\ f \text{ is periodic with period } k/n,\ \left\|f^{(s)}\right\|_{2,k/n} \leq R \right\}, \tag{1.12}$$

where $\mathcal{C}^s([0,1])$ denotes the set of functions f from $[0,1]$ with a continuous $s-$th order derivative denoted by $f^{(s)}$, and for θ in $(0,1)$, for g in $\mathbb{L}_2([0,1], \lambda)$, $\|g\|_{2,\theta}^2 = \theta^{-1}\int_0^\theta g^2 d\lambda$.

We first obtain the following lower bound.

Theorem 2 (Fromont, Lévy-Leduc, 2006)**.** *Given some fixed levels α in $(0,1)$ and β in $(0, 1-\alpha)$, there exist an absolute constant κ in $(0,1]$ and $C(s,\alpha,\beta)$ such that for every integer k in $[\kappa^{-1}, n/2]$,*

$$m\text{SR}_{d_2}^{\alpha,\beta}(\mathcal{S}_{s,2,k}(R)) \geq C(s,\alpha,\beta)\left(R^{\frac{1}{4s+1}}\left(\frac{k}{n}\right)^{\frac{s}{4s+1}}\left(\frac{\sigma^2}{n}\right)^{\frac{2s}{4s+1}} \wedge \left(\frac{\sqrt{k}\sigma^2}{n}\right)^{\frac{1}{2}} \wedge R\left(\frac{k}{n}\right)^s \right).$$

In particular, for all k in $[\kappa^{-1}, n/2]$, if

$$\sigma n^{-1/2}\left(\frac{k}{n}\right)^{-s} \leq R \leq \sigma n^{s-1/4}\left(\frac{k}{n}\right)^{1/4},$$

then

$$m\mathrm{SR}_{d_2}^{\alpha,\beta}\left(\mathcal{S}_{s,2,k}(R)\right) \geq C(s,\alpha,\beta)R^{\frac{1}{4s+1}}\left(\frac{k}{n}\right)^{\frac{s}{4s+1}}\left(\frac{\sigma^2}{n}\right)^{\frac{2s}{4s+1}}.$$

To establish the optimality of this lower bound, we construct a single test of (H_0) against (H_1), whose uniform separation rate over $\mathcal{S}_{s,2,k}(R)$ is of the same order as the lower bound, that is which is minimax over $\mathcal{S}_{s,2,k}(R)$.

For any function g defined on $[0,1]$, we set $\overline{g} = (g(1/n), g(2/n), \dots g(1))'$, and for any vector $v = (v_1, \dots, v_n)'$ in $\mathbb{R}^n$, we introduce

$$||v||_{\mathbb{R}^n}^2 = \frac{1}{n}\sum_{i=1}^{n} v_i^2.$$

Our minimax test is based on the estimation of $\left\|\Pi_{S_m}\left(\overline{f}\right)\right\|_{\mathbb{R}^n}$, where S_m is a subspace of $\mathbb{R}^n$ and Π_{S_m} is the orthogonal projection over S_m with respect to the Euclidean norm of $\mathbb{R}^n$ or with respect to $\|.\|_{\mathbb{R}^n}$. Considering the periodicity property of the signal f, a natural choice for S_m is a space based on the Fourier basis on the interval $[0,\theta]$ $(\theta > 0)$ defined by

$$\left\{\begin{array}{lcl} p_{\theta,0}(x) & = & 1, \\ p_{\theta,2l'-1}(x) & = & \sqrt{2}\cos\left(2\pi l' x/\theta\right), \forall l' \geq 1, \\ p_{\theta,2l'}(x) & = & \sqrt{2}\sin\left(2\pi l' x/\theta\right), \forall l' \geq 1. \end{array}\right. \tag{1.13}$$

For all D in $\mathbb{N}\backslash\{0\}$, we consider the subspace S_D of $\mathbb{R}^n$ spanned by the vectors $\left\{\overline{p_{k/n,l}},\ l = 0, \dots, D-1\right\}$, and we introduce

$$\phi_\alpha = \mathbb{1}_{\left\{n\|\Pi_{S_{D_k}}(Y)\|_{\mathbb{R}^n}^2 > F_{D_k}^{-1}(1-\alpha)\sigma^2\right\}},$$

where $D_k = k \wedge \inf\left\{D \in \mathbb{N},\ D \geq ((k/n)^{2s} n R^2/\sigma^2)^{2/(4s+1)}\right\}$, and $F_{D_k}^{-1}$ stands for the quantile function of the χ^2 distribution with D_k degrees of freedom. Then ϕ_α obviously satisfies $(\mathcal{P}_{\mathrm{level},\alpha})$ and if

$$\sigma n^{-1/2}\left(\frac{k}{n}\right)^{-s} \leq R \leq \sigma n^{\frac{8s^2-6s-1}{8s}}\left(\frac{k}{n}\right)^{-\frac{2s+1}{8s}}, \tag{1.14}$$

then

$$\mathrm{SR}_{d_2}^{\beta}\left(\phi_\alpha, \mathcal{S}_{s,2,k}(R)\right) \leq C(\alpha,\beta)R^{\frac{1}{4s+1}}\left(\frac{k}{n}\right)^{\frac{s}{4s+1}}\left(\frac{\sigma^2}{n}\right)^{\frac{2s}{4s+1}}.$$

Such a result is of course not useful at all in practice since the test constructed here clearly depends on s which is not known in applications. This test is thus only minimax over any given $\mathcal{S}_{s,2,k}(R)$, but not minimax adaptive. It is nevertheless important from a theoretical point of view, since assuming that k belongs to $[\kappa^{-1}, n/2]$ and that the radius R belongs to the range given by (1.14), it allows to conclude, combined with Theorem 2, that the minimax separation rate $m\mathrm{SR}_{d_2}^{\alpha,\beta}\left(\mathcal{S}_{s,2,k}(R)\right)$ is of order

$$R^{\frac{1}{4s+1}}\left(\frac{k}{n}\right)^{\frac{s}{4s+1}}\left(\frac{\sigma^2}{n}\right)^{\frac{2s}{4s+1}}.$$

Note that from an asymptotic point of view, where n tends to ∞ and k/n to a fixed period τ, since s is a positive integer, the range of (1.14) is not constraining: any positive R is allowed.

1.3.2 Minimax adaptive tests

From now on, we consider more realistic contexts where the variance σ^2 is not assumed to be known anymore, and where we aim at constructing minimax adaptive tests.
As the model $\mathcal{M}^{(1)}_{\text{regression}}$ is a particular case of the general regression model considered by Baraud, Huet and Laurent in [BHL03], we turn to the minimax adaptive aggregated test they propose.
Let $\{S_m, m \in \mathcal{M}\}$ be a collection of linear subspaces of $\mathbb{R}^n$, each S_m being of dimension D_m.
For every m in $\mathcal{M}$, and every u in $(0,1)$, we introduce the classical level u Fisher test of

$$(H_{0,m}) \quad \Pi_{S_m}\left(\overline{f}\right) = 0 \quad \text{against} \quad (H_{1,m}) \quad \Pi_{S_m}\left(\overline{f}\right) \neq 0,$$

defined by

$$\phi_{m,u} = \mathbb{1}_{\left\{\frac{(n-D_m)\|\Pi_{S_m}(\mathbf{X})\|^2_{\mathbb{R}^n}}{D_m\|\mathbf{X}-\Pi_{S_m}(\mathbf{X})\|^2_{\mathbb{R}^n}} > F^{-1}_{D_m,n-D_m}(1-u)\right\}},$$

where $F^{-1}_{D_m,n-D_m}$ is the Fisher quantile function with D_m and $n-D_m$ degrees of freedom.
From these single Fisher tests, given a fixed level α in $(0,1)$, as described in Section 1.1.2, Baraud, Huet and Laurent introduce the collection of tests Φ^{BHL}_α, and the corresponding aggregated test $\bar{\Phi}^{BHL}_\alpha$ defined by (1.4) and (1.1).

We here use this aggregated test by taking particular collections of linear subspaces $\{S_m, m \in \mathcal{M}\}$ fitting the periodicity assumption on the signal.

Periodic signal detection when the period is known. Let us assume that $n \geq 3$ and that f is periodic with period k/n for k in $\{3,\dots,n\}$. We consider the above aggregated test $\bar{\Phi}^{BHL}_\alpha$ with:

- $\mathcal{M} = \left\{2^J,\ J \in \{0,\dots,\lfloor \log_2(k/2)\rfloor\}\right\}$,
- $\{S_m,\ m \in \mathcal{M}\}$, where for $m = D$ in $\mathcal{M}$, S_D is still the linear subspace of $\mathbb{R}^n$ spanned by the vectors $\left\{\overline{p_{k/n,l}},\ l \in \{0,\dots,D-1\}\right\}$.

Proposition 1 (Fromont, Lévy-Leduc, 2006)**.** *For α and β in $(0,1)$, assume that n is large enough so that $\alpha \geq e^{-n/20}\log_2(n)$ and $\beta \geq 2e^{-n/42}$. For all s in $\mathbb{N}\setminus\{0\}$, there exists $C(s,\alpha,\beta)$ such that for all $R > 0$ satisfying*

$$\sigma n^{-1/2}\left(\frac{k}{n}\right)^{-s}(\ln\ln k)^{s+1/2} \leq R \leq \sigma n^{\frac{8s^2-6s-1}{8s}}\left(\frac{k}{n}\right)^{-\frac{2s+1}{8s}}(\ln\ln k)^{1/4}, \tag{1.15}$$

then

$$\mathrm{SR}^{\beta}_{d_2}\left(\bar{\Phi}^{BHL}_\alpha, \mathcal{S}_{s,2,k}(R)\right) \leq C(s,\alpha,\beta)R^{\frac{1}{4s+1}}\left(\frac{k}{n}\right)^{\frac{s}{4s+1}}\left(\frac{\sqrt{\ln\ln k}\sigma^2}{n}\right)^{\frac{2s}{4s+1}}.$$

According to the above results, this means that the testing procedure is rate optimal, up to a possible $(\ln\ln k)^{1/2}$ factor, over all the Sobolev balls $\mathcal{S}_{s,2,k}(R)$ such that (1.15) holds simultaneously. In view of the results due to Spokoiny [Spo96] in the Gaussian white noise model for Besov balls and Baraud [Bar02] in the Gaussian sequence model for families of nested ellipsoids, we can rightfully think that this loss of efficiency is unavoidable when we deal with an adaptive procedure. In this sense, one can say that considering

$$\overline{\mathcal{F}}_1 = \left\{\mathcal{S}_{s,2,k}(R),\ s > 0,\ R > 0 \text{ satisfying (1.15)}\right\},$$

in the notation and under conditions of Proposition 1, $\bar{\Phi}^{BHL}_\alpha$ satisfies $\left(\mathcal{P}_{\text{adaptive},\alpha,\beta,\overline{\mathcal{F}_1},d_2}\right)$.

Periodic signal detection when the period is unknown. We now consider the most general setup where the period of the signal is unknown. We consider the above aggregated test $\bar{\Phi}_\alpha^{BHL}$ again, but this time with:

— $\mathcal{M} = \big\{(k, 2^J), k \in \{2, \dots, n\}, J \in \{0, \dots, \lfloor \log_2(k/2) \rfloor\}\big\}$,
— $\{\, S_m,\ m \in \mathcal{M}\,\}$, where for $m = (k, D)$ in $\mathcal{M}$, $S_{(k,D)}$ is the linear subspace of $\mathbb{R}^n$ spanned by the vectors $\big\{\, \overline{p_{k/n,l}},\ l \in \{\,0, \dots, D-1\,\}\,\big\}$.

To avoid any confusion, the test is here denoted by $\bar{\tilde{\Phi}}_\alpha^{BHL}$. The following result gives upper bounds for the uniform separation rates of this test over some Sobolev balls described below.
Assume that $n \geq 3$. For $s \in \mathbb{N} \setminus \{\,0\,\}$, $R > 0$ and $\tau_1 \in [2/n, 1]$, let

$$\tilde{\mathcal{S}}_{s,2,\tau_1}(R) = \Big\{ f \in \mathcal{C}^s([0,1]),\ f \text{ is periodic with period } \tau(f) \in [2/n, \tau_1],\ \big\|f^{(s)}\big\|_{2,\tau(f)} \leq R \Big\}.$$

Proposition 2 (Fromont, Lévy-Leduc, 2006)**.** *For α and β in $(0,1)$, assume that n is large enough so that $\alpha \geq e^{-n/20} \log_2(n!)$ and $\beta \geq 2e^{-n/42}$. For all s in $\mathbb{N} \setminus \{\,0\,\}$, there exists $C(s, \alpha, \beta)$ such that for all τ_1 in $[2/n, 1]$, and $R > 0$ satisfying*

$$\sigma \tau_1^{-s} (\ln n)^{s+1/2} n^{-1/2} \leq R \leq \sigma \tau_1^{\frac{-4s^2+4s+1}{4s}} n^{1/8s} (\ln n)^{-\frac{2s+1}{8s}}, \tag{1.16}$$

then

$$\mathrm{SR}_{d_2}^{\beta}\left(\bar{\tilde{\Phi}}_\alpha^{BHL}, \tilde{\mathcal{S}}_{s,2,\tau_1}(R)\right) \leq C(s,\alpha,\beta) R^{\frac{1}{4s+1}} \tau_1^{\frac{s}{4s+1}} \left(\frac{\sqrt{\ln n}\sigma^2}{n}\right)^{\frac{2s}{4s+1}}.$$

The upper bound for the uniform separation rate of the test over $\tilde{\mathcal{S}}_{s,2,\tau_1}(R)$ when τ_1 and R satisfy (1.16) is similar to the one obtained when the period of f is known to be equal to k/n (see Proposition 1), but with a loss of efficiency of the order of a $(\ln n)^{1/2}$ factor instead of $(\ln \ln k)^{1/2}$. This is technically due to the fact that we choose a collection $\{S_m,\ m \in \mathcal{M}\}$ which is rich enough to ensure that it contains, for any function f with period $\tau(f)$ in $[2/n, 1]$, some subspace S_m close enough to f. We do not know if the consequent extra $(\ln n)^{1/2}$ factor is unavoidable, as explained in Section 1.2 where the aggregated test proposed for the problem of testing that f belongs to a translation/scale parametric family shows a similar behavior.

1.3.3 Links with model selection

The aggregated test proposed by Baraud, Huet and Laurent [BHL03] is closely related to the estimation of quadratic functionals, such as $\|f\|_2^2$, by Laurent and Massart [LM00].
Considering a collection of linear subspaces $\{\, S_m,\ m \in \mathcal{M}\,\}$ of $\mathbb{R}^n$, Laurent and Massart propose to estimate $\theta(f) = \|f\|_2^2$ by

$$\hat{\theta} = \sup_{m \in \mathcal{M}} \Big(\|\Pi_{S_m}(\mathbf{X})\|_{R^n}^2 - \mathrm{pen}(m) \Big),$$

where $\mathrm{pen}(m)$ is a well-chosen deterministic penalty term, leading to oracle inequalities and minimax adaptivity properties for $\hat{\theta}$.
If the variance σ^2 is equal to 1, when $f = 0$, one furthermore has that $\mathbb{E}\big[\hat{\theta}^2\big] \leq C/n^2$ for some $C > 0$. A level α aggregated test of $(\,H_0\,)$ against $(\,H_1\,)$ can thus be obtained with

$$\mathbb{1}_{\{\hat{\theta} > \sqrt{C}/(n\sqrt{\alpha})\}} = \sup_{m \in \mathcal{M}} \mathbb{1}_{\{\|\Pi_{S_m}(\mathbf{X})\|_{R^n}^2 > \mathrm{pen}(m) + \sqrt{C}/(n\sqrt{\alpha})\}}.$$

Since the penalty $\mathrm{pen}(m)$ depends on a constant that has to be calibrated in practice in the estimation problem, it is, in the present testing context, more appropriate to directly calibrate some individual critical values $c_{m,\alpha}$'s so that the aggregated test $\sup_{m \in \mathcal{M}} \mathbb{1}_{\{\|\Pi_{S_m}(\mathbf{X})\|_{R^n}^2 > c_{m,\alpha}\}}$, is exactly of level α.

When $f = 0$, the distribution of $\|\Pi_{S_m}(\mathbf{X})\|^2_{R^n}$ is here a simple chi-square distribution, so calibrating the critical values becomes rather easy. This idea is then generalized to the case where the variance σ^2 is unknown and thus has to be estimated, at the price that $\|\Pi_{S_m}(\mathbf{X})\|^2_{R^n}$ is replaced by the Fisher distributed variable $(n - D_m)\|\Pi_{S_m}(\mathbf{X})\|^2_{\mathbb{R}^n} / \big(D_m \|\mathbf{X} - \Pi_{S_m}(\mathbf{X})\|^2_{\mathbb{R}^n}\big)$.

1.3.4 Experimental results

Simulation study. The simulation study that we presented in [4] was specifically dedicated to the evaluation, from a practical point of view, of the performances of the proposed test when the period of the signal is unknown. In particular, we aimed at highlighting the sensitivity of the test to the periodicity assumption. To do this, we compared the estimated sizes and powers under various alternatives of the present test with the ones of the test initially proposed by Baraud, Huet and Laurent [BHL03], which does not take the periodicity assumption into account.
As expected from the theoretical study, our test show significant improvements as compared to the test of [BHL03], and this is particularly striking when the number of significant Fourier coefficients (or harmonics) in the expansion of the periodic signal alternative is large.

Application in laser vibrometry. This study was motivated by a real practical issue from the Thalès Optronique company, which partially supported it.
Consider targets vibrating under the effect of the vibrations of their motor for instance. One of the most topical issues in optronics is the identification of such a target, through the determination of some of its vibration parameters, like its vibration period. The use of coherent lasers has provided some progress in this field. After emission of a continuous coherent laser wave, reflection of it on a target composed of reflectors vibrating at the same frequency F_s (that is $1/F_s$ is the vibration period), reception and demodulation, one receives a complex valued signal of the form:

$$X_j = f(j/n) + \sigma(\varepsilon_{1,j} + i\,\varepsilon_{2,j}) \text{ for } j \in \{1, \dots, n\}. \tag{1.17}$$

Here, $i^2 = -1$, the $\varepsilon_{1,j}$'s and the $\varepsilon_{2,j}$'s are independent standard Gaussian random variables and f is of the form

$$f(x) = \sum_{m=1}^{M} a_m \exp\left[\frac{4i\pi\gamma_m}{\lambda}\cos\left(2\pi F_s x\right)\right], \tag{1.18}$$

when the target consists of M (which may be large: $M \approx 200$) punctual reflectors, a_m being the amplitude of the signal reflected by the reflector number m.
The issue of the estimation of the frequency F_s has been treated by Céline Lévy-Leduc and co-authors (see for instance [CLLM06] and references therein).

Our concern in the present work is the target detection step, which has to precede the estimation phase. Of course, at this step, the frequency F_s is unknown, and this is why we developed the above last periodic signal detection tests, that are easily adapted to the present complex valued Gaussian regression model.

An example of synthetic signal arising in laser vibrometry is displayed in [4]. In particular, one can see that for a corrupted signal with a signal to noise ratio $\text{SNR} = 10\ln\big(\sum_{m=1}^{M} a_m^2/2\sigma^2\big)$ equal to -27dB, the high level of noise makes the original signal and its harmonics visually undetectable.

We consider an alternative which corresponds to the signal (1.18) where $M = 1$, $F_s = 48$Hz, $\gamma_1 = 35 \times 10^{-6}$, $\lambda = 1.5 \times 10^{-6}$, with $\sigma^2 = 1$, and a number of observations $n = 2^{18}$. For $\alpha = 0.05$, estimated powers are presented for various SNR in Table 2 of [4]. It is shown that for an SNR $= -27$ dB, our test has an estimated power equal to 0.67, and for an SNR $= -24$ dB, which is still reasonable for the present application, it has an estimated power equal to 0.99. As a comparison, in this last case, the test of Brockwell and Davis [BD13], which is a classical test for periodic signal detection, has an estimated power of 0.06.

1.3.5 Sketch of proof

We here draw a sketch of proof for the lower bound in Theorem 2. For sake of simplicity in the notations, the dependencies on α, β, k and σ (which are fixed here) are not always specified.
The idea of the proof is rather general and can be summarized as constructing, for the Sobolev ball $\mathcal{S}_{s,2,k}(R)$, a class of alternatives $\Omega(r^*) \subset \{ f,\ d_2(f,0) = \|f\|_2 = r^* \}$, with r^* as large as possible (so that the final lower bound is more likely to be sharp), included in $\mathcal{S}_{s,2,k}(R)$, but for which

$$\inf_{\phi_\alpha \text{ satisfying } \left(\mathcal{P}_{\text{level},\alpha}\right)} \sup_{f \in \Omega(r^*)} P_f\left(\phi_\alpha = 0\right) \geq \beta.$$

1. The first step consists in introducing a collection of classes of alternatives $\Omega_D(r)$ for D in some $\mathcal{D}_k \subset \{1, \dots, k\}$, $r > 0$, given by

$$\Omega_D(r) = \left\{ f,\ f(x) = \sum_{l=1}^{D} \beta_l \varphi_D\left(\frac{nDx}{k} - l + 1\right),\ \beta_l \in \mathbb{R},\ \|f\|_2 = r \right\},$$

   where φ_D is a D-periodic function in $\mathcal{C}^\infty(\mathbb{R})$, positive on its support which is included in $\cup_{u \in \mathbb{Z}}(uD, uD+1)$.

2. The second and main step then consists in determining, for all D in $\mathcal{D}_k$, a maximal radius r_D such that for all $r \leq r_D$, then

$$\inf_{\phi_\alpha \text{ satisfying } \left(\mathcal{P}_{\text{level},\alpha}\right)} \sup_{f \in \Omega_D(r)} P_f\left(\phi_\alpha = 0\right) \geq \beta.$$

   This step is based on Bayesian techniques developed by Ingster [Ing93]: the idea is to find a probability measure μ_r on $\Omega_D(r)$ (the prior) such that $r \leq r_D$ implies that

$$\mathbb{E}_0\left[L^2_{\mu_r}(\mathbf{X})\right] \leq 1 + 4(1-\alpha-\beta)^2,$$

   where:
   — $P_{\mu_r} = \int P_f d\mu_r(f)$,
   — $L_{\mu_r}(x) = \frac{dP_{\mu_r}}{dP_0}(x)$,
   — $\mathbb{E}_0$ is the expectation with respect to P_0.
   We then use the following inequality (see [Bar02] for instance):

$$\inf_{\phi_\alpha \text{ satisfying } \left(\mathcal{P}_{\text{level},\alpha}\right)} \sup_{f \in \Omega_D(r)} P_f\left(\phi_\alpha = 0\right) \geq 1 - \alpha - \frac{1}{2}\left(\mathbb{E}_0\left[L^2_{\mu_r}(\mathbf{X})\right] - 1\right)^{1/2} \geq \beta.$$

   To construct such a prior distribution μ_r, we consider $f_\xi(x) = \lambda \sum_{l=1}^{D} \xi_l \varphi_D\left(nDx/k - l + 1\right)$ where $\xi = (\xi_1, \dots, \xi_D)$ is a sequence of i.i.d. Rademacher random variables (that is random variables taking the values 1 or -1 with probability $1/2$), and λ is chosen so that $||f_\xi||_2 = r$.

3. The third and final step is to determine for all D in $\mathcal{D}_k$ the maximal radius $r_{D,s,R} \leq r_D$ so that $\Omega_D(r_{D,s,R})$ is included in $\mathcal{S}_{s,2,k}(R)$, and to take $\Omega(r^*) = \Omega_{D^*}(r_{D^*,s,R})$ with

$$D^* = \text{argmax}_{D \in \mathcal{D}_k} r_{D,s,R}.$$

1.4 Tests of homogeneity for Poisson processes

As explained in the nice introduction of the book of Daley and Vere-Jones, historically the theory of point processes seems to go back to the study of the first life tables and renewal processes, and of counting problems with the work of Poisson [Poi37]. Since recently, point processes are largely encountered in the genetics and neuroscience literature. At the origin of this work, we were actually interested in some applications in genetics, where a point process of the real line may represent occurrences of words or motifs on the DNA sequence (see [RRS05] for instance). In this context, it is particularly important to be able to detect abnormal behaviors.
Assuming that the point process is a Poisson process on a bounded interval, which has been commonly admitted as a realistic model for occurrences of motifs on the DNA sequence, such an abnormal behaviors detection question may be viewed as a problem of testing the homogeneity of the Poisson process. More precisely, let us consider the following Poisson process model:

$\mathcal{M}^{(1)}_{\text{Poisson}}$	$\mathbf{X}$ is a (possibly inhomogeneous) Poisson process observed on the interval $\mathbb{X} = [0,1]$, with intensity f, with respect to some measure μ on $[0,1]$ such that $d\mu = nd\lambda$ for a fixed positive integer n.

In this model, we focus on the problem of testing

$$(H_0) \quad f \in \mathcal{F}_0 \quad \text{against} \quad (H_1) \quad f \notin \mathcal{F}_0,$$

where $\mathcal{F}_0$ is the set of constant functions on $\mathbb{X}$.
Notice that the present problem of testing is of course not dedicated to the only field of DNA sequences analysis. This study can also be used in all the other frameworks where the Poisson process model is realistic (and these frameworks are numerous, going from economics and finance to sport or music analysis for instance).

Let $\mathcal{X}$ be the set of the possible values for finite point processes on $\mathbb{X}$, that is the set of the countable subsets of $\mathbb{X}$. For all x in $\mathcal{X}$, all t in $\mathbb{R}$, let us define

$$N_x(t) = \int_{\mathbb{X}} \mathbb{1}_{\{u \leq t\}} dN_x(u), \tag{1.19}$$

where dN_x stands for the point measure associated with x, given for all measurable real-valued function g by

$$\int_{\mathbb{X}} g(u) dN_x(u) = \sum_{u \in x} g(u). \tag{1.20}$$

With such notation, remark that $\#\mathbf{X} = N_{\mathbf{X}}(1)$.

The model $\mathcal{M}^{(1)}_{\text{Poisson}}$ has in fact close links with the density model $\mathcal{M}^{(1)}_{\text{density}}$, and can even be defined via these links: $\mathbf{X}$ may be introduced as a set of random variables $X_1, \dots, X_{N_{\mathbf{X}}(1)}$ observed in $[0,1]$, such that

- the random variable $N_{\mathbf{X}}(1)$ has a Poisson distribution with parameter $\int_{\mathbb{X}} f d\mu = n \int_{\mathbb{X}} f d\lambda$,
- given $N_{\mathbf{X}}(1) = n_0$, $(X_1, \dots, X_{N_{\mathbf{X}}(1)})$ is conditionally distributed as a sample of n_0 i.i.d. random variables, with density $f / \int_{\mathbb{X}} f d\lambda$ with respect to the Lebesgue measure λ.

Given $N_{\mathbf{X}}(1) = n_0$, testing (H_0) against (H_1) therefore amounts to testing that $f = \mathbb{1}_{[0,1]}$ that is testing the uniformity on $[0,1]$ of the X_i's, based on the observation of $(X_1, \dots, X_{n_0})$. The minimax adaptive tests introduced in Section 1.2.1 can thus be used in practice, as well as every test of uniformity on $[0,1]$ in a density model, like the historical one of Kolmogorov-Smirnov for instance. However, studying such tests from the minimax point of view raises the difficulty of deconditioning the event

$N_{\mathbf{X}}(1) = n_0$, and separation rates are not so directly obtained. We thus turned towards more natural unconditional tests, considering the Poisson process by itself and using its convenient properties.

The present problem of testing the homogeneity of a Poisson process has been widely investigated both from theoretical and practical points of view (see Bain, Engelhardt, and Wright [BEW85] or Cohen and Sackrowitz [CS93] for a survey and Bhattacharjee, Deshpande, and Naik-Nimbalkar [BDNN04] for a more recent work). In these papers, the alternative intensities are monotonous. Another related topic is the problem of testing the hypothesis that a point process is a Poisson process with a given intensity. We can cite for instance the papers by Fazli and Kutoyants [FK05] where the alternative is also a Poisson process with a known intensity, Fazli [Faz07] where the alternatives are Poisson processes with one-sided parametric intensities, or Dachian and Kutoyants [DK06], where the alternatives are self-exciting point processes. The paper by Ingster and Kutoyants [IK07] is the closest one to the present work. The alternatives considered by Ingster and Kutoyants are Poisson processes with nonparametric intensities in a Sobolev or Besov space $\mathcal{B}_{s,2,q}(R)$ with $1 \leq q < +\infty$ and known smoothness parameter s: the minimax rate of testing with respect to the $\mathbb{L}_2$-metric over such a Sobolev or Besov ball is proved to be of order $n^{-2s/(4s+1)}$, which matches the classical minimax rate of testing in the density model.

However, in certain practical cases, such smooth alternatives cannot be considered. For instance, admitting that the Poisson process represents occurrences of motifs on the DNA sequence, its intensity may burst at a particular position of special interest for the biologist (see [GS05] for more details). The present problem thus deals with the question: "how can we distinguish a Poisson process with constant intensity from a Poisson process whose intensity has some small localized spikes?". This question had already been partially considered in a precursory work by Watson [Wat78], but without any precise study of the second kind error rate of the tests.

So, based on these reflexions, assuming that f belongs to $\mathbb{L}_2(\mathbb{X}, \lambda)$, we constructed in [7] new minimax adaptive tests of (H_0) against (H_1), considering classes of smooth alternatives such as Besov bodies, but also some classes of alternatives that may be much less smooth, such as weak Besov bodies that were, until this work, only studied in some estimation contexts (see [Riv02, Riv06]).

Since the minimax separation rates over such weak Besov bodies were not known, our purpose was first to provide lower bounds for these minimax separation rates, and then to construct level α minimax adaptive tests, based on the aggregation principle. A phenomenon, completely new with respect to the anterior minimax testing literature, then appeared: the minimax separation rates over particular subsets of weak Besov bodies are so large that there is no additional price to pay for adaptivity on the one hand (but such a phenomenon already occurs when the $\mathbb{L}_2$-metric is replaced by the $\mathbb{L}_\infty$-one), and that they are of the same order as the minimax estimation rates on the other hand.

1.4.1 Lower bounds for the minimax separation rates over Besov bodies

Let $\|.\|_2$, d_2 and $\langle ., . \rangle_2$ respectively denote the classical $\mathbb{L}_2$ norm, metric and scalar product of $\mathbb{L}_2(\mathbb{X}, \lambda)$. Let us introduce the Haar basis of $\mathbb{L}_2([0,1], \lambda)$, $\{\varphi_0,\ \psi_{(j,k)},\ j \in \mathbb{N},\ k \in \{0, \dots, 2^j - 1\}\}$ with $\varphi_0(x) = \mathbb{1}_{[0,1]}(x)$, and

$$\psi_{(j,k)}(x) = 2^{j/2}\psi(2^j x - k), \tag{1.21}$$

where $\psi(x) = \mathbb{1}_{[0,1/2)}(x) - \mathbb{1}_{[1/2,1)}(x)$.

We introduce the Besov bodies defined for $s > 0$, $R > 0$ by

$$\mathcal{B}_{s,2,\infty}(R) = \left\{ f = \alpha_0\varphi_0 + \sum_{j \in \mathbb{N}} \sum_{k=0}^{2^j-1} \alpha_{(j,k)}\psi_{(j,k)} \geq 0,\ \forall j \in \mathbb{N},\ \sum_{k=0}^{2^j-1} \alpha_{(j,k)}^2 \leq R^2 2^{-2js} \right\}. \tag{1.22}$$

The weak Besov bodies are defined for $s' > 0$ and $R' > 0$ by

$$w\mathcal{B}_{s'}(R') = \left\{ f = \alpha_0\varphi_0 + \sum_{j\in\mathbb{N}}\sum_{k=0}^{2^j-1} \alpha_{(j,k)}\psi_{(j,k)} \geq 0,\ \forall t > 0,\ \sum_{j\in\mathbb{N}}\sum_{k=0}^{2^j-1} \alpha_{(j,k)}^2 \mathbb{1}_{\left\{\alpha_{(j,k)}^2 \leq t\right\}} \leq R'^2 t^{\frac{2s'}{2s'+1}} \right\}. \tag{1.23}$$

Theorem 3 (Fromont, Laurent, Reynaud-Bouret, 2011)**.** *Assume that $R > 0$, $R' > 0$, and $R'' \geq 2$, and fix some levels α and β in $(0,1)$ such that $\alpha + \beta \leq 0.59$.*
(i) If $s \geq (s'/2) \vee (s'/(2s'+1))$, then

$$\liminf_{n\to+\infty} n^{\frac{2s}{4s+1}} m\mathrm{SR}_{d_2}^{\alpha,\beta}(\mathcal{B}_{s,2,\infty}(R) \cap w\mathcal{B}_{s'}(R') \cap \mathbb{L}_\infty(R'')) > 0.$$

(ii) If $s < s'/2$ and $s' > 1/2$, then

$$\liminf_{n\to+\infty} \left(\frac{n}{\ln n}\right)^{\frac{s'}{2s'+1}} m\mathrm{SR}_{d_2}^{\alpha,\beta}(\mathcal{B}_{s,2,\infty}(R) \cap w\mathcal{B}_{s'}(R') \cap \mathbb{L}_\infty(R'')) > 0.$$

(iii) If $s < s'/(2s'+1)$ and $s' \leq 1/2$, then

$$\liminf_{n\to+\infty} \left(n^{\frac{1}{4}} \wedge n^{\frac{2s'}{(4s+1)(2s'+1)}}\right) m\mathrm{SR}_{d_2}^{\alpha,\beta}(\mathcal{B}_{s,2,\infty}(R) \cap w\mathcal{B}_{s'}(R') \cap \mathbb{L}_\infty(R'')) > 0.$$

The lower bounds in (i) coincide with the minimax separation rates over Besov spaces, obtained by Ingster and Kutoyants [IK07] in a slightly different Poisson process model, and therefore also match with the rates in the more classical density model. As seen in Section 1.2, such rates can be achieved in the density model by some aggregated tests, closely linked with model selection. This is the principle of our first aggregated test. One now should focus on the lower bounds obtained in (ii), which are in fact equal to the minimax estimation rates on the maxisets of the thresholding estimation procedure, namely $\mathcal{B}_{\kappa s'/(2s'+1),2,\infty}(R) \cap w\mathcal{B}_{s'}(R')$ with some constant $\kappa < 1$ (see [KP00, Riv06, RBR10]). This means that it is at least as difficult to test as to estimate over such classes of functions. Following the idea that the minimax estimation rates on these classes are achieved by thresholding rules, the principle of our second aggregated test will be based on thresholding methods.

1.4.2 Minimax adaptive tests

Let us consider a collection of subspaces $\{S_m\ m \in \mathcal{M}\}$ of $\mathbb{L}_2([0,1],\lambda)$ such that for m in $\mathcal{M}$, S_m is spanned by $\{\varphi_0,\ \psi_{j,k},\ (j,k) \in \mathcal{L}_m\}$, where $\mathcal{L}_m$ is a subset of $\left\{(j,k),\ j \in \mathbb{N},\ k \in \left\{0,\dots,2^j-1\right\}\right\}$. Following the aggregation principle described in Section 1.1.2, we construct a collection of tests of

$$(H_{0,m}) \quad d_2\left(\Pi_{S_m}(f), \mathcal{F}_0\right) = 0 \quad \text{against} \quad (H_{1,m}) \quad d_2\left(\Pi_{S_m}(f), \mathcal{F}_0\right) \neq 0,$$

for m in $\mathcal{M}$, where Π_{S_m} denotes the orthogonal projection with respect to $\langle .,. \rangle_2$ onto S_m.
Noticing that for each m in $\mathcal{M}$, $d_2^2\left(\Pi_{S_m}(f), \mathcal{F}_0\right) = \sum_{(j,k)\in\mathcal{L}_m} \langle f, \psi_{j,k}\rangle_2^2$, a reasonable test statistic for $(H_{0,m})$ against $(H_{1,m})$ is given by the unbiased estimator of $d_2^2\left(\Pi_{S_m}(f), \mathcal{F}_0\right)$:

$$T_m = \frac{1}{n^2} \sum_{(j,k)\in\mathcal{L}_m} \sum_{i\neq i'=1}^{N_{\mathbf{X}}(1)} \psi_{(j,k)}(X_i)\psi_{(j,k)}(X_{i'}). \tag{1.24}$$

The construction of the critical value corresponding to each T_m depends on the chosen collection $\{S_m,\ m \in \mathcal{M}\}$. In particular, two main different choices are made for the considered collection of subspaces, leading to two different methods for the construction of the critical values.

As the distribution of T_m is not free from f under (H_0), both methods are nevertheless based on an instrumental conditional distribution, as described in Section 1.1.2, using the statistic $Z = N_{\mathbf{X}}(1)$. More precisely, both methods are based on the argument that under (H_0), for any n_0 in $\mathbb{N}\setminus\{0\}$, given $N_{\mathbf{X}}(1) = n_0$, T_m is conditionally distributed as

$$\tilde{T}_m^{(n_0)} = \frac{1}{n^2} \sum_{(j,k)\in\mathcal{L}_m} \sum_{i\neq i'=1}^{n_0} \psi_{(j,k)}\left(\tilde{X}_i\right) \psi_{(j,k)}\left(\tilde{X}_{i'}\right),$$

where $\left(\tilde{X}_1, \ldots, \tilde{X}_{n_0}\right)$ is an i.i.d. sample from the uniform distribution on $[0,1]$.
Therefore, denoting by $q_m^{(n_0)}$ the quantile function of the distribution of $\tilde{T}_m^{(n_0)}$, the critical value associated with T_m is chosen as

$$q_m^{(N_{\mathbf{X}}(1))}\left(1 - u_{m,\alpha}^{(N_{\mathbf{X}}(1))}\right),$$

where $u_{m,\alpha}^{(N_{\mathbf{X}}(1))}$ has to be correctly calibrated.

Collection of nested spaces. Let $\mathcal{M}$ be a subset of $\mathbb{N}\setminus\{0\}$, and for $m = J$ in $\mathcal{M}$, $\mathcal{L}_J = \big\{(j,k),\ j \in \{0,\ldots,J-1\},\ k \in \{0,\ldots,2^j-1\}\big\}$, so that S_J is spanned by $\big\{\varphi_0,\ \psi_{j,k},\ j \in \{0,\ldots,J-1\},\ k \in \{0,\ldots,2^j-1\}\big\}$.
Then, we consider the test $\bar{\Phi}_\alpha^{nested}$ aggregating as in (1.1) the tests

$$\mathbb{1}_{\left\{T_m > q_m^{(N_{\mathbf{X}}(1))}\left(1-u_{m,\alpha}^{(N_{\mathbf{X}}(1))}\right)\right\}}, \tag{1.25}$$

where

$$u_{m,\alpha}^{(n_0)} = w_m \sup\left\{u,\ \mathbb{P}_{(H_0)}\left(\exists m \in \mathcal{M},\ T_m > q_m^{(n_0)}(1-w_m u)\middle| N_{\mathbf{X}}(1) = n_0\right) \le \alpha\right\}. \tag{1.26}$$

This test satisfies $(\mathcal{P}_{\text{level},\alpha})$, and in [7], it is proved to satisfy an oracle type result. Since any function f in $\mathcal{B}_{s,2,\infty}(R)$ is well approximated by its projections onto the chosen nested subspaces, in the sense that $\|f - \Pi_{S_J}(f)\|_2^2 \le C(s)R^2 2^{-2Js}$, this oracle type result leads to the following theorem.

Theorem 4 (Fromont, Laurent, Reynaud-Bouret, 2011)**.** *Given α and β in $(0,1)$, let $\bar{\Phi}_\alpha^{nested}$ be the test defined above with $\mathcal{M} = \big\{1,\ldots,\big\lfloor \log_2(n^2/(\ln\ln n)^3)\big\rfloor\big\}$ and $w_m = 1/\#\mathcal{M}$ for every m in $\mathcal{M}$. For every $s > 0$, there exist $C(s)$ and $C(\alpha,\beta,R,R'',s)$ such that if $n > C(s)$, then*

$$\mathrm{SR}_{d_2}^{\beta}\left(\bar{\Phi}_\alpha^{nested}, \mathcal{B}_{s,2,\infty}(R)\cap\mathbb{L}_\infty(R'')\right) \le C(\alpha,\beta,R,R'',s)\left(\frac{\sqrt{\ln\ln n}}{n}\right)^{\frac{2s}{4s+1}}.$$

Note that a more precise bound is given in [7], in particular giving the precise dependence with respect to the radius R of the Besov body $\mathcal{B}_{s,2,\infty}(R)$.

Collection of nonnested two-dimensional spaces. Let $\mathcal{M}$ be now defined from a fixed integer $\bar{J} \ge 1$ by $\mathcal{M} = \big\{(j,k),\ j \in \{0,\ldots,\bar{J}-1\},\ k \in \{0,\ldots,2^j-1\}\big\}$, and for $m = (j,k)$ in $\mathcal{M}$, $\mathcal{L}_{(j,k)} = \{(j,k)\}$, so that $S_{(j,k)}$ is spanned by $\{\varphi_0,\ \psi_{j,k}\}$. Each individual problem of testing $(H_{0,m})$ against $(H_{1,m})$ then consists in detecting a nonnull coefficient $\langle f, \psi_{j,k}\rangle_2$ in the expansion of f with respect to the Haar basis.
Then, we consider the test $\bar{\Phi}_\alpha^{nonnested}$ aggregating as in (1.1) the tests

$$\mathbb{1}_{\left\{T_{(j,k)} > q_{(j,k)}^{(N_{\mathbf{X}}(1))}\left(1-u_{(j,k),\alpha}^{(N_{\mathbf{X}}(1))}\right)\right\}}, \tag{1.27}$$

where

$$u^{(n_0)}_{(j,k),\alpha} = (2^j\bar{J})^{-1} \sup\Big\{u,\ \mathbb{P}_{(H_0)}\Big(\exists (j,k)\in\mathcal{M},\ T_{(j,k)} > q^{(n_0)}_{(j,k)}\left(1-(2^j\bar{J})^{-1}u\right)\Big| N_{\mathbf{X}}(1)=n_0\Big)\le\alpha\Big\}. \tag{1.28}$$

This test also satisfies $(\mathcal{P}_{\text{level},\alpha})$ and, from the oracle type result obtained for $\bar{\Phi}^{nested}_{\alpha}$, the following upper bound is derived.

Theorem 5 (Fromont, Laurent, Reynaud-Bouret, 2011)**.** *Given some levels α and β in $(0,1)$, let $\bar{\Phi}^{nonnested}_{\alpha}$ be the aggregated test defined above with $\bar{J} = \lfloor \log_2(n/\ln n)\rfloor$. For every $s>0$ and $s'>0$, such that $s \ge s'/(2s'+1)$, there exist some positive constants $C(s,s')$ and $C(\alpha,\beta,R,R',R'',s,s')$ such that if $n > C(s,s')$, then*

$$\mathrm{SR}^{\beta}_{d_2}\left(\bar{\Phi}^{nonnested}_{\alpha}, \mathcal{B}_{s,2,\infty}(R)\cap w\mathcal{B}_{s'}(R')\cap\mathbb{L}_{\infty}(R'')\right) \le C(\alpha,\beta,R,R',R'',s,s')\left(\frac{\ln n}{n}\right)^{\frac{s'}{2s'+1}}.$$

In view of the above results, considering

$$\overline{\mathcal{F}_1}^{(i)} = \left\{\mathcal{B}_{s,2,\infty}(R)\cap w\mathcal{B}_{s'}(R')\cap\mathbb{L}_{\infty}(R''),\ s\ge (s'/2)\vee(s'/(2s'+1))\right\},$$

for n large enough, the test $\bar{\Phi}^{nested}_{\alpha}$ of Theorem 4 satisfies $\left(\mathcal{P}_{\text{adaptive},\alpha,\beta,\overline{\mathcal{F}_1}^{(i)},d_2}\right)$, with a price for adaptivity of the order of a $(\ln\ln n)^{1/2}$ factor. Now consider

$$\overline{\mathcal{F}_1}^{(ii)'} = \left\{\mathcal{B}_{s,2,\infty}(R)\cap w\mathcal{B}_{s'}(R')\cap\mathbb{L}_{\infty}(R''),\ s'/(2s'+1)\le s< s'/2\right\}\subset\overline{\mathcal{F}_1}^{(ii)},$$

where $\overline{\mathcal{F}_1}^{(ii)}$ is the collection corresponding to the parameter set of the case (ii) in Theorem 3, that is

$$\overline{\mathcal{F}_1}^{(ii)} = \left\{\mathcal{B}_{s,2,\infty}(R)\cap w\mathcal{B}_{s'}(R')\cap\mathbb{L}_{\infty}(R''),\ s<s'/2,\ s'>1/2\right\}.$$

For n large enough, the test $\bar{\Phi}^{nonnested}_{\alpha}$ of Theorem 5 satisfies $\left(\mathcal{P}_{\text{adaptive},\alpha,\beta,\overline{\mathcal{F}_1}^{(ii)'},d_2}\right)$, with no price for adaptivity.

Simply combining the two above tests as defined in Theorem 4 and Theorem 5 (with a Bonferroni correction), that is taking the test $\bar{\Phi}_{\alpha} = \bar{\Phi}^{nested}_{\alpha/2}\vee\bar{\Phi}^{nonnested}_{\alpha/2}$, allows to obtain adaptivity over the collection of classes coming from both $\overline{\mathcal{F}_1}^{(i)}$ and $\overline{\mathcal{F}_1}^{(ii)'}$. More precisely, for n large enough, $\bar{\Phi}_{\alpha}$ satisfies $\left(\mathcal{P}_{\text{adaptive},\alpha,\beta,\overline{\mathcal{F}_1}^{(i)}\cup\overline{\mathcal{F}_1}^{(ii)'},d_2}\right)$. Notice that the parameter set considered in $\overline{\mathcal{F}_1}^{(i)}\cup\overline{\mathcal{F}_1}^{(ii)'}$ is the set $\{(s,s'),\ s\ge s'/(2s'+1)\}$. We now know, from the posterior work [10] (see [11] for more details), that some aggregated tests could be constructed, so that they are adaptive over $\overline{\mathcal{F}_1}^{(i)}\cup\overline{\mathcal{F}_1}^{(ii)}$.

We still however do not know what exactly happens for the parameter set considered in the case (iii) of Theorem 3.

1.4.3 Links with model selection and thresholding

Based on the idea that the Poisson process is closely related with the density model, the links between the first aggregated test $\bar{\Phi}^{nested}_{\alpha}$ based on a collection of nested subspaces and model selection become obvious, and are explained in Section 1.2.1.

Using a collection of extremely nonnested subspaces such as the one considered in our second aggregated test $\bar{\Phi}^{nonnested}_{\alpha}$ is more in the spirit of thresholding approaches, where each coefficient in the expansion of f is taken into account by itself, independently of the other ones, than model selection approaches, though bridges are well known between these two approaches.

In the above notation, following the thresholding principle, one could roughly estimate the quadratic functional $d_2^2\left(\Pi_{S_J}(f),\mathcal{F}_0\right)$ by $\sum_{(j,k)\in\mathcal{L}_J}\hat{\beta}^2_{(j,k)}\mathbb{1}_{|\hat{\beta}_{(j,k)}|\ge\lambda_{(j,k)}}$, where $\hat{\beta}_{(j,k)} = n^{-1}\sum_{i=1}^{N_{\mathbf{X}}(1)}\psi_{(j,k)}(X_i)$.

A test of (H_0) against (H_1) could then be introduced as

$$\mathbb{1}_{\left\{\sum_{(j,k)\in\mathcal{L}_J}\hat{\beta}^2_{(j,k)}\mathbb{1}_{|\hat{\beta}_{(j,k)}|\geq\lambda_{(j,k)}}>c_{J,\alpha}\right\}},$$

where the $\lambda_{(j,k)}$'s and $c_{J,\alpha}$ have to be correctly calibrated so that the test is of level α. This test is in fact equivalent to

$$\mathbb{1}_{\left\{\exists(j,k)\in\mathcal{L}_J,\ \hat{\beta}^2_{(j,k)}>c_{(j,k),\alpha}\right\}},$$

for well chosen critical values $c_{(j,k),\alpha}$'s. Replacing the rough estimator $\hat{\beta}^2_{(j,k)}$ by an unbiased estimator of $\langle f,\psi_{(j,k)}\rangle_2^2$ and correctly calibrating the critical values leads to the test $\bar{\Phi}_\alpha^{nonnested}$.
Notice that the idea is not completely new. Such aggregation of tests coming from wavelet shrinkage has already been proposed to construct adaptive tests in various statistical models. One can cite for instance the papers by Spokoiny ([Spo96] and [Spo98]) in Gaussian white noise models or by Butucea and Tribouley [BT06] in the density model. Combining some tools from model selection and wavelet analysis was proposed by Baraud, Huet and Laurent [BHL03] in a Gaussian regression framework, and in [5] in the density framework. However, in these tests, the individual levels of the tests (namely the $u_{m,\alpha}$'s) are identical for all the aggregated single tests. Such a choice does not allow to recover the minimax separation rates over weak Besov spaces. As seen above, to obtain minimax adaptivity on such spaces, a weight has to be correctly attributed to each individual level.

1.4.4 Experimental results

A simulation study is provided in [7], which aims at comparing the performance of the proposed tests in practice, through the estimated sizes and powers of the tests, with the one of some historical tests: the conditional Kolmogorov Smirnov test, the unconditional Laplace test, and the Z test studied, among others, in [BEW85, CS93]. The critical values involved in our tests are approximated by Monte Carlo methods, which are quite complex from an algorithmic point of view, as the simulations have to be done given each number of observed points in the Poisson process. The parameter n is chosen equal to 100, and the alternative intensities such that $\int_{[0,1]} f d\lambda = 1$, that is $N_{\mathbf{X}}(1)$ has a Poisson distribution with parameter $n = 100$. These alternatives, which are either rather smooth, or very irregular, are displayed in [7, Section 5], and one can see that the inhomogeneity is not really visible, just looking at histograms, especially when the alternatives are similar to a uniform density with very localized spikes. When the alternatives are not increasing, our tests have estimated powers significantly larger than the Laplace and Z tests that are designed for increasing alternatives, but also than Kolmogorov and Smirnov's test.
As expected, the test $\bar{\Phi}_\alpha^{nested}$ performs better than $\bar{\Phi}_\alpha^{nonnested}$ when the alternatives are smooth, and the trend is in general reversed when the alternatives are very irregular. But we also found some cases for which this rule is not valid. The combination of both tests is therefore recommended in any practical situation.
As for the increasing alternatives, the specific Laplace and Z tests remain the most powerful ones (as they are precisely designed to be the most powerful ones in a parametric context), especially when the alternatives are smooth. In this case however, the performance of our tests could surely be significantly improved by using the Fourier basis instead of the Haar basis (see [5] for instance).

1.4.5 Tools and sketches of proofs

Sketch of proof for the lower bound in Theorem 3. The proof follows similar arguments as the ones for the lower bound explained in Section 1.3.5. The idea is therefore to construct a class of alternatives $\Omega(r^*) \subset \{f,\ d_2(f,\mathcal{F}_0) = r^*\}$, with r^* as large as possible, included in $\mathcal{B}_{s,2,\infty}(R)\cap w\mathcal{B}_{s'}(R')$,

but for which

$$\inf_{\phi_\alpha \text{ satisfying } (\mathcal{P}_{\text{level},\alpha})} \sup_{f \in \Omega(r^*)} P_f\left(\phi_\alpha = 0\right) \geq \beta.$$

1. The first step is to introduce the class of alternatives

$$\Omega_{M,D}(r) = \left\{ f = C\varphi_0 + r\sqrt{\frac{M}{D}} \sum_{l=1}^{M} \beta_l \varphi\left(M. - l + 1\right),\ \beta_l \in \{-1, 0, 1\},\ \sum_{l=1}^{M} \mathbb{1}_{\beta_l \neq 0} = D \right\},$$

   where φ is a function on $[0,1]$ such that $\int_{[0,1]} \varphi d\lambda = 0$, $\int_{[0,1]} \varphi^2 d\lambda = 1$, and $\|\varphi\|_\infty \leq C$. Note that for every f in $\Omega_{M,D}(r)$, as soon as $r^2 \leq D/M$, f is positive, and $d_2(f, \mathcal{F}_0) = r$.

2. The second step is then to determine a maximal radius $r_{M,D} \leq \sqrt{D/M}$ such that for all $r \leq r_{M,D}$,

$$\inf_{\phi_\alpha \text{ satisfying } (\mathcal{P}_{\text{level},\alpha})} \sup_{f \in \Omega_{M,D}(r)} P_f\left(\phi_\alpha = 0\right) \geq \beta.$$

   This step is still based on the Bayesian techniques developed by Ingster [Ing93], and described in Section 1.3.5. The prior μ_r on $\Omega_{M,D}(r)$ is here constructed by considering

$$f_{\xi,\Delta}(x) = C\varphi_0(x) + r\sqrt{\frac{M}{D}} \sum_{l=1}^{M} \Delta_l \xi_l \varphi\left(Mx - l + 1\right),$$

   where $\xi = (\xi_1, \ldots, \xi_M)$ is a sample of i.i.d. Rademacher variables, and $\Delta = (\Delta_1, \ldots, \Delta_M)$ is a random vector, independent of ξ and defined by $\Delta_l = \mathbb{1}_{l \in \mathcal{L}}$, where $\mathcal{L}$ is a set of D indices drawn at random from $\{1, \ldots, M\}$ without replacement.

3. The third step is to find the maximal radius $r_{M,D,s,s',R,R'} \leq r_{M,D}$ so that $\Omega_{M,D}(r_{M,D,s,s',R,R'})$ is included in $\mathcal{B}_{s,2,\infty}(R) \cap w\mathcal{B}_{s'}(R')$, and to take $\Omega(r^*) = \Omega_{M^*,D^*}(r_{M^*,D^*,s,s',R,R'})$ with appropriate M^* and D^*.

Notice that the computations, which are here more difficult than the ones when only considering classical Besov bodies, heavily rely on the independence properties of the Poisson process. Therefore, the proof is not straightforwardly exportable in the density model.

Tools for upper bounds. A first point is to see that the two considered tests can be rewritten with a same expression. They indeed both take the value 1 if and only if

$$\exists \mathcal{L} \in \mathcal{C},\ \sum_{(j,k) \in \mathcal{L}} \frac{1}{n^2} \sum_{i \neq i' = 1}^{N_{\mathbf{X}}(1)} \psi_{(j,k)}(X_i) \psi_{(j,k)}(X_{i'}) > c_{\mathcal{L},\alpha}^{(N_{\mathbf{X}}(1))},$$

where:

— for $\bar{\Phi}_\alpha^{nested}$, $\mathcal{C} = \{\mathcal{L}_J,\ J \in \mathcal{M}\}$ and $c_{\mathcal{L}_J,\alpha}^{(N_{\mathbf{X}}(1))} = q_J^{(N_{\mathbf{X}}(1))}\big(1 - u_{J,\alpha}^{(N_{\mathbf{X}}(1))}\big)$, with $u_{J,\alpha}^{(N_{\mathbf{X}}(1))}$ given by (1.26);

— for $\bar{\Phi}_\alpha^{nonnested}$, $\mathcal{C} = \{\mathcal{L},\ \mathcal{L} \subset \mathcal{L}_{\bar{J}}\}$ and $c_{\mathcal{L},\alpha}^{(N_{\mathbf{X}}(1))} = \sum_{(j,k) \in \mathcal{L}} q_{(j,k)}^{(N_{\mathbf{X}}(1))}\big(1 - u_{(j,k),\alpha}^{(N_{\mathbf{X}}(1))}\big)$, with $u_{(j,k),\alpha}^{(N_{\mathbf{X}}(1))}$ given by (1.28).

From this shared expression, an oracle type result is obtained for both tests at the same time, following rather similar arguments as in the proof of Theorem 1 proved in [5]. In particular, concentration

inequalities for U-statistics of order 2 for a Poisson process in [HRB03, Theorem 4.2] are used since

$$\sum_{(j,k)\in\mathcal{L}} \frac{1}{n^2} \sum_{i\neq i'=1}^{N_{\mathbf{X}}(1)} \psi_{(j,k)}(X_i)\psi_{(j,k)}(X_{i'})$$
$$= \frac{1}{n^2} \sum_{(j,k)\in\mathcal{L}} \left[\left(\int_{\mathbb{X}} \psi_{(j,k)}(x) \left(dN_{\mathbf{X}}(x) - f(x) n d\lambda(x) \right) \right)^2 - \int_{\mathbb{X}} \psi^2_{(j,k)}(x) dN_{\mathbf{X}}(x) \right]$$
$$+ \frac{2}{n} \int_{\mathbb{X}} \left(\Pi_{S_{\mathcal{L}}}(f)(x) - \langle f, \varphi_0 \rangle_2 \varphi_0(x) \right) \left(dN_{\mathbf{X}}(x) - f(x) n d\lambda(x) \right) + d_2^2(f, \mathcal{F}_0) - \| f - \Pi_{S_{\mathcal{L}}}(f) \|_2^2 ,$$

where $S_{\mathcal{L}}$ is naturally the space spanned by $\{\varphi_0,\ \psi_{(j,k)},\ (j,k) \in \mathcal{L}\}$.
A more tricky point is here the control of the random critical values $c_{\mathcal{L},\alpha}^{(N_{\mathbf{X}}(1))}$. Using again concentration inequalities for U-statistics, but for a sample of uniform real random variables, from [HRB03, Theorem 3.4], allows to upper bound $c_{\mathcal{L},\alpha}^{(n_0)}$ for every given integer n_0. Then, a control of the obtained upper bound, but replacing n_0 by the random variable $N_{\mathbf{X}}(1)$, is computed thanks to Bernstein's inequality, thus leading to the expected oracle type result.
Theorem 4 and Theorem 5 are rather straightforward consequences of this oracle type result, by standard arguments from the approximation theory as concerns classical Besov bodies. The arguments are of course more difficult when considering weak Besov bodies.

1.5 Perspectives

Only short-term perspectives which are already the objects of current works are stated in the present dissertation.

It is rather clear that the minimax separation rates and the minimax adaptive tests presented in Section 1.4 and Section 3.2 could be adapted to handle goodness-of-fit testing problems in the density model for instance, but also signal detection problems. The issue is however not so obvious, as the lower bounds for the minimax separation rates over weak Besov bodies, obtained here, strongly rely on the very convenient independence properties of the Poisson processes.

Other interesting points would be to extend the aggregated tests to general kernels based tests, in the spirit of [9] and [10] (see Chapter 3), thus exploiting the properties of the RKHS based on reproducing kernels. The recent works in the statistical learning literature on the links between classical weak distances between distributions and the MMD distance (see [SSGF13] and references therein) open up a wide field of research, giving new insights to many historical tests such as Kolmogorov-Smirnov type ones.

Furthermore, our work on multiple testing, and in particular on the parallel between multiple tests and aggregated tests (see Chapter 5), should help us to improve the present aggregation scheme. Some improvements have already been made with the introduction of some weights in the individual levels in [7, 10] (with respect to the tests introduced in [BHL03]), so that the aggregated tests can now be minimax adaptive over classes of very irregular alternatives such as weak Besov bodies. But one can also imagine new refinements, taking advantage of knowledge on multiple step-down testing procedures, to increase the power of the tests.

Chapter 2

Contributions to classification

2.1 Introduction

In this chapter, we deal with the usual problem of binary classification in statistical learning, where the observed random variable is distributed according to the following model.

$\mathcal{M}^{(1)}_{\text{classification}}$	$\mathbf{X} = \mathbf{X}_n = (X_1, \dots, X_n)$ is a sample of n i.i.d. random variables $X_i = (Y_i, Z_i)$ with the same distribution P as a pair of random variables (Y, Z), defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, with values in a measurable space $\mathbb{X} = \mathbb{Y} \times \{0, 1\}$.

In the above model, the marginal distribution of Y is denoted by P_Y, and the expectation with respect to P_Y by $\mathbb{E}_Y$. The set of possible probability distributions P is denoted by $\mathcal{P}$.

The purpose of classification is to construct a (measurable) function, called a classification rule or a classifier, $\hat{f}_n : \mathbb{Y} \to \{0, 1\}$ based on $\mathbf{X}_n$, which allows to predict the value of Z from the observation of Y. When the value of Z is not fully determined by Y and $\mathbf{X}_n$, the prediction suffers from the classification error defined by $L(\hat{f}_n) = \mathbb{P}(\hat{f}_n(Y) \neq Z | \mathbf{X}_n)$. The function f^* minimizing the classification error $L(f) = \mathbb{P}\,(\,f(Y) \neq Z\,)$, over the set $\mathcal{F}$ of all possible measurable functions $f : \mathbb{Y} \to \{0, 1\}$, is called the Bayes classifier. From the regression function $\eta : y \in \mathbb{Y} \mapsto \mathbb{P}\,(\,Z = 1 | Y = y\,)$, the Bayes classifier is expressed as $f^*(y) = \mathbb{1}_{\{\eta(y) > 1/2\}}$. In statistical terms, classification corresponds to the estimation of this Bayes classifier f^* from the sample $\mathbf{X}_n$, and the theoretical performance of any estimator $\hat{f}_n$ can be evaluated by comparing $\mathbb{E}\big[L(\hat{f}_n)\big]$ with $L(f^*)$. In particular, $\hat{f}_n$ is said to be universally consistent if $\mathbb{E}[L(\hat{f}_n)]$ tends to $L(f^*)$ as n tends to $+\infty$ for every P in $\mathcal{P}$. So many references are devoted to universal consistency of classifiers, that they cannot be all cited, all the more as the methods of construction of classifiers are highly varied, but some review of the historical results at least can be found in [DGL96]. We here investigate classifiers obtained as penalized empirical classification error minimizers on the one hand, and plug-in rules combined with a hold-out or data-splitting device on the other hand, from a nonasymptotic point of view based on oracle inequalities, leading to minimax adaptivity properties.

Although the present chapter does not deal with nonparametric testing as all the other ones of this dissertation, it is in fact closely linked to most of the ideas developed in the next chapter on two-sample problems, mixing classical minimax adaptive tests with bootstrap and statistical learning approaches such as (reproducing) kernels and k-Nearest Neighbors methods.

2.1.1 Nonasymptotic minimax adaptivity

From a nonasymptotic point of view, the difference between $\mathbb{E}[L(\hat{f}_n)]$ and $L\,(\,f^*\,)$, for a given sample size n, usually defined as the excess risk of the estimator $\hat{f}_n$, allows to introduce the following definition.

Definition 3 (Risk of a classifier)**.** Let $\hat{f}_n$ be a classifier based on $\mathbf{X}_n$, that is an estimator of the Bayes classifier f^*. For a given subset $\mathcal{Q}$ of $\mathcal{P}$, the risk of $\hat{f}_n$ over $\mathcal{Q}$ is defined by:

$$R_n\left(\hat{f}_n, \mathcal{Q}\right) = \sup_{P \in \mathcal{Q}} \left(\mathbb{E}\left[L\left(\hat{f}_n\right)\right] - L\left(f^*\right)\right).$$

Definition 4 (Minimax risk, minimax (adaptive) classifier)**.** Let $\overline{\mathcal{Q}}$ be a collection of subsets of $\mathcal{P}$. The minimax risk over some probability distributions set $\mathcal{Q}$ in $\overline{\mathcal{Q}}$ is defined as

$$mR_n\left(\mathcal{Q}\right) = \inf_{\hat{f}_n} R_n\left(\hat{f}_n, \mathcal{Q}\right) = \inf_{\hat{f}_n} \sup_{P \in \mathcal{Q}} \left(\mathbb{E}\left[L\left(\hat{f}_n\right)\right] - L\left(f^*\right)\right),$$

where the infimum is taken over all the possible classifiers, and f^* is the Bayes classifier.
A classifier $\hat{f}_n$ based on $\mathbf{X}_n$ is said to be minimax over $\mathcal{Q}$ if $R_n(\hat{f}_n, \mathcal{Q})$ achieves $mR_n\left(\mathcal{Q}\right)$, possibly up to a multiplicative constant.
It is said to be minimax adaptive over the collection $\overline{\mathcal{Q}}$ if $R_n(\hat{f}_n, \mathcal{Q})$ achieves, or nearly achieves, $mR_n\left(\mathcal{Q}\right)$ for every $\mathcal{Q}$ in $\overline{\mathcal{Q}}$ simultaneously. This property is formalized in the following as

$$\left(\mathcal{P}_{\text{adaptive},\overline{\mathcal{Q}}}\right) \qquad \Big| \; R_n(\hat{f}_n, \mathcal{Q}) \text{ achieves or nearly achieves } mR_n\left(\mathcal{Q}\right), \text{ for every } \mathcal{Q} \text{ in } \overline{\mathcal{Q}}.$$

In the case where $\mathbb{Y} \subset \mathbb{R}^d$, the subsets $\mathcal{Q}$ of $\mathcal{P}$, over which the risk or the minimax risk is evaluated, are usually defined thanks to complexity and margin assumptions.

For instance, consider $\mathcal{Q} = \mathcal{P}_{\mathcal{C}} = \{ P \in \mathcal{P},\ f^* \in \{ \mathbb{1}_C,\ C \in \mathcal{C} \} \}$, where $\mathcal{C}$ is a Vapnik-Chervonenkis (VC) class of subsets of $\mathbb{Y}$, that is a class of subsets of $\mathbb{Y}$ with finite VC dimension:

$$V(\mathcal{C}) = \sup \left\{ k \geq, \max_{y_1,\ldots,y_k \in \mathbb{Y}} \# \{ \{ y_1, \ldots, y_n \} \cap C,\ C \in \mathcal{C} \} = 2^k \right\} < \infty.$$

A now well-known lower bound for the minimax risk over $\mathcal{P}_{\mathcal{C}}$ was obtained by Vapnik and Chervonenkis [VC74]: there exist some positive constants κ_1 and κ_2 such that

$$mR_n\left(\mathcal{P}_{\mathcal{C}}\right) \geq \kappa_1 \sqrt{V(\mathcal{C})/n} \quad \text{for } n \geq \kappa_2 V(\mathcal{C}).$$

This lower bound was then proved to be achieved, up to a multiplicative constant, by the Empirical Risk Minimizer on $\{ \mathbb{1}_C,\ C \in \mathcal{C} \}$ (see (2.2) below) in [Lug02]. The construction of minimax adaptive classifiers has been the purpose of many papers, and the penalized ERM approach account for a large part in these references. Section 2.2 below deal with this issue, and in particular with the idea of general weighted bootstrap penalization, first introduced in [1, 2, 6] as an extension of the Rademacher penalization due to Koltchinskii [Kol01] and [BBL02].
Considering the optimistic situation, called the zero-error case, where the Bayes classification error is assumed to be equal to zero, Vapnik and Chervonenkis [VC74] and Haussler et al. [HLW94] gave a lower bound for the minimax risk over the smaller set $\mathcal{Q} = \mathcal{P}_{\mathcal{C},\text{ZE}} = \{ P \in \mathcal{P},\ f^* \in \{ \mathbb{1}_C, C \in \mathcal{C} \},\ L\left(f^*\right) = 0 \}$. They proved that there exist some positive constants κ_3 and κ_4, such that

$$mR_n\left(\mathcal{P}_{\mathcal{C},\text{ZE}}\right) \geq \kappa_3 V(\mathcal{C})/n \quad \text{for } n \geq \kappa_4 V(\mathcal{C}).$$

This lower bound is achieved, up to a $(\ln n)$ factor, by some ERM classifiers (see [DW76] and [Vap82]). In the following, the minimax risk $mR_n\left(\mathcal{P}_{\mathcal{C},\text{ZE}}\right)$ is referred to as the *zero-error minimax risk* for the VC class $\mathcal{C}$, by contrast to the minimax risk $mR_n\left(\mathcal{P}_{\mathcal{C}}\right)$ referred to as the *global minimax risk* for $\mathcal{C}$.
The main point here is that the zero-error minimax risk is of smaller order of magnitude than the global one, and that the difference is really significant ($V(\mathcal{C}) \ln n/n$ at most instead of $\sqrt{V(\mathcal{C})/n}$). This lead to the intuition that the minimax risk can be acutely analyzed when restricting to probability distribution such that the Bayes classification error is not necessarily exactly equal to zero but very

small. Devroye and Lugosi [DL95] and Lugosi [Lug02] gave such a refined analysis in a case which can be viewed as a kind of interpolation between the global case and the zero-error one. In their proofs, the behavior of the regression function η around $1/2$ turns out to be crucial. Mammen and Tsybakov [MT99] first investigated the influence of this behavior by introducing margin assumptions, and then were followed by many authors (see [Tsy04, MN06, Kol06, AT07, Lec07a, AB11] among many other references).
Several kinds of complexity and margin assumptions have been considered in the literature.
As for complexity assumptions, a review can be found in [Kol06] notably including assumptions expressed in terms of VC dimension (as in [MN06]) or entropy with bracketing (as in [Tsy04]) for instance. Another kind of complexity assumption is studied in [AT07].
As for margin assumptions, we introduce, for h in $[0,1]$, and $\theta \geq 1$, the following general margin assumption deduced from [MT99] and [Tsy04]:

$$\mathrm{GMA}(\theta, h): \ L(f) - L(f^*) \geq \left(h\mathbb{E}_Y[|f(Y) - f^*(Y)|]\right)^\theta, \ \forall f \in \mathcal{F}.$$

Notice that if there exists $\kappa > 0$, such that

$$\mathrm{MA}(\alpha): \ \mathbb{P}[|\eta(Y) - 1/2| \leq t] \leq \kappa t^\alpha, \ \forall t > 0,$$

then $\mathrm{GMA}((1+\alpha)/\alpha, h)$ is satisfied for some h in $[0,1]$. Moreover, if

$$\mathrm{MA}(\infty): \ |2\eta(y) - 1| \geq h, \ \forall y \in \mathbb{Y},$$

then $\mathrm{GMA}(1, h)$ is satisfied.

Let $\mathcal{P}_{\mathcal{C},\mathrm{GMA}(\theta,h)} = \{ P \in \mathcal{P}, \ f^* \in \{ \mathbb{1}_C, \ C \in \mathcal{C} \}, \ P \text{ satisfies } \mathrm{GMA}(\theta, h) \}$, for some VC class $\mathcal{C}$ satisfying an appropriate condition of separability denoted by (M) in [MN06]. Then Massart and Nédélec [MN06] prove that the ERM classifier $\hat{f}_n$ over $\{ \mathbb{1}_C, \ C \in \mathcal{C} \}$ satisfies, for every $h \geq (V(\mathcal{C})/n)^{1/(2\theta)}$,

$$R_n\left(\hat{f}_n, \mathcal{P}_{\mathcal{C},\mathrm{GMA}(\theta,h)}\right) \leq \kappa_5 \left(V(\mathcal{C})(1 + \ln(nh^{2\theta}/V(\mathcal{C})))/(nh)\right)^{\frac{\theta}{2\theta-1}}. \tag{2.1}$$

They also prove that this ERM classifier is optimal in the following (weak) minimax sense, when $\theta = 1$. If $\mathcal{C}$ is a VC class of subsets of $\mathbb{Y}$ such that $2 \leq V(\mathcal{C}) \leq n$, then

$$\inf_{\hat{f}_n \in \{ \mathbb{1}_C, \ C \in \mathcal{C} \}} R_n\left(\hat{f}_n, \mathcal{P}_{\mathcal{C},\mathrm{GMA}(1,h)}\right) \geq \kappa_6 \left\{ (V(\mathcal{C})/(nh)) \wedge \sqrt{V(\mathcal{C})/(nh)} \right\}.$$

This means in particular that no ERM classifier over any subset of $\{ \mathbb{1}_C, \ C \in \mathcal{C} \}$ can have a risk with faster rate of convergence than n^{-1} (super-fast rate). This however does not mean that no classifier can have a risk with such super-fast rate of convergence, as proved by Audibert and Tsybakov [AT07].

Let now for ρ, h in $(0,1)$, $\theta \geq 1$, and some set $\mathcal{F}'$ of measurable functions from $\mathbb{Y}$ to $\{0,1\}$,

$$\mathcal{P}_{\mathcal{F}',\rho,\mathrm{GMA}(\theta,h)} = \left\{ P \in \mathcal{P}, \ f^* \in \mathcal{F}', \ H(\varepsilon, \mathcal{F}', \mathbb{L}_1(\mathbb{Y}, P_Y)) = O\left(\varepsilon^{-\rho}\right) \forall \varepsilon \in (0,1), \ P \text{ satisfies } \mathrm{GMA}(\theta, h) \right\},$$

where $H(\varepsilon, \mathcal{F}', \mathbb{L}_1(\mathbb{Y}, P_Y))$ is the ε-entropy with bracketing of the set $\mathcal{F}'$ with respect to the $\mathbb{L}_1(\mathbb{Y}, P_Y)$ norm. Considering the ERM classifier $\hat{f}_n$ computed on a set depending on ρ, Massart and Nédélec [MN06] prove that

$$R_n\left(\hat{f}_n, \mathcal{P}_{\mathcal{F},\rho,\mathrm{GMA}(\theta,h)}\right) \leq \kappa_7 \left\{ \left((1-\rho)^2 n h^{1-\rho}\right)^{-\frac{\theta}{2\theta+\rho-1}} \wedge (1-\rho)^{-1} n^{-1/2} \right\},$$

which slightly refines Tsybakov's [Tsy04] result as it gives the precise dependence with respect to h. This upper bound is proved to be optimal in a minimax sense by Tsybakov [Tsy04] when $\mathbb{Y} = [0,1]^d$, and in a slightly weaker minimax sense by Massart and Nédélec [MN06] when $\theta = 1$. Minimax adaptive classifiers over some collection of such classes $\mathcal{P}_{\mathcal{F},\rho,\mathrm{GMA}(\theta,h)}$ are constructed in Tsybakov [Tsy04], laying the foundation of the approach of optimal aggregation of estimators, now largely developed.

2.1.2 Penalized ERM or model selection by penalization

As seen above, minimax classifiers can be obtained by the historical Empirical Risk Minimization (ERM) approach introduced in learning issues by [VC71]. Given a set $\mathcal{F}'$ of measurable functions from $\mathbb{X}$ to $\{0,1\}$, the ERM classifier over $\mathcal{F}'$ is defined as

$$\hat{f}_{n,\mathcal{F}'} \in \operatorname{argmin}_{f\in\mathcal{F}'} L_n(f), \tag{2.2}$$

where

$$L_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{f(Y_i)\neq Z_i}.$$

The question of minimax adaptivity here becomes a question of choice of an appropriate function set $\mathcal{F}'$. The famous method of Structural Risk Minimization (SRM) initiated by Vapnik [Vap82] and also known as complexity regularization [Bar91] consists in selecting, among a given collection of function sets, the set $\mathcal{F}'$ minimizing the penalized criterion $L_n(\hat{f}_{n,\mathcal{F}'}) + \mathrm{pen}(\mathcal{F}')$. The penalty term $\mathrm{pen}(\mathcal{F}')$ usually involves a quantity measuring the complexity of $\mathcal{F}'$, such as for instance the VC dimension of the class of subsets associated with $\mathcal{F}'$. Considering a collection $\{\mathcal{F}_m,\ m \in \mathbb{N}\setminus\{0\}\}$ of subsets of $\mathcal{F}$ such that each $\mathcal{C}_m = \{\{y \in \mathbb{X},\ f(y) = 1\},\ f \in \mathcal{F}_m\}$ is a VC class with VC dimension $V(\mathcal{C}_m)$, Lugosi and Zeger [LZ96] use some penalties of order $\kappa\sqrt{(V(\mathcal{C}_m)\ln n + m)/n}$. They prove that if the sequence $(V(\mathcal{C}_m))_{m\in\mathbb{N}\setminus\{0\}}$ is strictly increasing, and if the Bayes classifier f^* belongs to the union of the $\mathcal{F}_m$'s, there exists an integer k such that the risk of the SRM or penalized ERM classifier is upper bounded by $\sqrt{V(\mathcal{C}_k)\ln n/n}$, up to a multiplicative constant, that is the classifier is minimax over $\mathcal{F}_k$ up to a logarithmic factor.

This approach can be viewed as a classical model selection by penalization approach (see [Mas07]) for the estimation of the Bayes classifier f^*, where each function set $\mathcal{F}_m$ corresponds to a model in a given collection $\{\mathcal{F}_m,\ m \in \mathcal{M}\}$ and where L_n plays the role of the empirical contrast, associated with the contrast defined for every f in $\mathcal{F}$ by $\gamma(f,(y,z)) = \mathbb{1}_{f(y)\neq z}$. An ERM classifier $\hat{f}_{n,m} := \hat{f}_{n,\mathcal{F}_m}$ can thus be viewed as a minimum contrast estimator of f^*. For every m in $\mathcal{M}$, such a classifier, or more generally an approximate ERM classifier over $\mathcal{F}_m$, or an approximate minimum contrast estimator of f^* over $\mathcal{F}_m$, $\hat{f}_{n,m}$, is considered, that is

$$L_n\left(\hat{f}_{n,m}\right) \leq \inf_{f\in\mathcal{F}_m} L_n(f) + \rho_n/2, \tag{2.3}$$

for some $\rho_n \geq 0$. Denoting by $l(f,g) = L(g) - L(f)$, for all f and g in $\mathcal{F}$, the excess risk of $\hat{f}_{n,m}$ is given by $\mathbb{E}[l(f^*,\hat{f}_{n,m})]$. Ideally, we would like to select some element $\bar{m}$ (the oracle) in $\mathcal{M}$ minimizing

$$\mathbb{E}\left[l\left(f^*,\hat{f}_{n,m}\right)\right] = l\left(f^*,f_{n,m}\right) + \mathbb{E}\left[l\left(f_{n,m},\hat{f}_{n,m}\right)\right],$$

where $f_{n,m}$ denotes some function in $\mathcal{F}_m$ such that $l\left(f^*,f_{n,m}\right) = \inf_{f\in\mathcal{F}_m} l\left(f^*,f\right)$. However, such an oracle $\bar{m}$ necessarily depends on the unknown distribution P. The original idea of model selection by penalization is to select, only from the data, an element $\hat{m}$ in $\mathcal{M}$ mimicking the oracle. Considering some penalty function $\mathrm{pen} : \mathcal{M} \to \mathbb{R}_+$, $\hat{m}$ is chosen such that:

$$L_n\left(\hat{f}_{n,\hat{m}}\right) + \mathrm{pen}\left(\hat{m}\right) \leq \inf_{m\in\mathcal{M}}\left\{L_n\left(\hat{f}_{n,m}\right) + \mathrm{pen}(m)\right\} + \rho_n/2,$$

and the selected estimator $\hat{f}_{n,\hat{m}}$ is called the approximate minimum penalized contrast estimator. A challenge is to determine a penalty function such that

$$\begin{aligned}\mathbb{E}\left[l\left(f^*,\hat{f}_{n,\hat{m}}\right)\right] &\leq C\inf_{m\in\mathcal{M}}\left(l\left(f^*,f_{n,m}\right) + \mathbb{E}\left[l\left(f_{n,m},\hat{f}_m\right)\right]\right) + R_n\\ &\leq C\mathbb{E}\left[l\left(f^*,\hat{f}_{n,\bar{m}}\right)\right] + R_n,\end{aligned}$$

where R_n is a residual term. Such an inequality is called an oracle inequality.
The models and the penalty function are usually chosen so that from this oracle inequality, the selected classifier $\hat{f}_{n,\hat{m}}$ is proved to satisfy minimax adaptivity properties.
With this purpose in view, when considering the global minimax risk, the following inequality may be sufficient instead:

$$\mathbb{E}\left[l\left(f^*, \hat{f}_{n,\hat{m}}\right)\right] \leq C \inf_{m \in \mathcal{M}} \left(l\left(f^*, f_{n,m}\right) + \sqrt{\frac{V\left(\mathcal{C}_m\right)}{n}} \right) + R_n, \tag{2.4}$$

with a residual term of order at most $\sqrt{V\left(\mathcal{C}_m\right)/n}$.
The various strategies to construct adequate penalty functions in this global minimax context are all connected with the calibration of an upper bound for $\sup_{f \in \mathcal{F}_m} \left(L(f) - L_n(f)\right)$ (see Section 2.2). The penalties based on the VC dimension, such as the one of [LZ96] are deterministic and have the disadvantage to overestimate this supremum for specific data distributions. This remark is in favor of data-driven penalization approaches, such as Rademacher penalization ones introduced by Koltchinskii [Kol01] and Bartlett, Boucheron, Lugosi [BBL02].

Although Rademacher variables were already used in the wild bootstrap approach (see [Mam92] for instance), the connection between such Rademacher penalization approaches in classification, and classical model selection with bootstrap penalization, was neither studied, nor even clearly stated until the work which is presented in Section 2.2. The purpose of [1, 2, 6] was actually to establish this connection, and thereby to extend the Rademacher penalties to a wider family of penalties based on the general weighted bootstrap approach.

2.1.3 Plug-in classifiers

A plug-in classifier is of the form $\hat{f}_n(y) = \mathbb{1}_{\hat{\eta}_n(y) \geq 1/2}$, for all y in $\mathbb{Y}$, where $\hat{\eta}_n$ is a nonparametric estimator of the regression function η. The most simple plug-in classifiers are probably the kernel and k-Nearest Neighbors (k-NN) rules, which have been largely studied from both theoretical and practical points of view, in particular when the input data space $\mathbb{Y}$ is assumed to be equal to $\mathbb{R}^d$. In this case, conditions leading to universal consistency, and even strong universal consistency are well-known (see [DGL96] and [GKKW02] for an overview). Most of these results are based on the same inequality proved in [DGL96, Theorem 2.2]:

$$\mathbb{E}\left[L\left(\hat{f}_n\right)\right] - L\left(f^*\right) \leq 2\mathbb{E}\left[\int \left|\hat{\eta}_n(y) - \eta(y)\right| dP_Y(y)\right],$$

which enables to see that closeness between $\hat{\eta}_n$ and η implies closeness between $\hat{f}_n$ and f^* in terms of risk. The papers [CH67], [Sto77] and [DGKL94] for instance deal with consistency of kernel and k-NN regression function estimators, while [KK06] give more precise rates of convergence of $\mathbb{E}[L(\hat{f}_n)] - L\left(f^*\right)$ towards 0, under regularity conditions on η. Audibert and Tsybakov [AT07] study the plug-in classifiers from the minimax point of view, considering the margin assumption $\mathrm{MA}(\alpha)$, and complexity assumptions expressed as regularity assumptions on the regression function. For instance, introducing

$$\mathcal{P}_{\beta,\mathrm{MA}(\alpha)} = \big\{P \in \mathcal{P},\ \eta \text{ belongs to a Hölder class of functions of order } \beta,\ P \text{ satisfies } \mathrm{MA}(\alpha),\ P_Y \text{ satisfies the strong density assumption}\big\},$$

(see [AT07, Definition 2.2] for the definition of the strong density assumption), Audibert and Tsybakov construct a plug-in classifier whose risk over $\mathcal{P}_{\beta,\mathrm{MA}(\alpha)}$ is at most of order $n^{-\beta(1+\alpha)/(2\beta+d)}$. This risk is then proved to be optimal when $\alpha\beta < d$. Audibert and Tsybakov [AT07] thus give evidence that plug-in classifiers can achieve faster rates of convergence (named super-fast rates when $\alpha\beta > d$) than ERM classifiers under margin assumptions. They further notify that such super-fast rates can not be

achieved when the strong density assumption is relaxed to a mild density assumption. Nevertheless, fast-rates of order n^{-1} are still achievable by a "hybrid" plug-in/ERM type classifier.
The paper by Kohler and Krzyżak [KK06] completes the picture, focusing on k-NN rules, that are at the core of the work presented in Section 2.3.

2.2 Model selection by bootstrap penalization

This section is devoted to a work on bootstrap penalization, that was initiated in [1] and [2], then improved and published in a longer version in [6].
Still considering the above problem of binary classification from an observed sample $\mathbf{X} = \mathbf{X}_n$ distributed according to the model $\mathcal{M}^{(1)}_{\text{classification}}$, and keeping the same notation, we introduce a collection $\{\mathcal{F}_m,\ m \in \mathcal{M}\}$ of models, that is a collection of subsets of $\mathcal{F}$, such that for all m in $\mathcal{M}$, $\mathcal{C}_m = \{\{y \in \mathbb{Y},\ f(y) = 1\},\ f \in \mathcal{F}_m\}$ is a VC class with VC dimension $V(\mathcal{C}_m)$. For every m in $\mathcal{M}$, $\hat{f}_{n,m}$ denotes an approximate ERM classifier satisfying (2.3). For some penalty function pen : $\mathcal{M} \to \mathbb{R}_+$, recall that $\hat{m}$ is chosen such that:

$$L_n\left(\hat{f}_{n,\hat{m}}\right) + \text{pen}\,(\hat{m}) \leq \inf_{m\in\mathcal{M}}\left\{L_n\left(\hat{f}_{n,m}\right) + \text{pen}(m)\right\} + \rho_n/2,$$

and consider the selected estimator $\hat{f}_{n,\hat{m}}$.
As explained in the above section, when considering the nonasymptotic minimax point of view with the global minimax risk, one should choose a penalty function such that (2.4) holds.

The various strategies to determine adequate penalty functions with this purpose in mind rely on the same following basic inequality.
Let us fix m in $\mathcal{M}$ and introduce the centered empirical contrast defined by:

$$\forall f \in \mathcal{F},\ \overline{L_n}(f) = L_n(f) - L(f). \tag{2.5}$$

By definition,

$$l\left(f_{n,m}, \hat{f}_{n,\hat{m}}\right) = \overline{L_n}\,(f_{n,m}) - L_n\,(f_{n,m}) - \overline{L_n}\left(\hat{f}_{n,\hat{m}}\right) + L_n\left(\hat{f}_{n,\hat{m}}\right).$$

Noticing that

$$L_n\left(\hat{f}_{n,\hat{m}}\right) + \text{pen}(\hat{m}) \leq L_n\left(\hat{f}_{n,m}\right) + \text{pen}(m) + \rho_n/2 \leq L_n\,(f_{n,m}) + \text{pen}(m) + \rho_n,$$

we derive

$$l\left(f^*, \hat{f}_{n,\hat{m}}\right) \leq l\,(f^*, f_{n,m}) + \overline{L_n}\,(f_{n,m}) + \text{pen}(m) - \overline{L_n}\left(\hat{f}_{n,\hat{m}}\right) - \text{pen}(\hat{m}) + \rho_n, \tag{2.6}$$

which holds whatever the penalty function. Since $\mathbb{E}\left[\overline{L_n}\,(f_{n,m})\right] = 0$, one way to obtain (2.4) is to choose a penalty such that $\text{pen}(\hat{m})$ compensates for the fluctuations of $-\overline{L_n}\big(\hat{f}_{n,\hat{m}}\big)$ and such that $\mathbb{E}\left[\text{pen}(m)\right]$ is of order at most $\sqrt{V\left(\mathcal{C}_m\right)/n}$. In this case, one needs to control $-\overline{L_n}(f)$ uniformly for f in $\mathcal{F}_m$ and m in $\mathcal{M}$ and concentration inequalities for the supremum $\sup_{f\in\mathcal{F}_m}\big(-\overline{L_n}(f)\big)$ appear as the appropriate tools.
Since the considered contrast γ is bounded, McDiarmid's [McD89] concentration inequality can be used to see that for all m in $\mathcal{M}$, $\sup_{f\in\mathcal{F}_m}\big(-\overline{L_n}(f)\big)$ concentrates around its expectation. A well-chosen estimator of an upper bound for $\mathbb{E}\big[\sup_{f\in\mathcal{F}_m}\big(-\overline{L_n}(f)\big)\big]$, with expectation of order $\sqrt{V\left(\mathcal{C}_m\right)/n}$, may therefore be a good penalty.

2.2.1 Rademacher and symmetrization based penalties

Starting from symmetrization tools used in the empirical processes theory, Koltchinskii [Kol01] and Bartlett, Boucheron and Lugosi [BBL02] propose a penalty based on the random variable

$$\hat{R}_m = \mathbb{E}\left[\sup_{f\in\mathcal{F}_m} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \mathbb{1}_{f(Y_i)\neq Z_i} \middle| \mathbf{X}_n \right],$$

where $(\varepsilon_1,\ldots,\varepsilon_n)$ is a sample of independent identically distributed Rademacher variables such that $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$ and the ε_i's are independent of $\mathbf{X}_n$. More precisely, they take $\mathcal{M} = \mathbb{N}\setminus\{0\}$ and they consider $\hat{f}_{n,\hat{m}}$ with $\mathrm{pen}(m) = 2\hat{R}_m + \kappa_1\sqrt{\ln m/n}$, for some absolute, positive constant κ_1. They prove that there exists some constant $\kappa_2 > 0$ such that

$$\mathbb{E}\left[l\left(f^*, \hat{f}_{n,\hat{m}}\right)\right] \leq \inf_{m\in\mathcal{M}} \{ l(f^*, f_{n,m}) + \mathbb{E}[\mathrm{pen}(m)]\} + \frac{\kappa_2}{\sqrt{n}} + \rho_n.$$

Moreover, Koltchinskii notes that this leads to

$$\mathbb{E}\left[l\left(f^*, \hat{f}_{n,\hat{m}}\right)\right] \leq \kappa \inf_{m\in\mathcal{M}} \left\{ l(f^*, f_{n,m}) + \left(\sqrt{\frac{V(\mathcal{C}_m)}{n}} + \sqrt{\frac{\ln m}{n}} + \frac{1}{\sqrt{n}} \right) \right\} + \rho_n.$$

Our aim in [1], [2] and in [6] was to extend this study by investigating penalty functions based on random variables of the form

$$\hat{R}_m^{W_n} = \mathbb{E}\left[\sup_{f\in\mathcal{F}_m} \frac{1}{n}\sum_{i=1}^{n} (1 - W_{n,i}) \mathbb{1}_{f(Y_i)\neq Z_i} \middle| \mathbf{X}_n \right], \tag{2.7}$$

with various vectors of random weights $W_n = (W_{n,1},\ldots,W_{n,n})$.
Remark that $\hat{R}_m$ corresponds the particular case where $W_n = 2(B_1,\ldots,B_n)$ is a sample of i.i.d. random variables, the B_i's being Bernoulli random variables with parameter $1/2$.

To avoid dealing with measurability issues, let us assume in all the sequel that any considered set of measurable functions from $\mathbb{Y}$ to $\{0,1\}$ is at most countable.

Noticing that the symmetrization trick used by Koltchinskii [Kol01] and Bartlett, Boucheron, Lugosi [BBL02] can be applied to any symmetric (and not only Rademacher) variable with a finite first order moment, we first introduced penalties based on $\hat{R}_m^{W_n}$ defined in (2.7) where $(1 - W_{n,1},\ldots,1 - W_{n,n})$ is a sample of n i.i.d. symmetric random variables such that $\mathbb{E}[|W_{n,1}|] < +\infty$. Notice that this generalization allows to consider, besides Rademacher penalization, the Gaussian complexity measure defined by Bartlett and Mendelson [BM03].

2.2.2 Weighted bootstrap penalties

The general weighted bootstrap approach. Bootstrap methods were introduced by Efron [Efr79] whose aim was to generalize and improve the ideas of the jackknife from Quenouille [Que49] and Tukey [Tuk58]. These methods were originally developed for a sample $\mathbf{X}_n = (X_1,\ldots,X_n)$ of i.i.d. real valued random variables or vectors with distribution P, and a root $R_n = R_n(\mathbf{X}_n; P)$, defined as a functional of the sample $\mathbf{X}_n$ and the common distribution P, whose probabilistic characteristics having a particular interest from a statistical point of view (distribution, or expectation, variance, quantiles, etc.) are unknown, and have to be estimated. Denoting by P_n the empirical measure associated with $\mathbf{X}_n$ defined by $P_n = n^{-1}\sum_{i=1}^{n}\delta_{X_i}$, Efron's original idea was to replace in the expression of $R_n(\mathbf{X}_n; P)$, P by P_n, and $\mathbf{X}_n$ by an i.i.d. sample from P_n denoted by $\mathbf{X}_n^* = (X_{n,1}^*,\ldots,X_{n,n}^*)$ and called a bootstrap sample from $\mathbf{X}_n$. The conditional distribution of the resulting bootstrapped root $R_n^* = R_n(\mathbf{X}_n^*; P_n)$ given $\mathbf{X}_n$ is then proposed as an estimator of the distribution of R_n.

This intuitive estimation method was justified theoretically by asymptotic arguments that were first specific to the considered root and its probabilistic characteristics of interest (see [Efr79] and for instance papers on the bootstrap of the mean, then linear or related statistics, like [Sin81], [BF81], [Ath87], [GZ89] among others).
These arguments are generally based on a result of consistency as:

$$\mathcal{L}(R_n^*|\mathbf{X}_n) \simeq \mathcal{L}(R_n),$$

meaning that the conditional distribution of R_n^* given $\mathbf{X}_n$ converges in probability or almost surely to the asymptotic distribution of R_n. Then, considering the empirical process $\mathbb{G}_n = \sqrt{n}(P_n - P)$, general results on the consistency of the bootstrapped empirical process $\mathbb{G}_n^* = \sqrt{n}(P_n^* - P_n)$ where $P_n^* = n^{-1}\sum_{i=1}^n \delta_{X_{n,i}^*}$, were obtained in [BF81], [GZ90], or [KW92] (see [VdVW96, ST95] for instance for theoretical reviews).
From a practical point of view, taking advantage from the fact that a realization of the bootstrap sample $\mathbf{X}_n^*$ given $\mathbf{X}_n = (x_1, \ldots, x_n)$ can be simulated by simply taking n values with replacement in the set $\{x_1, \ldots, x_n\}$, statisticians often do not strive to compute exactly the probabilistic characteristics of the bootstrapped root R_n^*, but rather approximate them by Monte Carlo procedures. This explains the frequent confusion between the term of bootstrap and the one of resampling, which is more related to the mechanism at stake in the Monte Carlo procedures following the bootstrap estimation (see [BB95] for instance where this point is underlined). If we introduce for every $i = 1 \ldots n$, the random variable $M_{n,i}$ defined as the number of times that X_i appears in the bootstrap sample $\mathbf{X}_n^*$, it is easy to see that the bootstrapped empirical process satisfies:

$$\mathbb{G}_n^* = \sqrt{n}(P_n^* - P_n) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (M_{n,i} - 1)\delta_{X_i},$$

and that any linear root R_n^* can be expressed as a function of $\mathbf{X}_n$ and the $M_{n,i}$'s only. The random vector $M_n = (M_{n,1}, \ldots, M_{n,n})$ which has a multinomial distribution with parameters $(n, n^{-1}, \ldots, n^{-1})$ is viewed as a resampling plan, and the $M_{n,i}{}'s$ as the resampling weights of the bootstrap method.

Starting from this observation, many authors proposed to study other types of resampling weights, and to replace $M_n = (M_{n,1}, \ldots, M_{n,n})$ by any exchangeable random (or not) vector $W_n = (W_{n,1}, \ldots, W_{n,n})$, independent of $\mathbf{X}_n$. This allowed to see some well-known methods such as Fisher's permutation, jackknife, subsampling or cross validation as bootstrap methods (see [Rom89], [Præ95], and [Arl07, Arl09] for more details). This also led to various new types of bootstrap methods such as the m out of n bootstrap introduced by Bretagnolle [Bre83], the Bayesian bootstrap of Rubin [Rub81] and Lo [Lo87], whose resampling weights have a Dirichlet distribution, Weng's [Wen89] bootstrap, and the wild bootstrap whose weights are i.i.d. variables with expectation and variance equal to 1, and which is detailed in [Mam92]. Præstgaard and Wellner [PW93] proved an analogue of Giné and Zinn's theorem from [GZ90] for the general exchangeable weighted bootstrapped empirical process under appropriate conditions on the weights $W_n = (W_{n,1}, \ldots, W_{n,n})$. When the root R_n can be expressed from $\mathbb{G}_n$ exclusively, as linear roots for instance, the weighted bootstrapped root is denoted by $R_n^{W_n}$ and defined replacing $\mathbb{G}_n$ in the expression of R_n by $\mathbb{G}_n^{W_n} = \sqrt{n}\left(P_n^{W_n} - P_n\right)$, with $P_n^{W_n}(t) = n^{-1}\sum_{i=1}^n W_{n,i}t(X_i)$.

About the choice of the exchangeable weights, we refer to the book of Barbe and Bertail [BB95], where an asymptotic analysis of a large family of such exchangeable weights is given, based on Edgeworth expansions.

Weighted bootstrap penalties. From the global minimax point of view, any well-chosen estimator of an upper bound for $\mathbb{E}\left[\sup_{f\in\mathcal{F}_m}\left(-\overline{L_n}(f)\right)\right]$, with expectation of order $\sqrt{V\left(\mathcal{C}_m\right)/n}$, may be an appropriate penalty. Yet

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_m}\left(-\overline{L_n}(f)\right)\right] = \mathbb{E}\left[\sup_{t\in\mathcal{T}_m}(P - P_n)(t)\right],$$

with $\mathcal{T}_m = \left\{ t : \mathbb{X} = \mathbb{Y} \times \{0,1\} \to \{0,1\},\ t(y,z) = \mathbb{1}_{f(y) \neq z},\ f \in \mathcal{F}_m \right\}$

Extrapolating the bootstrap paradigm described above to the sample $\mathbf{X}_n$ in the model $\mathcal{M}^{(1)}_{\text{classification}}$, and considering the root $R_n(\mathbf{X}_n; P) = \sup_{t \in \mathcal{T}_m} (P - P_n)(t)$, a natural weighted bootstrap estimator of the expectation $\mathbb{E}\left[R_n(\mathbf{X}_n; P) \right]$ is

$$\hat{R}_m^{W_n} = \mathbb{E}\left[\sup_{t \in \mathcal{T}_m} (P_n - P_n^{W_n})(t) \middle| \mathbf{X}_n \right] = \mathbb{E}\left[\sup_{f \in \mathcal{F}_m} \frac{1}{n} \sum_{i=1}^{n} \left(1 - W_{n,i} \right) \mathbb{1}_{f(Y_i) \neq Z_i} \middle| \mathbf{X}_n \right],$$

which is introduced in (2.7), as a generalization of the Rademacher quantity $\hat{R}_m$.

Here are considered several kinds of random weights $W_n = (W_{n,1}, \ldots, W_{n,n})$:

- $W_n^{Rad} = 2(B_1, \ldots, B_n)$, where $(B_1, \ldots, B_n)$ is a sample of n i.i.d. Bernoulli random variables with parameter $1/2$, so that $1 - 2B_i$ is a Rademacher variable, which corresponds to the case investigated by Koltchinskii [Kol01] and Bartlett, Boucheron, Lugosi [BBL02],
- $W_n^{Sym} = (W_{n,1}^{Sym}, \ldots, W_{n,n}^{Sym})$ is a sample of n i.i.d. real random variables such that $\left(1 - W_{n,1}^{Sym} \right)$ has a symmetric distribution and such that $\mathbb{E}\left[|W_{n,1}^{Sym}| \right] < +\infty$,
- $W_n^{EB} = (M_{n,1}, \ldots, M_{n,n})$, where $(M_{n,1}, \ldots, M_{n,n})$ is a multinomial random vector with parameters $(n, , n^{-1}, \ldots, n^{-1})$, which corresponds to Efron's bootstrap,
- $W_n^{BB} = \left(V_1/\overline{V_n}, \ldots, V_n/\overline{V_n} \right)$, where $(V_1, \ldots, V_n)$ is a sample of n i.i.d. positive random variables, which corresponds to the Bayesian bootstrap investigated in [Rub81, Lo87, Wen89] when the V_i's are exponential random variables with parameter 1.

Now, the question is whether $\mathbb{E}\left[\sup_{f \in \mathcal{F}_m} \left(-\overline{L_n}(f) \right) \right] = \mathbb{E}\left[\sup_{t \in \mathcal{T}_m} (P - P_n)(t) \right]$ is definitely well estimated by $\hat{R}_m^{W_n}$. Anterior asymptotic results in the literature, and in particular the ones of Giné and Zinn [GZ90] and Præstgaard and Wellner [PW93], were encouraging, but we in fact aimed at obtaining nonasymptotic oracle inequalities and minimax adaptivity properties.

Regarding (2.6), the more fundamental issue from the nonasymptotic point of view is in fact to see how $\hat{R}_m^{W_n}$ is close to the supremum $\sup_{f \in \mathcal{F}_m} \left(-\overline{L_n}(f) \right) = \sup_{t \in \mathcal{T}_m} (P - P_n)(t)$ itself.

To this end, we established in [6] in particular (see also [1] and [2] for less sharp results) new exponential inequalities, that we describe in the next section.

2.2.3 Exponential inequalities

The scope of this section is well beyond the classification framework: here, $\mathbf{X}_n = (X_1, \ldots, X_n)$ denotes any sample of n i.i.d. random variables defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$, with values in $\mathbb{X}$, and with common distribution P on $\mathbb{X}$. Let P_n be the corresponding empirical process defined by $P_n(t) = n^{-1} \sum_{i=1}^{n} t(X_i)$, and $P(t) = \mathbb{E}\left[t(X_1) \right]$, for every measurable function t from $\mathbb{X}$ to $[0,1]$. Let $\mathcal{T}$ be a countable set of measurable functions from $\mathbb{X}$ to $[0,1]$.

We propose here generalizations of the exponential inequality for $\sup_{t \in \mathcal{T}_m} (P - P_n)(t) - \hat{R}_m^{W_n}$ of [Kol01] and [BBL02], based on comparisons of expectations combined with McDiarmid's [McD89] inequality.

Expectation inequality based on symmetrization. Let $W_n = (W_{n,1}, \ldots, W_{n,n})$ be a vector of n i.i.d. random variables such that $(W_{n,1} - 1)$ (or equivalently $(1 - W_{n,1})$) is a real symmetric random variable such that $\mathbb{E}\left[|W_{n,1}| \right] < +\infty$. For every measurable function t from $\mathbb{X}$ to $[0,1]$, we set $P_n^{W_n}(t) = n^{-1} \sum_{i=1}^{n} W_{n,i} t(X_i)$. As explained above, a first point in our analysis was to compare the expectations of the quantities at hand, that is of $\sup_{t \in \mathcal{T}} \left(P - P_n \right)(t)$ and $\mathbb{E}\left[\sup_{t \in \mathcal{T}} \left(P_n - P_n^{W_n} \right)(t) | \mathbf{X}_n \right]$.

The following inequality is obtained via a symmetrization trick developed in the empirical processes theory by Koltchinskii [Kol81], Pollard [Pol82] and especially Giné and Zinn [GZ84].

Introducing some independent copy $\mathbf{X}'_n = (X'_1, \ldots, X'_n)$ of $\mathbf{X}_n$ and denoting by P'_n the corresponding empirical process, one obtains by Jensen's inequality:

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in\mathcal{T}}\left(P-P_n\right)(t)\right] &= \mathbb{E}\left[\sup_{t\in\mathcal{T}}\mathbb{E}\left[\left(P'_n-P_n\right)(t)\,\middle|\,\mathbf{X}_n\right]\right]\\ &\leq \mathbb{E}\left[\sup_{t\in\mathcal{T}}\left(P'_n-P_n\right)(t)\right].\end{aligned}$$

Let $(\varepsilon_1, \ldots, \varepsilon_n)$ be a sample of n i.i.d. Rademacher variables independent of $\mathbf{X}_n$, $\mathbf{X}'_n$ and W_n. Since for any symmetric random variable S independent of ε_1, the variables $\varepsilon_1 S$, $\varepsilon_1|S|$ and S are identically distributed, one gets:

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in\mathcal{T}}\left(P-P_n\right)(t)\right] &\leq \mathbb{E}\left[\sup_{t\in\mathcal{T}}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\left(t(X'_i)-t(X_i)\right)\right]\\ &\leq \frac{2}{n}\mathbb{E}\left[\sup_{t\in\mathcal{T}}\sum_{i=1}^n\varepsilon_i t(X_i)\right]\\ &\leq \frac{2}{n\mathbb{E}\left[|W_{n,1}-1|\right]}\mathbb{E}\left[\sup_{t\in\mathcal{T}}\mathbb{E}\left[\sum_{i=1}^n\varepsilon_i|W_{n,i}-1|t(X_i)\,\middle|\,\varepsilon,\mathbf{X}_n\right]\right]\\ &\leq \frac{1}{n\mathbb{E}\left[\left(W_{n,1}-1\right)_+\right]}\mathbb{E}\left[\sup_{t\in\mathcal{T}}\sum_{i=1}^n\varepsilon_i|W_{n,i}-1|t(X_i)\right].\end{aligned}$$

Using the same symmetrization argument as above finally leads to

$$\mathbb{E}\left[\sup_{t\in\mathcal{T}}\left(P-P_n\right)(t)\right]\leq\frac{1}{\mathbb{E}\left[\left(W_{n,1}-1\right)_+\right]}\mathbb{E}\left[\sup_{t\in\mathcal{T}}\left(P_n-P_n^{W_n}\right)(t)\right]. \tag{2.8}$$

For several other interesting inequalities, we refer the reader to the recent monograph of Giné and Nickl [GN15].

Expectation inequality for the exchangeably weighted bootstrap. Let us here consider a vector $W_n = (W_{n,1}, \ldots, W_{n,n})$ of n exchangeable and nonnegative random variables independent of $\mathbf{X}_n$ and satisfying $\sum_{i=1}^n W_{n,i} = n$, and denote by $P_n^{W_n}$ the corresponding exchangeably weighted bootstrap empirical process defined by $P_n^{W_n}(t) = n^{-1}\sum_{i=1}^n W_{n,i}t(X_i)$. We here aim at obtaining a result similar to (2.8) but since we do not deal with symmetric random variables any more, we here need to replace the symmetrization trick by another argument.
Using Jensen's inequality again allows to obtain:

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in\mathcal{T}}(P-P_n)(t)\right] &\leq \mathbb{E}\left[\sup_{t\in\mathcal{T}}\frac{1}{n}\sum_{i=1}^n\frac{\mathbb{E}\left[(W_{n,i}-1)\mathbb{1}_{W_{n,i}\geq 1}\right]}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}(P(t)-t(X_i))\right]\\ &\leq \frac{1}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\mathbb{E}\left[\sup_{t\in\mathcal{T}}\frac{1}{n}\sum_{i=1}^n(W_{n,i}-1)\mathbb{1}_{W_{n,i}\geq 1}(P(t)-t(X_i))\right].\end{aligned}$$

It is known that if U and V are independent random variables such that for all g in a class of functions $\mathcal{G}$, $\mathbb{E}\left[g(V)\right]=0$, then

$$\mathbb{E}\Big[\sup_{g\in\mathcal{G}}g(U)\Big]\leq\mathbb{E}\Big[\sup_{g\in\mathcal{G}}(g(U)+g(V))\Big]. \tag{2.9}$$

Applying this inequality to the random variables $\sum_{i=1}^n(W_{n,i}-1)\mathbb{1}_{W_{n,i}\geq 1}(P(t)-t(X_i))$ and $\sum_{i=1}^n(W_{n,i}-1)\mathbb{1}_{W_{n,i}<1}(P(t)-t(X_i))$ conditionally given W_n leads to the same inequality as (2.8), that is

$$\mathbb{E}\left[\sup_{t\in\mathcal{T}}(P-P_n)(t)\right]\leq\frac{1}{\mathbb{E}\left[\left(W_{n,1}-1\right)_+\right]}\mathbb{E}\left[\sup_{t\in\mathcal{T}}\left(P_n-P_n^{W_n}\right)(t)\right].$$

General exponential inequality. From the above computations, one can state the following common result.

Proposition 3 (Fromont, 2007)**.** *Let $W_n = (W_{n,1}, \dots, W_{n,n})$ be a vector, independent of $\mathbf{X}_n$, of either:*
- *n i.i.d. random variables such that $(W_{n,1} - 1)$ (or equivalently $(1 - W_{n,1})$) is a real symmetric random variable such that $\mathbb{E}[|W_{n,1}|] < +\infty$ or,*
- *n exchangeable and nonnegative random variables such that $\sum_{i=1}^n W_{n,i} = n$.*

If, for every measurable function t from $\mathbb{X}$ to $[0,1]$, $P_n^{W_n}(t) = n^{-1}\sum_{i=1}^n W_{n,i} t(X_i)$, then

$$\mathbb{E}\left[\sup_{t\in\mathcal{T}}(P - P_n)(t)\right] \leq \frac{1}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\mathbb{E}\left[\sup_{t\in\mathcal{T}}\left(P_n - P_n^{W_n}\right)(t)\right].$$

Combining Proposition 3 with McDiarmid's [McD89] inequality now leads to the following general exponential inequality.

Proposition 4 (Fromont, 2007)**.** *With the notation of Proposition 3, for any $x > 0$, the following inequality holds:*

$$\mathbb{P}\Bigg(\sup_{t\in\mathcal{T}}(P - P_n)(t) - \frac{1}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\mathbb{E}\left[\sup_{t\in\mathcal{T}}\left(P_n - P_n^{W_n}\right)(t)\,\middle|\,\mathbf{X}_n\right] \geq \left(1 + \frac{\mathbb{E}\left[|W_{n,1}-1|\right]}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\right)\sqrt{\frac{x}{2n}}\Bigg) \leq e^{-x}.$$

Note that in the symmetric case, that is when $(W_{n,1} - 1)$ is a real symmetric random variable, $\left(1 + \left(\mathbb{E}\left[|W_{n,1}-1|\right]\right)/\mathbb{E}\left[(W_{n,1}-1)_+\right]\right) = 3$ which simplifies the above inequality. As for the Rademacher case, another concentration inequality is given in [LN11].

2.2.4 Main theoretical results

Let us now come back to the binary classification problem at hand in the model $\mathcal{M}^{(1)}_{\text{classification}}$.
We consider the three possible choices of weights W_n^{Sym}, W_n^{EB} and W_n^{BB} defined in Section 2.2.2. Recall that for W_n equal to any of these W_n^{Sym}, W_n^{EB}, W_n^{BB},

$$\hat{R}_m^{W_n} = \mathbb{E}\left[\sup_{t\in\mathcal{T}_m}(P_n - P_n^{W_n})(t)\,\middle|\,\mathbf{X}_n\right] = \mathbb{E}\left[\sup_{f\in\mathcal{F}_m}\frac{1}{n}\sum_{i=1}^n (1 - W_{n,i})\,\mathbb{1}_{f(Y_i)\neq Z_i}\,\middle|\,\mathbf{X}_n\right].$$

From Proposition 4, we deduce the following result.

Theorem 6 (Fromont, 2007)**.** *Let $(x_m)_{m\in\mathcal{M}}$ be a family of positive numbers such that $\sum_{m\in\mathcal{M}} e^{-x_m} \leq \kappa$, for some constant κ. In the above notation, choose a penalty such that*

$$\text{pen}(m) = \frac{1}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\hat{R}_m^{W_n} + \left(1 + \frac{\mathbb{E}\left[|W_{n,1}-1|\right]}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\right)\sqrt{\frac{x_m}{2n}},$$

for every m in $\mathcal{M}$. The approximate minimum penalized contrast estimator $\hat{f}_{n,\hat{m}}$ satisfies:

$$\mathbb{E}\left[l\left(f^*, \hat{f}_{n,\hat{m}}\right)\right] \leq \inf_{m\in\mathcal{M}}\left\{l\left(f^*, f_{n,m}\right) + \mathbb{E}\left[\text{pen}(m)\right]\right\} + \left(1 + \frac{\mathbb{E}\left[|W_{n,1}-1|\right]}{\mathbb{E}\left[(W_{n,1}-1)_+\right]}\right)\frac{\kappa}{2}\sqrt{\frac{\pi}{2n}} + \rho_n.$$

In order to obtain adaptivity properties from the global minimax point of view, computing a sharp upper bound for $\mathbb{E}\left[\text{pen}(m)\right]$ is now necessary.

When the vector of weights W_n^{Sym} satisfy the following moment assumption

$$\forall k \geq 2,\ \mathbb{E}\left[|W_{n,1}-1|^k\right] \leq \frac{k!}{2} v c^{k-2}, \tag{2.10}$$

for some positive numbers v and c, an upper bound for $\mathbb{E}\big[\hat{R}_m^{W_n^{Sym}}\big]$, and therefore for $\mathbb{E}\left[\mathrm{pen}(m)\right]$, is computed in [6, Theorem 4] (for i.i.d. weights, which are not necessary symmetric), using Haussler's bound [Hau95] for the metric entropy of VC classes and chaining arguments.
Assuming that $n \geq 4$, when W_n^{Sym} satisfies (2.10), there exist positive constants κ_1 and κ_2 such that

$$\mathbb{E}\left[\hat{R}_m^{W_n^{Sym}}\right] \leq \kappa_1 \sqrt{v}\sqrt{\frac{V\left(\mathcal{C}_m\right)}{n}} + \kappa_2 c \frac{V\left(\mathcal{C}_m\right)}{n}\ln^2 n,$$

and the constants v and c can be respectively replaced by 1 and 0 when the weights have the subgaussian behavior: $\mathbb{E}\left[e^{\lambda(1-W_{n,1})}\right] \leq e^{\lambda^2/2}$ for any $\lambda > 0$, which is typically the case of Rademacher weights.
The control of $\mathbb{E}\big[\hat{R}_m^{W_n^{EB}}\big]$ is obtained using [6, Theorem 4] as above, but as the exchangeable weights are not i.i.d. anymore, an additional tool has to be introduced, which is well-known in the empirical process theory. This tool consists in a Poissonization trick, whose idea relies on the following arguments.
For every i in $\{1,\ldots,n\}$, $W_{n,i}$ can be written as $W_{n,i} = \sum_{j=1}^n \mathbb{1}_{U_j \in ((i-1)/n, i/n]}$, where $U = (U_1,\ldots,U_n)$ is a sample of n i.i.d. random variables uniformly distributed on $(0,1)$. If N is a Poisson variable with parameter n independent of $\mathbf{X}_n$ and for every i in $\{1,\ldots,n\}$, $N_i = \sum_{j=1}^N \mathbb{1}_{U_j \in ((i-1)/n, i/n]}$, then the N_i's are i.i.d. with Poisson distribution with parameter 1. As N has mean n, thanks to the concentration principle, its fluctuations are concentrated around n. As a consequence, the N_i's are close to the $W_{n,i}$'s but having the further advantage to be independent.
As for the Bayesian bootstrap, assuming that the V_i's satisfy the moment assumption

$$\forall k \geq 2,\ \mathbb{E}\left[V_i^k\right] \leq \frac{k!}{2} v c^{k-2}, \tag{2.11}$$

Bernstein's concentration inequality, with rather technical computations, allow to obtain a similar control of $\mathbb{E}\big[\hat{R}_m^{W_n^{BB}}\big]$, to lead to the final following result.

Corollary 2 (Fromont, 2007)**.** *Let $n \geq 4$, and W_n be one of the three following vectors of weights.*
— *$W_n = W_n^{Sym}$ satisfying the moment assumption* (2.10)*,*
— *$W_n = W_n^{EB}$,*
— *$W_n = W_n^{BB}$, with some V_i's satisfying the moment assumption* (2.11)*.*
In the notation of Theorem 6, there is $C = C(v, c, \mathbb{E}\left[V_1\right], \mathbb{E}\left[|V_1/\overline{V_n}-1|\right], \mathbb{E}\left[(V_1/\overline{V_n}-1)_+\right])$ such that

$$\mathbb{E}\left[l\left(f^*, \hat{f}_{n,\hat{m}}\right)\right] \leq \inf_{m\in\mathcal{M}}\left\{ l\left(f^*, f_{n,m}\right) + C\left(\sqrt{\frac{V\left(\mathcal{C}_m\right)}{n}} + \frac{V\left(\mathcal{C}_m\right)}{n}\ln^2 n + \sqrt{\frac{x_m}{n}}\right)\right\} + \rho_n.$$

Moreover, in the first case, if the weights in $W_n = W_n^{Sym}$ have a subgaussian behavior, the term $\left(V\left(\mathcal{C}_m\right)\ln^2 n\right)/n$ can be removed in the above inequality.

The above risk upper bound clearly generalizes Koltchinskii [Kol01] and Bartlett, Boucheron, Lugosi's [BBL02] one for Rademacher penalization.
Furthermore, it allows to prove that when ρ_n is smaller than $n^{-1/2}$, and $\ln n \leq V\left(\mathcal{C}_m\right) \leq n/\ln^4 n$ for every m in $\mathcal{M}$, $\hat{f}_{n,\hat{m}}$ satisfies $\left(\mathcal{P}_{\text{adaptive},\overline{\mathcal{F}}}\right)$ over the whole collection of models $\overline{\mathcal{F}} = \{\mathcal{F}_m,\ m \in \mathcal{M}\}$.
The constant $1/\mathbb{E}\left[\left(W_{n,1}-1\right)_+\right]$ in front of $\hat{R}_m^{W_n}$ in the penalty was discussed in [6] as a possible pessimistic choice due to technical reasons. Regarding the asymptotic theory of [GZ90] and [PW93], better choices for the constant seemed to be 1 in the Rademacher case W_n^{Rad} , as well as in Efron's bootstrap case W_n^{EB}, $\left(\mathbb{E}\left[V_1^2\right]/\mathrm{Var}\left[V_1\right]\right)^{-1/2}$ in the Bayesian bootstrap case W_n^{BB}. This conjecture

seemed to be empirically confirmed by the simulation study (see also [Loz00] for the Rademacher case).
Arlot furthermore focused on this particular issue in [Arl07, Chapter 9] and proved that under very strong assumptions on the distribution P (only satisfied in an unrealistic toy framework), the constant $1/\mathbb{E}\left[(W_{n,1}-1)_+\right]$ can indeed be relaxed in the result of Proposition 3. Besides, some examples are exhibited for which the constant cannot be relaxed, which let us think that there is here a real gap between the asymptotic and the nonasymptotic theory, that has to be understood, and that the precise calibration of the constant is a real difficult and, up to our knowledge, still open question.

2.2.5 Experimental results

A simulation study has been performed in [6] in order to investigate the above general bootstrap penalization from a practical point of view. As Lozano [Loz00] and Bartlett, Boucheron and Lugosi [BBL02], this study focuses on the issue which is usually referred to as the intervals model selection problem and which can be described as follows.
Let us set for all u, v in $\mathbb{N}$, $u \leq v$, $[\![u,v]\!] = [u,v] \cap \mathbb{N}$. We consider $\mathbb{Y} = \{1,\dots,2^N\}$ and some partition $\{[\![u_l,v_l]\!],\ l \in \mathcal{L}\}$ of $\mathbb{Y}$. Let Y be a random variable uniformly distributed on $\mathbb{Y}$ and Z a $\{0,1\}-$valued random variable such that $\mathbb{P}(Z=1|Y\in S_0) = 1/2+h$ and $\mathbb{P}(Z=1|Y\notin S_0) = 1/2-h$, where h is some margin parameter in $(0,1/2)$ and $S_0 = \cup_{l\in\mathcal{L}_0\subset\mathcal{L}}[\![u_l,v_l]\!]$. Then the target is the piecewise constant function defined by $f^*(y) = \mathbb{1}_{S_0}(y)$ for y in $\mathbb{Y}$. Two cases are considered. The first one is based on regular partitions of $\mathbb{Y}$ such that $S_0 = \cup_{k\in\{2p+1,\ p\in\mathbb{N},\ 2p+1\leq 2^{J_0}-1\}}[\![(k-1)2^{N-J_0}+1, k2^{N-J_0}]\!]$, with $N=8$, $J_0=2$, $h=0.05$ first, $N=8$, $J_0=6$, $h=0.1$ then. The second one is based on some irregular partition of $\mathbb{Y}$ such that $S_0 = \cup_{k\in\{2p+1,\ p\in\mathbb{N},\ 2p+1\leq 2^{J_0}-1\}}[\![U_{k-1},U_k]\!]$, with $U_0=1$ and $U_1,\dots,U_{2^{J_0}-1}$ randomly chosen on $\{1,\dots,2^N-1\}$ in such a way that $1\leq U_1<\ldots<U_{2^{J_0}-1}<2^N$.
In the first case, the collection of models is chosen such that $\mathcal{M} = \{2,2^2,\dots,2^N\}$ and for $m=2^J$ in $\mathcal{M}$,

$$\mathcal{F}_{2^J} = \left\{ f:\mathbb{Y}\to\{0,1\},\ f = \sum_{k=1}^{2^J} c_k \mathbb{1}_{[\![(k-1)2^{N-J}+1,k2^{N-J}]\!]},\ c_1,\dots,c_{2^J}\in\{0,1\} \right\}.$$

The sequence $(x_m)_{m\in\mathcal{M}}$ is chosen as $x_m = \ln m$ for all m in $\mathcal{M}$.
In the second case, the collection of models is chosen so that it contains, for each complexity D in $\{2,2^2,\dots,2^N\}$, all the models based on the partitions of $\mathbb{Y}$ with D pieces.
The principle of the algorithm used to compute the ERM classifiers as well as Monte Carlo approximations of the bootstrap penalties is detailed in [6].

The results obtained for the estimated risks of the penalized ERM classifiers (that is the approximate penalized contrast estimators with $\rho_n = 0$) for sample sizes varying in $[\![200,2000]\!]$ as well as percentages of good model (or complexity) selection give preference to Rademacher and Efron's bootstrap penalization in complex problems of classification (based on a partitions with many pieces), to the Bayesian bootstrap with some V_i's whose distribution is Gamma with parameter 4 in simple problems (based on a regular partition with very few pieces).

In all the study, as explained above, special attention is paid on the multiplicative constant in front of $\hat{R}_m^{W_n}$ involved in the penalty term, looking at the ratio $\mathbb{E}\left[\hat{R}_m^{W_n}\right]/\mathbb{E}\left[\mathrm{pen}_{id}(m)\right]$, where $\mathrm{pen}_{id}(m) = \sup_{f\in\mathcal{F}_m}\left(-\overline{L_n}(f)\right) = \sup_{t\in\mathcal{T}_m}(P-P_n)(t)$. In nearly all the studied cases, the conjecture that the constant $1/\mathbb{E}\left[(W_{n,1}-1)_+\right]$ in the penalty should be replaced by 1 in the symmetric and Efron's bootstrap cases, by $\left(\mathbb{E}\left[V_1^2\right]/\mathrm{Var}\left[V_1\right]\right)^{1/2}$ in the Bayesian bootstrap case was confirmed, but this goes against the above theoretical study and Arlot's developments in [Arl07, Chapter 9]. Some mild assumptions on the distribution P and the chosen collection of models that would be satisfied in the present studied cases, and under which the constant could be relaxed, should therefore be possible, but this remains an open question.

2.2.6 Posterior works

We considered in [1], [2] and [6] the binary classification problem from the global minimax point of view, and therefore we only evaluated the relevance of the studied data-driven penalties as estimators for the global ideal penalty $\mathrm{pen}_{id}(m)$. It is now well known that this global ideal penalty is not the ideal choice when one considers the optimality from the minimax point of view under margin assumptions. Recall that the challenge of model selection by penalization is to determine a penalty such that the risk of $\hat{f}_{n,\hat{m}}$ is upper bounded by $\inf_{m\in\mathcal{M}} \mathbb{E}\big[l\big(f^*, \hat{f}_{n,m}\big)\big]$, or even better by $\mathbb{E}\big[\inf_{m\in\mathcal{M}} l\big(f^*, \hat{f}_{n,m}\big)\big]$ (for discussions on different kinds of oracle inequalities, see e.g., [Arl07] where stronger pathwise oracle inequalities are obtained), up to a multiplicative constant ideally close to 1.
In the same notation as in the above sections, by definition, for any m in $\mathcal{M}$,

$$l\left(f^*, \hat{f}_{n,\hat{m}}\right) = l\left(f^*, \hat{f}_{n,m}\right) + \overline{L_n}\left(\hat{f}_{n,m}\right) - L_n\left(\hat{f}_{n,m}\right) - \overline{L_n}\left(\hat{f}_{n,\hat{m}}\right) + L_n\left(\hat{f}_{n,\hat{m}}\right).$$

It is easy to see that

$$l\left(f^*, \hat{f}_{n,\hat{m}}\right) \leq l\left(f^*, \hat{f}_{n,m}\right) + \overline{L_n}\left(\hat{f}_{n,m}\right) + \mathrm{pen}(m) - \overline{L_n}\left(\hat{f}_{n,\hat{m}}\right) - \mathrm{pen}(\hat{m}) + \rho_n/2. \tag{2.12}$$

The ideal, but unknown, penalty when the aim is to obtain strong oracle inequalities or minimax adaptivity properties (under margin assumptions) is therefore $\mathrm{pen}_{id}^{loc}(m) = -\overline{L_n}\big(\hat{f}_{n,m}\big)$. Note that $\mathrm{pen}_{id}(m)$, which was estimated in our work, is a (rough) upper bound for $\mathrm{pen}_{id}^{loc}(m)$. Any good estimator of $\mathrm{pen}_{id}^{loc}(m)$ itself instead of its upper bound $\mathrm{pen}_{id}(m)$ may therefore be a better penalty choice. Such estimators are obtained via the local Rademacher complexities or averages which are now widely used in the statistical learning theory (see for instance [KP99, BMP04, LW04, BBM05, Kol06]). Arlot's work [Arl07, Arl09] complete the picture with new local penalties, which are easier to compute and calibrate, based on more general bootstrap schemes than the ones considered in our work, including several cross validation approaches.
Some of these local bootstrap penalization approaches are proved to lead to minimax adaptivity properties under margin conditions such as the ones described in Section 2.1.1, as well as different versions of the general margin assumption introduced by [Kol06] (see [Arl07, Arl09] and especially [AB11]) in various learning frameworks.
An alternative to these approaches to obtain minimax adaptivity under margin assumptions is the aggregation of estimators (see e.g., [Tsy04, AT07, Lec07a] and all the recent work of Lecué and co-authors). In some cases, aggregation of estimators outperforms model selection from a specific minimax point of view (see [Lec07b]), but let us remark that the oracle inequalities obtained there are weaker than the ones generally achieved in model selection. Interesting discussions about this topic can also be found in [Arl07, AB11].
The question of the constant calibration in the penalties is still challenging, though there have been considerable advances for a few years, notably around the slope heuristics (see [Mas07, AM09] for instance).

2.3 Functional classification under margin assumptions

In this section, we still consider the binary classification problem in the model $\mathcal{M}_{\text{classification}}^{(1)}$, but focusing on the case where $\mathbb{Y}$ is an infinite dimensional, or functional, separable space. This particular case fits with many real-world applications where the data are more accurately represented by discretized functions than by standard vectors.
Historically, the most simple and popular classifiers in this context are plug-in ones such as kernel and k-Nearest Neighbors (k-NN) rules. Although these rules are known to be universally consistent and even strongly universally consistent when $\mathbb{Y}$ is finite dimensional (see [CH67, Sto77, DGKL94] and [DGL96, GKKW02] for overviews), this is not so clear when $\mathbb{Y}$ is a functional space, and many

papers have been devoted to this issue in the last fifteen years (see for instance the monograph of Ferraty and Vieu [FV06], [BCG10], and references therein). Universal consistency results are usually established under regularity assumptions on the regression function η or under some assumption on small balls probabilities for instance. In a recent note, Azaïs and Fort [AF13] proved that some of these assumptions on small balls probabilities in fact imply that $\mathbb{Y}$ has a finite Haussdorff dimension. Anyway, it is clear that direct approaches, such as kernel or k-NN rules straightforwardly applied on the functional data themselves, suffer from the phenomenon commonly referred to as the curse of dimensionality, and are therefore not expected to achieve good rates of convergence. To overcome this difficulty, most of the traditional effective methods for $\mathbb{R}^d$-valued data analysis have been adapted to handle functional data under the general name of Functional Data Analysis. A key reference on this topic is the series of books by Ramsay and Silverman [RS02, RS05].

The work presented in this section, in collaboration with Christine Tuleau, followed a paper by Biau, Bunea and Wegkamp [BBW05], where the authors propose to filter the functional data Y_i in the Fourier basis and to apply the k-Nearest Neighbors rule to the first d coefficients of the expansion. The choice of the dimension d and the number of neighbors k is made automatically by minimization of a penalized empirical classification error, performed after some hold-out device. The resulting classifier satisfies an oracle type inequality, and hence is universally consistent. As noted by the authors, similar results could be obtained for other universally consistent classification procedures in finite dimension. In this spirit, the Support Vector Machines are investigated by Rossi and Villa [RV05].

The approach of Biau, Bunea and Wegkamp [BBW05] had however let two issues unsolved. The authors indeed underlined their preferring to implement the procedure based on the minimization of the empirical classification error without penalization, but they did not give any theoretical justification of this choice. They also pointed out the problem of instability of the hold-out device, which is well-known by practitioners (see e.g., [HMLW02] for the related question of bandwidth selection in local linear regression smoothers). The authors propose to use a cross validation technique to overcome this problem, but still without any theoretical background.

In [3], both issues were addressed. We first proposed a theoretical justification for the improvements observed when choosing to minimize the nonpenalized empirical classification error rather than the penalized one. The result actually shows that the penalty of order $n^{-1/2}$ considered in [BBW05] is too large to obtain minimax adaptivity under margin assumptions. This suggests to take a penalty equal to zero, or possibly of order smaller than $n^{-1/2}$. Then, an experimental method of stabilization for the hold-out approach is proposed, with illustrations on simple simulated data, as well as real data coming from speech recognition or food industry contexts.

2.3.1 Functional classification via (non)penalized criteria

Let us first present the classification approach of Biau, Bunea and Wegkamp [BBW05], which is used all along our work.

The input data space $\mathbb{Y}$ is assumed to be an infinite dimensional separable space, equipped with a complete system denoted by $\{\psi_j,\ j \in \mathbb{N}\setminus\{0\}\}$. For every i in $\{1,\dots,n\}$, Y_i can thus be expressed as a series expansion $Y_i = \sum_{j=1}^{\infty} Y_{i,j}\psi_j$ and for d in $\mathbb{N}\setminus\{0\}$, we set $\mathbf{Y}_i^d = (Y_{i,1},\dots,Y_{i,d})$. In the same way, $\mathbf{y}^d$ denotes the first d coefficients in the expansion of any new element y in $\mathbb{Y}$. The procedure developed in [BBW05] is described as follows.

— The data are split into a training set $\mathbf{X}^{(\mathcal{T}_l)} = \{X_i = (Y_i, Z_i),\ i \in \mathcal{T}_l\}$ of length l ($1 \le l \le n-1$) and its associated validation set $\mathbf{X}^{(-\mathcal{T}_l)} = \{X_i = (Y_i, Z_i),\ i \notin \mathcal{T}_l\}$ of length $(n-l)$.

— For each k in $\{1,\dots,l\}$, d in a subset $\mathcal{D}$ of $\mathbb{N}\setminus\{0\}$, let $\hat{p}_{l,k,d}$ be the k-Nearest Neighbors rule on $\mathbb{R}^d$ constructed from the set $\{(\mathbf{Y}_i^d, Z_i),\ i \in \mathcal{T}_l\}$. Let $\mathbf{y}$ be an element of $\mathbb{R}^d$. The set $\{(\mathbf{Y}_i^d, Z_i),\ i \in \mathcal{T}_l\}$ is reordered according to increasing Euclidean distances $\|\mathbf{Y}_i^d - \mathbf{y}\|_2$, and the reordered variables are denoted by $(\mathbf{Y}_{(1)}^d(\mathbf{y}), Z_{(1)}(\mathbf{y})),\dots,(\mathbf{Y}_{(l)}^d(\mathbf{y}), Z_{(l)}(\mathbf{y}))$. Thus $\mathbf{Y}_{(k)}^d(\mathbf{y})$ is the k-th nearest neighbor of $\mathbf{y}$ amongst $\{\mathbf{Y}_i^d,\ i \in \mathcal{T}_l\}$. When $\|\mathbf{Y}_{i_1}^d - \mathbf{y}\|_2 = \|\mathbf{Y}_{i_2}^d - \mathbf{y}\|_2$, $\mathbf{Y}_{i_1}^d$

is declared closer to $\mathbf{y}$ if $i_1 < i_2$ (the indexes rule in case of equality). Then $\hat{p}_{l,k,d}(\mathbf{y})$ is defined by

$$\hat{p}_{l,k,d}(\mathbf{y}) = \begin{cases} 0 & \text{if } \sum_{i=1}^{k} \mathbb{1}_{Z_{(i)}(\mathbf{y})=0} \geq \sum_{i=1}^{k} \mathbb{1}_{Z_{(i)}(\mathbf{y})=1} \\ 1 & \text{otherwise.} \end{cases}$$

We introduce the corresponding functional classifier defined by

$$\hat{f}_{l,k,d}(y) = \hat{p}_{l,k,d}(\mathbf{y}^d) \text{ for all } y \text{ in } \mathbb{Y}.$$

— The appropriate k and d are simultaneously selected from the validation set by minimizing a penalized empirical classification error:

$$(\hat{k}, \hat{d}) = \text{argmin}_{k\in\{1,\ldots,l\},\ d\in\mathcal{D}} \left\{ \frac{1}{m} \sum_{i\in\{1,\ldots,n\}\setminus\mathcal{T}_l} \mathbb{1}_{\hat{f}_{l,k,d}(Y_i)\neq Z_i} + \text{pen}(d) \right\}, \tag{2.13}$$

where $\text{pen}(d)$ is a positive penalty term that can be equal to zero.

— The final classifier is defined by

$$\hat{f}_n(y) = \hat{f}_{l,\hat{k},\hat{d}}(y) \text{ for all } y \text{ in } \mathbb{Y}. \tag{2.14}$$

Considering $\mathcal{D} = \mathbb{N} \setminus \{0,\}$, a family $\{w_d,\ d \in \mathcal{D}\}$ of positive numbers such that $\sum_{d\in\mathcal{D}} e^{-w_d} \leq \kappa$ for some constant κ, and a penalty of the form $\text{pen}(d) = \sqrt{w_d/(2(n-l))}$ (penalized case), Biau, Bunea and Wegkamp [BBW05] proved the following oracle type result. When $l > 1/\kappa$, there exists $C(\kappa)$ such that the classifier $\hat{f}_n$ defined by (2.14) satisfies

$$\mathbb{E}\left[L\left(\hat{f}_n\right)\right] - L\left(f^*\right) \leq \inf_{d\in\mathcal{D}} \left\{ L_d^* - L\left(f^*\right) + \inf_{1\leq k\leq l} \left\{ \mathbb{E}\left[L\left(\hat{f}_{l,k,d}\right)\right] - L_d^* \right\} + \text{pen}(d) \right\} + C(\kappa)\sqrt{\frac{\ln l}{n-l}}, \tag{2.15}$$

where L_d^* is the minimal classification error when the feature space is $\mathbb{R}^d$. The authors notify that the same result holds when $\hat{f}_n$ is defined by (2.14) with $\mathcal{D} = \{1,\ldots,d_n\}$ and $\text{pen}(d) = 0$ (nonpenalized case), but at the price that the last term $C(\kappa)\sqrt{\ln l/(n-l)}$ is replaced by $C(\kappa)\sqrt{\ln l d_n/(n-l)}$.
The quantity $L_d^* - L\left(f^*\right)$ can be viewed as an approximation term, which tends to 0 as d tends to $+\infty$. Some classical martingale arguments allow to see this. Moreover, from Stone's [Sto77] consistency result in $\mathbb{R}^d$, one deduces that for d in $\mathbb{N} \setminus \{0\}$, $\mathbb{E}[L(\hat{f}_{l,k,d})]$ tends to L_d^* as $l \to \infty$, $k \to \infty$, $k/l \to 0$ whatever the distribution P. The classifier $\hat{f}_n$ is thus universally consistent.
From the minimax point of view, (2.15) also allows to see that the risk of $\hat{f}_n$ nearly achieves the global minimax risk over any set of distributions based on a VC class. However, (2.15) is not sufficient to see whether the risk of $\hat{f}_n$ achieves the minimax risk under margin assumptions. On the one hand, the order of magnitude of the penalty term $(\sqrt{w_d/(2(n-l))})$ in the penalized case is too large as compared to the rates faster than $1/\sqrt{n-l}$ that are expected under margin assumptions. On the other hand, in the nonpenalized case, the right hand side of the inequality (2.15) makes a term of order $\sqrt{\ln l d_n/(n-l)}$ appear. This term can not be seen as a residual term when considering any margin assumption anymore, and hence the oracle type inequality (2.15) had to be refined.

The key point in the proof of the result (2.15) of Biau, Bunea and Wegkamp is the following inequality, which is similar to (2.12), but adapted to the present context. Setting $L_{n,l}(f) = (n-l)^{-1} \sum_{i\in\{1,\ldots,n\}\setminus\mathcal{T}_l} \mathbb{1}_{f(Y_i)\neq Z_i}$ for all measurable function $f : \mathbb{Y} \to \{0,1\}$, one has

$$L\left(\hat{f}_n\right) - L\left(f^*\right) \leq L\left(\hat{f}_{l,k,d}\right) - L\left(f^*\right) + \text{pen}(d) - \text{pen}(\hat{d}) + L\left(\hat{f}_{l,\hat{k},\hat{d}}\right) - L_{n,l}\left(\hat{f}_{l,\hat{k},\hat{d}}\right) - L\left(\hat{f}_{l,k,d}\right) + L_{n,l}\left(\hat{f}_{l,k,d}\right). \tag{2.16}$$

Since $L(\hat{f}_{l,k,d}) - L_{n,l}(\hat{f}_{l,k,d})$ is centered, (2.15) is obtained by choosing a penalty such that $\mathrm{pen}(\hat{d})$ is large enough to compensate for the quantity $L(\hat{f}_{l,\hat{k},\hat{d}}) - L_{n,l}(\hat{f}_{l,\hat{k},\hat{d}})$, but such that $\mathrm{pen}(d)$ is small enough (of order at most $1/\sqrt{n-l}$) to fit the minimax risk bounds in the global case. The main issue is then to evaluate the fluctuations of $L(\hat{f}_{l,\hat{k},\hat{d}}) - L_{n,l}(\hat{f}_{l,\hat{k},\hat{d}})$, and to this end, Hoeffding's concentration inequality is used in [BBW05].

In [3], we propose to use instead of Hoeffding's inequality a Bernstein type inequality which allows to control the fluctuations of the whole quantity

$$L\left(\hat{f}_{l,\hat{k},\hat{d}}\right) - L_{n,l}\left(\hat{f}_{l,\hat{k},\hat{d}}\right) - L\left(\hat{f}_{l,k,d}\right) + L_{n,l}\left(\hat{f}_{l,k,d}\right),$$

by taking its variance into account. We then prove the following result.

Proposition 5 (Fromont, Tuleau-Malot, 2006)**.** *Assume that $n \geq 2$ and let $\hat{f}_n$ be the classifier defined by (2.14) with a finite subset $\mathcal{D}$ of $\mathbb{N} \setminus \{0\}$ and with penalty terms $\mathrm{pen}(d)$ that can be equal to zero. For any $\varepsilon > 0$, if GMA(θ, h) holds with $\theta \geq 1$ and h in $[0,1]$, then*

$$\mathbb{E}\left[L\left(\hat{f}_n\right) - L\left(f^*\right)\middle|\mathbf{X}^{(\mathcal{T}_l)}\right] \leq (1+\varepsilon) \inf_{k\in\{1,\dots,l\},\ d\in\mathcal{D}} \left\{L\left(\hat{f}_{l,k,d}\right) - L\left(f^*\right) + \mathrm{pen}(d)\right\} + C(\varepsilon)\frac{1+\ln\left(l\#\mathcal{D}\right)}{((n-l)h)^{\frac{\theta}{2\theta-1}}}. \quad (2.17)$$

Taking $pen(d) = 0$, the obtained result proves the efficiency of the classifier in its nonpenalized version with a risk upper bound of smaller order than $(n-l)^{-1/2}$. It thus gives a theoretical justification to the conjecture of Biau, Bunea and Wegkamp that the nonpenalized classifier has better performance that the penalized one with a penalty term of order $(n-l)^{-1/2}$.
Note that the last term in the right hand side of the inequality (2.17) is at most of the same order as the upper bound of (2.1), so it may be viewed as a residual term. Therefore, considering the classifier $\hat{f}_n$ in its penalized form, but with a penalty small enough, that is with a smaller order of magnitude than this residual term or than $\inf_{1\leq k\leq l}\left\{\mathbb{E}\left[L\left(\hat{f}_{l,k,d}\right)\right] - L_d^*\right\}$, the order of the risk bound is not altered. Kohler and Krzyżak [KK06] have proved that under some local Lipschitz condition on the regression function η, assuming that the margin condition MA(α) is satisfied,

$$\inf_{1\leq k\leq l}\left\{\mathbb{E}\left[L\left(\hat{f}_{l,k,d}\right)\right] - L_d^*\right\} \leq C(\ln(n-l))^{\frac{2(1+\alpha)}{d}}(n-l)^{-\frac{1+\alpha}{2+d}}.$$

This allows to consider penalties such as $\mathrm{pen}_0(d) = 0$, $\mathrm{pen}_1(d) = \ln d/(n-l)$, or $\mathrm{pen}_2(d) = \sqrt{d}/(n-l)$ for instance. Some experimental results are presented in [3] in order to see the effect of each of these various penalties on the performance of the classifiers on the one hand, and on the stability of the approach (with respect to the data-splitting device) on the other hand.
Remark however that there is, up to our knowledge, no lower bound for the minimax risk which would exactly guarantee that the classifier $\hat{f}_n$ is optimal from the minimax point of view under the margin assumption GMA(θ, h).

2.3.2 Experimental results

As explained above, we investigate from a practical point of view in [3] the classifier $\hat{f}_n$ defined by (2.14) with a penalty equal to pen_0 pen_1, pen_2, or the penalty $\mathrm{pen}_B(d) = \ln d/\sqrt{n-l}$ proposed by Biau, Bunea and Wegkamp [BBW05].
The performance of $\hat{f}_n$ is evaluated by estimating the classification error as follows: the data (of size m) are randomly split into three parts of respective sizes $l = m/4$, $n-l = m/2$ and $m/4$. The first part is used as training data set $\mathbf{X}^{(\mathcal{T}_l)}$ to construct the collection of classifiers $\{\hat{f}_{l,k,d},\ k \in \{1,\dots,l\},\ d \in \mathcal{D}\}$, the second part is used as validation data set $\mathbf{X}^{(-\mathcal{T}_l)}$ to select $\hat{k}$ and $\hat{d}$, and the third one is used to

estimate the classification error. The performance of $\hat{f}_n$ is further studied in terms of stability of the selected dimension $\hat{d}$.
For our experiments, both real data coming from a speech recognition problem and food industry problems and simple simulated data are used.
Assuming that $\mathbb{Y} = \mathbb{L}_2(\mathbb{R}, \lambda)$, the Fourier basis is chosen for the complete system $\{\psi_j,\ j \in \mathbb{N}\setminus\{0\}\}$. For each input data Y_i, the coefficients of the Fourier series expansion are evaluated using a Fast Fourier Transform.
The data coming from the speech recognition problem were created by Biau and considered in [BBW05]. The data set coming from the food industry, is the Tecator dataset, available on the StatLib repository. This data set is now well-known and often used by the machine learning community. The simulated data are described in [Tul05].

The conclusions are the same for all the experiments. Our study confirms that the penalty pen_B first proposed in the theoretical results of [BBW05] is not relevant since the estimated risk of the corresponding classifier is always significantly larger than the risk of the other ones. Moreover, the selected dimension $\hat{d}$ with pen_B is always very small ($\hat{d} = 1$ for more than 90% of the experiments), which corroborates the idea that this penalty is too heavy, and that the nonpenalized procedure will be more appropriate. However, a refined study of the other penalization schemes shows that it can be interesting to consider procedures with some penalties of small order. Using some penalized procedure with pen_1 or pen_2 indeed improves the stability of the dimension selection process, whereas it does not alter the risk too much.

From the posterior work of Arlot on V-fold and hold-out penalization (see [Arl07, AL15]), we could think that a better solution to stabilize the procedure would be to replace the hold-out device by a classical V-fold cross validation procedure or a cross validation penalization approach.
Interesting related works are the ones of Delaigle and Hall [DH12], where linear plug-in rules are investigated, and [HPS08], where a bootstrap choice for k is proposed.

Chapter 3

Two-sample problems

3.1 Introduction

Close links are now established between the problem of binary classification presented in Chapter 2 and the two-sample problem in the density model, usually referred to as the problem of testing homogeneity. Indeed, homogeneity tests may be integrated in some clustering or outliers detection procedures for instance, and conversely, binary classification procedures can be used to construct homogeneity tests, as explained in [GBR+12, Remark 20].

The issue of two-sample problems is thus central in my recent research, as it actually combines knowledge from the minimax adaptive testing field, with usual methods in binary classification such as kernel and k-Nearest Neighbors ones, and nonasymptotic bootstrap approaches, that I studied separately during or in the years following my PhD.

The present chapter is devoted to this issue, that is to problems of testing the null hypothesis that two independent sets of random variables are equally distributed. Three different models are considered: the independent sets of random variables are either sets of i.i.d. random variables from a density model such as $\mathcal{M}^{(1)}_{\text{density}}$, or sets of independent random variables from a heteroscedastic regression model, or inhomogeneous Poisson processes like in $\mathcal{M}^{(1)}_{\text{Poisson}}$.
Many articles deal with the two-sample problem in the density model, from the historical tests of Kolmogorov-Smirnov, Cramer von Mises, Wald and Wolfovitz [WW40] and their extensions, to the more recent tests in [GBR+08, GFHS10, GBR+12, SSGF13, SFG+10], based on statistical learning kernel methods. As for the problem of testing the equality of two signals in nonparametric regression, among many other papers, one can cite the ones by Hall and Hart [HH90], King et al. [KHW91], or Franke and Halim [FH07]. Note that most of signal detection tests can also be used to this purpose, and this is in particular the case of the tests by Durot and Rozenholc [DR06] and Arlot, Blanchard, and Roquain [ABR10]. When inhomogeneous Poisson processes are considered, Bovett and Saw [BS80] and Deshpande et al. [DMNN99] respectively propose conditional and unconditional tests for the two-sample problem for a restrictive alternative hypothesis.

We present here some works in collaboration with Béatrice Laurent and Patricia Reynaud-Bouret [10], with Béatrice Laurent, Matthieu Lerasle, and Patricia Reynaud-Bouret [9], and with Christine Tuleau-Malot [16], where new aggregated tests, based on either kernel or k-Nearest Neighbors methods, combined with nonasymptotic permutation or bootstrap approaches, are proposed.
Thus, although the three models considered in these works are of course different and dedicated to different applications, the developed tests are all based on common general ideas, that we describe in this introduction.

A general two-sample problem can be expressed as follows.

Let $\mathbf{X} = (\mathbf{X}^1, \mathbf{X}^2)$ be a pair of independent sets of random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and observed on a measurable space $\mathbb{X}$, whose - possibly random - cardinalities are respectively denoted by N_1 and N_2, and whose distributions respectively depend on unknown real valued functions f_1 and f_2 in some linear space $\mathcal{F}$. In the following, P_{f_1,f_2} denotes the joint distribution of $\mathbf{X} = (\mathbf{X}^1, \mathbf{X}^2)$, and $\mathbb{E}_{f_1,f_2}$ the expectation with respect to P_{f_1,f_2}.
The corresponding two-sample problem is defined as the problem of testing

$$(H_0) \quad f_1 = f_2 \quad \text{against} \quad (H_1) \quad f_1 \neq f_2,$$

from the observation of $\mathbf{X} = (\mathbf{X}^1, \mathbf{X}^2)$, with $\mathbf{X}^1 = \{X_1^1, \ldots, X_{N_1}^1\}$ and $\mathbf{X}^2 = \{X_1^2, \ldots, X_{N_2}^2\}$.

For any event $\mathcal{E}$ based on $\mathbf{X}$, $\mathbb{P}_{(H_0)}(\mathcal{E})$ then denotes as usual $\sup_{(f_1,f_2),\ f_1=f_2} P_{f_1,f_2}(\mathcal{E})$. The pooled set $\bar{\mathbf{X}} = \mathbf{X}^1 \cup \mathbf{X}^2$ is of utmost importance in the following. Its cardinality is denoted by $N = N_1 + N_2$, and its elements by $\{X_1, \ldots, X_N\}$.

3.1.1 Nonasymptotic minimax adaptivity

The point of view that is adopted here to evaluate the considered tests is still mainly nonasymptotic, and, in the Poisson framework, based on minimax adaptivity properties.
So, given a first kind error level α in $(0,1)$, any of our tests ϕ based on $\mathbf{X}$ is firstly required to be of level α, that is to satisfy the property (2), that can also be expressed with the present notation as

$$(\mathcal{P}_{\text{level},\alpha}) \qquad \Big| \ \mathbb{P}_{(H_0)}(\phi = 1) \leq \alpha.$$

Then, if possible, given a second kind error level β in $(0,1)$, our tests are secondly required to achieve, over several classes of alternatives simultaneously, the minimax separation rate, whose definition given below is deduced from Baraud's [Bar02] one, but adapted to the present two-sample problem.

Definition 5 (Uniform and minimax separation rate)**.** Let d be a metric over $\mathcal{F}$, and a subset $\mathcal{F}_1^{(2)}$ of $\mathcal{F}^2$. Let α and β be fixed in $(0,1)$, and a test ϕ_α of (H_0) against (H_1) satisfying $(\mathcal{P}_{\text{level},\alpha})$.
The uniform separation rate of ϕ_α over $\mathcal{F}_1^{(2)}$ with prescribed second kind error rate β, for the metric d, is defined by

$$\mathrm{SR}_d^\beta\left(\phi_\alpha, \mathcal{F}_1^{(2)}\right) = \inf\left\{ r > 0, \sup_{(f_1,f_2)\in\mathcal{F}_1^{(2)},\ d(f_1,f_2)\geq r} P_{f_1,f_2}(\phi_\alpha = 0) \leq \beta \right\}.$$

The corresponding minimax separation rate over $\mathcal{F}_1^{(2)}$ with prescribed error rates α and β, for the metric d, is defined by

$$m\mathrm{SR}_d^{\alpha,\beta}\left(\mathcal{F}_1^{(2)}\right) = \inf_{\phi_\alpha \text{ satisfying } (\mathcal{P}_{\text{level},\alpha})} \mathrm{SR}_d^\beta\left(\phi_\alpha, \mathcal{F}_1^{(2)}\right),$$

where the infimum is taken over all possible level α tests.

Definition 6 (Minimax (adaptive) test)**.** Let d be a metric over $\mathcal{F}$, and a collection $\overline{\mathcal{F}_1^{(2)}}$ of subsets $\mathcal{F}_1^{(2)}$ of $\mathcal{F}$. A level α test ϕ_α is said to be minimax over a class $\mathcal{F}_1^{(2)}$ of the collection $\overline{\mathcal{F}_1^{(2)}}$ for the metric d if $\mathrm{SR}_d^\beta(\phi_\alpha, \mathcal{F}_1^{(2)})$ achieves $m\mathrm{SR}_d^{\alpha,\beta}(\mathcal{F}_1^{(2)})$, possibly up to a multiplicative constant depending on α and β. It is said to be minimax adaptive over $\overline{\mathcal{F}_1^{(2)}}$ if $\mathrm{SR}_d^\beta(\phi_\alpha, \mathcal{F}_1^{(2)})$ achieves, or nearly achieves, $m\mathrm{SR}_d^{\alpha,\beta}(\mathcal{F}_1^{(2)})$, for every $\mathcal{F}_1^{(2)}$ in $\overline{\mathcal{F}_1^{(2)}}$ simultaneously, without knowing in advance to which class of the collection (f_1, f_2) may belong. This property is formalized in the following as

$$\left(\mathcal{P}_{\text{adaptive},\alpha,\beta,\overline{\mathcal{F}_1^{(2)}},d}\right) \ \Big| \ \mathrm{SR}_d^\beta(\phi_\alpha, \mathcal{F}_1^{(2)}) \text{ achieves or nearly achieves } m\mathrm{SR}_d^{\alpha,\beta}(\mathcal{F}_1^{(2)}), \text{ for every } \mathcal{F}_1^{(2)} \text{ in } \overline{\mathcal{F}_1^{(2)}} \text{ simultaneously.}$$

Few references deal with the considered two-sample problems from the minimax point of view.
In the density model, up to our knowledge, the only one is the paper by Butucea and Tribouley [BT06], where a minimax adaptive test is constructed, based on wavelet thresholding methods and a bootstrap approach.
As for regression models, let us notice that in particular fixed design regression models for instance, when f_1 and f_2 respectively denote the true signals of two observed noisy signals, any signal detection procedure applied to the differences of the observed signals can be used to test the null hypothesis $f_1 - f_2 = 0$. Moreover, some minimax separation rates can be determined directly from the minimax separation rates in a signal detection problem. All the references given in Chapter 1 on signal detection in such fixed design regression models are thus also relevant in the two-sample context, which offers numerous possibilities. In the heteroscedastic model that we consider in this chapter, no stringent assumption on the noise (such as a Gaussian distribution) is made, and in this sense, the minimax adaptive test of Durot and Rozenholc [DR06] which only assumes that the noise has a symmetric distribution, and which is based on the aggregation principle described in Chapter 1, is probably the closest one to the test we introduce here.
As for the inhomogeneous Poisson process model, no minimax result was available until our paper [10]. However, lower bounds for the minimax separation rates in the two-sample problem can be easily deduced from the proofs of the lower bounds in the homogeneity testing problem considered in [7] (see [11, Section 5] for more details). The same arguments could in fact be applied in the density model, thus providing lower bounds for the minimax separation rates in the two-sample problem, from the ones in the goodness-of-fit testing problem.

In the papers [9, 10, 16] presented in this chapter, we only obtained minimax adaptivity results in the Poisson process model, and this is why we particularly focus on this model in the following.
In the other models, our tests are constructed from the same aggregation scheme as in the Poisson process model, and are therefore expected to be also minimax adaptive, but this is the topic of a current work. Some of these tests are based on the aggregation of kernel or k-Nearest Neighbors testing procedures, that are known to be consistent against any alternative, and so are also consistent against any alternative.

Following the Neyman-Pearson principle, it has nevertheless to be recalled that the priority issue when constructing a testing procedure is that it satisfies $(\mathcal{P}_{\text{level},\alpha})$.
When considering single tests in one-sample problems such as in Chapter 1, this issue is obvious as soon as a test statistic, whose distribution is completely known or easy to simulate under the null hypothesis, is available. It is still rather simple when considering aggregated tests based on such single tests, as explained in Section 1.1.2. It becomes clearly more difficult in two-sample problems where, in general, the distributions of the considered test statistics are not free from the unknown function $f_1 = f_2$ under (H_0). Therefore, even if we hoped to obtain minimax adaptive properties for our tests, our main concern was to construct aggregated tests (thus expected to be minimax adaptive), which achieve $(\mathcal{P}_{\text{level},\alpha})$ despite nonfree distributions of the test statistics under (H_0). To this end, permutation and bootstrap approaches are introduced, and then adapted to the aggregation scheme.

3.1.2 Aggregated tests with permutation and bootstrap approaches

As explained in Chapter 2, Section 2.2, Efron's bootstrap was originally used in statistical problems where some probabilistic characteristic of a root, that is a functional of an observed sample and its distribution, has to be estimated. The case of linear roots has been widely studied in the literature, and in this particular case, Efron's bootstrap has even been generalized to more general weighted bootstrap approaches. The idea is to create a bootstrapped root whose conditional distribution, given the observed sample, is (asymptotically) close to the distribution of the initial root.
In testing problems, the same principle can be used either to approximate the distribution of a test statistic T under the null hypothesis (whatever the true hypothesis), or to mimic, under the null

hypothesis, the distribution of T. Asymptotic tests can thus be obtained (see Chapter 4 for instance), which are of prescribed asymptotic size α in $(0,1)$ and consistent against reasonable alternatives.
When considering the tests from a nonasymptotic point of view, the tests are required to be of level α that is to satisfy $(\mathcal{P}_{\text{level},\alpha})$ for any observed sample size. To this end, *exact bootstrap approaches* have to be introduced, so that, for instance, the conditional distribution of the bootstrapped test statistic given another particular statistic Z, is equal (and not only close) to the conditional distribution of the initial test statistic given Z. In two-sample problems, Z is generally taken as the pooled set $\bar{\mathbf{X}}$, and a historical example of exact bootstrap approach is the permutation approach which goes back to the thirties with Fisher's precursor work [Fis35], and which has been largely developed then with a huge family of permutation tests (see [PS10] for a review). Among the numerous references on permutation tests, one can cite at least Hoeffding's theoretical study [Hoe52], which lays the foundations of properly justified permutation tests, and which allows to generalize the principle by considering a general group of transformations instead of the only permutation group. The exact bootstrap approaches based on symmetrization arguments, such as the ones introduced by Durot, Rozenholc [DR06] and Arlot, Blanchard, Roquain [ABR10] can for instance be viewed as such a generalized permutation approach, where the considered group of transformations is $\{\, g_\epsilon : (x_1,\dots,x_n) \in \mathbb{R}^n \mapsto (\epsilon_1 x_1,\dots,\epsilon_n x_n),\ \epsilon = (\epsilon_1,\dots,\epsilon_n) \in \{-1,1\}^n \,\}$.
Despite the incontestable assets of the original permutation principle (starting from its simplicity), it suffers from a high computational cost, and this is why most of authors have given preference to purely asymptotic tests, based on the limit distribution of the test statistic under (H_0), when the sample size is large, or even moderate. The introduction of Monte Carlo methods to approximate the permutation based quantiles, as well as the development of computer facilities, now make possible the use of permutation tests within a reasonable computing time (see for instance [15] or Chapter 5 of this dissertation). But, to our knowledge, such a use of Monte Carlo permutation methods have been theoretically justified from a nonasymptotic point of view, that is have been proved to respect the property $(\mathcal{P}_{\text{level},\alpha})$, only since Romano and Wolf's key lemma [RW05, Lemma 1] (see Lemma 4 below).

In the present chapter, we consider, in the density and the heteroscedastic regression models, a classical permutation approach. In the Poisson process model, we introduce an exact bootstrap approach, inspired from the wild bootstrap developed by Mammen [Mam92], and based on Rademacher random variables as in [DR06] and [ABR10].
Both approaches are in fact rather different in spirit, as the bootstrap approach in the Poisson process model could only be viewed as a conditional generalized permutation approach, given the event $(N_1,N_2)=(n_1,n_2)$ for fixed integers $n_1,n_2 \geq 1$. They nevertheless share the common expected *conditional invariance property*: for any considered test statistic T, under (H_0), the conditional distribution of the permuted or bootstrapped test statistic T^ε given $\bar{\mathbf{X}}$, is equal to the conditional distribution of T given $\bar{\mathbf{X}}$. The notation T^ε for the permuted or the bootstrapped statistic comes from the classical notation for Rademacher random variables, which are used in the Poisson process model.

Considering a test statistic T, whose large values lead to reject (H_0), the corresponding critical value can be taken as the $(1-\alpha)$ quantile of the conditional distribution of T^ε given $\bar{\mathbf{X}}$, denoted by $q^{(\bar{\mathbf{X}})}(1-\alpha)$. Indeed, from the above conditional invariance property, one deduces that given α in $(0,1)$, under (H_0),

$$P_{f_1,f_2}\left(T > q^{(\bar{\mathbf{X}})}(1-\alpha)\middle|\bar{\mathbf{X}}\right) \leq \alpha,$$

which implies that the single test $\phi_\alpha = \mathbb{1}_{\left\{T > q^{(\bar{\mathbf{X}})}(1-\alpha)\right\}}$ satisfies $(\mathcal{P}_{\text{level},\alpha})$.
The test ϕ_α can even be slightly modified with a randomization tool so that it is exactly of size α, but as we use in practice ϕ_α and more precisely its Monte Carlo version (that is, with Monte Carlo approximations of $q^{(\bar{\mathbf{X}})}(1-\alpha)$), it seems quite useless to study the randomized version of ϕ_α. Romano and Wolf's lemma then allows to prove that replacing $q^{(\bar{\mathbf{X}})}(1-\alpha)$ in ϕ_α by a Monte Carlo approximation $q^{MC(\bar{\mathbf{X}})}(1-\alpha)$ in fact leads to a test ϕ_α^{MC} still satisfying $(\mathcal{P}_{\text{level},\alpha})$ (see Section 3.2.1).

Our tests are constructed according to the principle of aggregation described in Section 1.1.2. Therefore, are considered a collection $\left\{ \mathcal{F}_{0,m}^{(2)},\ m \in \mathcal{M} \right\}$ of subsets of $\mathcal{F}^2$ such that

$$\mathcal{F}_0^{(2)} := \{(f_1, f_2) \in \mathcal{F}^2,\ f_1 = f_2\} \subset \cap_{m\in\mathcal{M}} \mathcal{F}_{0,m}^{(2)},$$

and the collection of associated hypotheses $\{ (H_{0,m}),\ m \in \mathcal{M} \}$ such that $(H_{0,m})$ $(f_1, f_2) \in \mathcal{F}_{0,m}^{(2)}$. Then, for every m in $\mathcal{M}$ and every level u in $(0, 1)$, a single test $\phi_{m,u}$ of

$$(H_{0,m})\ (f_1, f_2) \in \mathcal{F}_{0,m}^{(2)} \quad \text{against} \quad (H_{1,m})\ (f_1, f_2) \notin \mathcal{F}_{0,m}^{(2)},$$

satisfying $(\mathcal{P}_{\text{level},u})$, is constructed as above:

$$\phi_{m,u} = \mathbb{1}_{\left\{ T_m > q_m^{(\bar{\mathbf{X}})}(1-u) \right\}},$$

where T_m is a test statistic of $(H_{0,m})$ against $(H_{1,m})$, and $q_m^{(\bar{\mathbf{X}})}(1-u)$ is the $(1-u)$ quantile of the conditional distribution of T_m^{ε} given $\bar{\mathbf{X}}$. Given a family of positive weights $(w_m)_{m\in\mathcal{M}}$ such that $\sum_{m\in\mathcal{M}} w_m \leq 1$, consider now the collection of single tests

$$\Phi_\alpha^{bootFLR} = \left\{ \phi_{m,u_{m,\alpha}^{(\bar{\mathbf{X}})}},\ m \in \mathcal{M} \right\} = \left\{ \mathbb{1}_{\left\{ T_m > q_m^{(\bar{\mathbf{X}})}\left(1-u_{m,\alpha}^{(\bar{\mathbf{X}})}\right) \right\}},\ m \in \mathcal{M} \right\}, \tag{3.1}$$

with

$$u_{m,\alpha}^{(\bar{\mathbf{X}})} = w_m \sup \left\{ u,\ \mathbb{P}\left(\exists m \in \mathcal{M},\ T_m^{\varepsilon} > q_m^{(\bar{\mathbf{X}})}(1 - w_m u) \middle| \bar{\mathbf{X}} \right) \leq \alpha \right\}.$$

The aggregated tests that we introduce and study are of the form (1.1) based on $\Phi_\alpha^{bootFLR}$, that is

$$\bar{\Phi}_\alpha^{bootFLR} = \sup_{m\in\mathcal{M}} \phi_{m,u_{m,\alpha}^{(\bar{\mathbf{X}})}} = \sup_{m\in\mathcal{M}} \mathbb{1}_{\left\{ T_m > q_m^{(\bar{\mathbf{X}})}\left(1-u_{m,\alpha}^{(\bar{\mathbf{X}})}\right) \right\}}. \tag{3.2}$$

Note that these tests, which are constructed so that they satisfy $(\mathcal{P}_{\text{level},\alpha})$, generalize the aggregated test based on an instrumental conditional distribution $\bar{\Phi}_\alpha^{condFLR}$ defined in Section 1.1.2.

3.1.3 Single hypotheses based on kernels

Assume that $\mathcal{F} = \mathbb{L}_1(\mathbb{Y}, \nu) \cap \mathbb{L}_\infty(\mathbb{Y}) \subset \mathbb{L}_2(\mathbb{Y}, \nu)$ for some space $\mathbb{Y}$, equal to $\mathbb{X}$ in the density and the Poisson process frameworks, to the space where the covariates are observed in the regression framework, and some measure ν on $\mathbb{Y}$. Let $\|.\|_2$ and $\langle ., . \rangle_2$ be the usual norm and scalar product of $\mathbb{L}_2(\mathbb{Y}, \nu)$.
As in Chapter 1, a classical choice for the collection of single hypotheses $\{ (H_{0,m}),\ m \in \mathcal{M} \}$ could be defined from a collection of subspaces $\{ S_m,\ m \in \mathcal{M} \}$ of $\mathbb{L}_2(\mathbb{Y}, \nu)$ by $(H_{0,m})$ $\Pi_{S_m}(f_1 - f_2) = 0$, where Π_{S_m} denotes the orthogonal projection onto S_m with respect to $\langle ., . \rangle_2$.
In our papers [9, 10, 16], this choice of hypotheses is only viewed as a particular case of a much more general kind of hypotheses, leading to single tests not only closely linked with model selection and thresholding estimation methods as in Chapter 1, but also based on kernel estimators and reproducing kernel methods coming from the statistical learning theory, like in [GBR+08, GFHS10, GBR+12, SSGF13]. Thus, in the following, we consider some collections of hypotheses $\{ (H_{0,m}),\ m \in \mathcal{M} \}$ based on general symmetric kernels, that is on functions $K_m : \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}$ such that $K_m(y, y') = K_m(y', y)$ for every y, y' in $\mathbb{Y}$. Denoting by $\diamond$ the integral operator defined for every symmetric kernel K and every function g in $\mathbb{L}_2(\mathbb{Y}, \nu)$ by

$$K \diamond g(y) = \langle K(., y), g \rangle_2,\ \forall y \in \mathbb{Y},$$

$(H_{0,m})$ is expressed as

$$(H_{0,m})\ \langle K_m \diamond (f_1 - f_2), f_1 - f_2 \rangle_2 = 0.$$

Three kinds of symmetric kernels are of particular interest.

[Projection kernel] Considering an orthonormal family $\{\varphi_l,\ l \in \mathcal{L}\}$ for the scalar product $\langle .,. \rangle_2$, such that $\sup_{y,y' \in \mathbb{Y}} \sum_{l \in \mathcal{L}} |\varphi_l(y)\varphi_l(y')| < +\infty$, a projection kernel is defined by

$$K(y,y') = \sum_{l \in \mathcal{L}} \varphi_l(y)\varphi_l(y').$$

Note that for every function g in $\mathbb{L}_2(\mathbb{Y},\nu)$, $K \diamond g = \Pi_{S_\mathcal{L}}(g)$, where $S_\mathcal{L}$ is the subspace of $\mathbb{L}_2(\mathbb{Y},\nu)$ generated by $\{\varphi_l,\ l \in \mathcal{L}\}$, and therefore $\langle K \diamond (f_1 - f_2), f_1 - f_2 \rangle_2 = \|\Pi_{S_\mathcal{L}}(f_1 - f_2)\|_2^2$.

[Approximation kernel] When $\mathbb{Y} = \mathbb{R}^d$ and ν is the Lebesgue measure, considering a usual kernel function k in $\mathbb{L}_2(\mathbb{R}^d,\lambda)$ such that $k(-y) = k(y)$ for every y in $\mathbb{R}^d$, and a vector $h = (h_1,\ldots,h_d)$ of d positive bandwidths, an approximation kernel is defined by

$$K(y,y') = \frac{1}{\prod_{i=1}^d h_i} k\left(\frac{y_1 - y'_1}{h_1},\ldots,\frac{y_d - y'_d}{h_d}\right),$$

for $y = (y_1,\ldots,y_d)$, $y' = (y'_1,\ldots,y'_d)$ in $\mathbb{Y}$. Note that for every function g in $\mathbb{L}_2(\mathbb{Y},\nu)$, $K \diamond g = k_h * g$, where

$$k_h(u_1,\ldots,u_d) = \frac{1}{\prod_{i=1}^d h_i} k\left(\frac{u_1}{h_1},\ldots,\frac{u_d}{h_d}\right),$$

and $*$ is the usual convolution operator with respect to the measure ν.

[Reproducing kernel] A reproducing kernel (see [SS02] for instance) is such that

$$K(y,y') = \langle \theta(y), \theta(y') \rangle_{\mathcal{H}_K},$$

where θ and $\mathcal{H}_K$ are a representation function and a RKHS associated with K. Here, $\langle .,. \rangle_{\mathcal{H}_K}$ denotes the scalar product of $\mathcal{H}_K$. Every considered reproducing kernel K is furthermore assumed to satisfy

$$\int_{\mathbb{Y}} K^2(y,y')(f_1+f_2)(y)(f_1+f_2)(y')d\nu(y)d\nu(y') < +\infty, \tag{3.3}$$

so that by Cauchy-Schwartz inequality, $\langle K \diamond (f_1 - f_2), f_1 - f_2 \rangle_2$ is well-defined.

Remark that as f_1 and f_2 are assumed to be in $\mathcal{F} = \mathbb{L}_1(\mathbb{Y},\nu) \cap \mathbb{L}_\infty(\mathbb{Y})$, (3.3) is also satisfied by projection kernels based on finite orthonormal families and approximation kernels.

A notable point is that $\langle K \diamond (f_1 - f_2), f_1 - f_2 \rangle_2 = \|m_{f_1} - m_{f_2}\|_{\mathcal{H}_K}^2$, where $m_{f_1} = K \diamond f_1 = \int \theta(y) f_1(y) d\nu(y)$ and $m_{f_2} = K \diamond f_2 = \int \theta(y) f_2(y) d\nu(y)$.

In the density model where f_1 and f_2 are densities w.r.t. the measure ν, m_{f_1} and m_{f_2} are the mean embeddings in the RKHS $\mathcal{H}_K$ of the distributions $f_1 d\nu$ and $f_2 d\nu$ respectively (see [BTA04] or [SGF+10] for instance). The distance $\|m_{f_1} - m_{f_2}\|_{\mathcal{H}_K}$ is moreover known as the Maximum Mean Discrepancy on the unit ball in the RKHS $\mathcal{H}_K$ (see [GBR+08, GFHS10, GBR+12, SSGF13]) .

When the kernel is characteristic (see [SGF+10, SFL11]), the map which assigns its mean embedding in $\mathcal{H}_K$ to any probability distribution is injective by definition. Hence, $\langle K \diamond (f_1 - f_2), f_1 - f_2 \rangle_2 = 0$ if and only if $f_1 = f_2$, which means that any hypothesis $(H_{0,m})$ defined from a characteristic reproducing kernel K_m is equivalent to (H_0).

For any symmetric kernel K_m chosen as in *[Projection kernel]*, *[Approximation kernel]*, or *[Reproducing kernel]*, and satisfying (3.3), a reasonable test statistic of $(H_{0,m})$ against $(H_{1,m})$ can be obtained with an unbiased estimator of $\langle K_m \diamond (f_1 - f_2), f_1 - f_2 \rangle_2$.

Such test statistics are introduced in each considered model in the following sections, and permutation or bootstrap-based corresponding critical values are proposed. The obtained single tests are then integrated in an aggregated test of the form $\bar{\Phi}_\alpha^{bootFLR}$ defined in (3.2).

In statistical learning, in particular in binary classification problems with finite dimensional input data, kernel and k-Nearest Neighbors plug-in rules are often studied in parallel, as they share many convergence properties.
In this spirit, we furthermore consider in the density model with $\mathbb{X} = \mathbb{R}^d$, single tests inspired from the k-Nearest Neighbors classification rule (see Chapter 2) that were introduced in [Sch86] and [Hen88]. In [16], we underline the links between these k-NN tests and the kernel based ones, and we propose to aggregate them in the same way. The resulting Nearest Neighbors aggregated test can then be adapted to handle functional data, following ideas from [BBW05] and [3].

3.2 Kernel methods in the Poisson process model

Let us consider in this section the following model.

$\mathcal{M}^{(2)}_{\text{Poisson}}$	$\mathbf{X} = (\mathbf{X}^1, \mathbf{X}^2)$ is a pair of independent Poisson processes $\mathbf{X}^1 = \{X^1_1, \dots, X^1_{N_1}\}$ and $\mathbf{X}^2 = \{X^2_1, \dots, X^2_{N_2}\}$, observed on a measurable space $\mathbb{X}$, with respective intensities f_1 and f_2, with respect to some measure μ on $\mathbb{X}$ such that $d\mu = nd\nu$, for a fixed nonatomic σ-finite measure ν and a fixed positive integer n.

The measure ν may typically be the Lebesgue measure λ when $\mathbb{X}$ is a measurable subset of $\mathbb{R}^d$.
We assume that f_1 and f_2 both belong to $\mathbb{L}_1(\mathbb{X}, \nu) \cap \mathbb{L}_\infty(\mathbb{X}) \subset \mathbb{L}_2(\mathbb{X}, \nu)$, endowed with its classical norm $\|.\|_2$ and scalar product $\langle ., . \rangle_2$ as in Section 3.1.
We consider the problem of testing, from the observation of $\mathbf{X} = (\mathbf{X}^1, \mathbf{X}^2)$,

$$(H_0) \quad f_1 = f_2 \quad \text{against} \quad (H_1) \quad f_1 \neq f_2.$$

3.2.1 Single tests with a wild bootstrap approach

Let K_m be a symmetric kernel as in *[Projection kernel]*, *[Approximation kernel]*, or *[Reproducing kernel]*, and satisfying (3.3).
As explained above, a single test statistic of $(H_{0,m})$ against $(H_{1,m})$ and therefore of (H_0) against (H_1) can be obtained with any unbiased estimator of $\langle K_m \diamond (f_1 - f_2), f_1 - f_2 \rangle_2$.
When the kernel K_m is a projection kernel based on an orthonormal family $\{\varphi_l,\ l \in \mathcal{L}_m\}$ for $\langle ., . \rangle_2$, one knows that $\langle K_m \diamond (f_1 - f_2), f_1 - f_2 \rangle_2 = \left\| \Pi_{S_{\mathcal{L}_m}} (f_1 - f_2) \right\|_2^2$. In this case, it is rather easy to see that an unbiased estimator of $n^2 \langle K_m \diamond (f_1 - f_2), f_1 - f_2 \rangle_2$ is given by

$$T_m = \sum_{l \in \mathcal{L}_m} \left(\left(\int_{\mathbb{X}} \varphi_l dN_{\mathbf{X}^1} - \int_{\mathbb{X}} \varphi_l dN_{\mathbf{X}^2} \right)^2 - \int_{\mathbb{X}} \varphi_l^2 dN_{\bar{\mathbf{X}}} \right),$$

where dN_x stands for the point measure associated with x, defined by (1.20).
The statistic T_m can therefore be taken as test statistic. Noticing that it can also be expressed as

$$\sum_{i,j \in \{1,\dots,N\},\ i \neq j} \left(\sum_{l \in \mathcal{L}_m} \varphi_l(X_i) \varphi_l(X_j) \right) \varepsilon_i^0 \varepsilon_j^0,$$

where the ε_i^0's are some marks on $\bar{\mathbf{X}}$, such that $\varepsilon_i^0 = 1$ if X_i belongs to $\mathbf{X}^1$, $\varepsilon_i^0 = -1$ if X_i belongs to $\mathbf{X}^2$, this test statistic can be generalized as

$$T_m = \sum_{i,j \in \{1,\dots,N\},\ i \neq j} K_m(X_i, X_j) \varepsilon_i^0 \varepsilon_j^0. \tag{3.4}$$

A wild bootstrap approach as exact bootstrap approach. The main point now is to define an exact bootstrap approach, that is a bootstrap approach satisfying the conditional invariance property defined in Section 3.1. A basic permutation approach would not be appropriate here, since it would not take the randomness of N_1 and N_2 into account. Therefore, we turn to another bootstrap type approach, starting from the remark that under (H_0), the test statistic T_m is a degenerate U-statistic of order 2. Bootstrapping degenerate U-statistic of order 2 is not an obvious question. A naive application of Efron's original bootstrap indeed fails in this case (see [Bre83]), since it leads the bootstrapped statistic to lose the degeneracy property. Apart from the m out of n bootstrap introduced by Bretagnolle [Bre83] to overcome this difficulty, Arcones and Giné [AG92] proposed a solution based on Efron's original bootstrap, but combined with a centering trick to force the bootstrapped statistic to satisfy the degeneracy property. The results of Arcones and Giné were then generalized to other kinds of bootstrap methods, and in particular Bayesian and wild bootstrapped U-statistics were introduced in [HJ93], [Jan94] and [DM94].
Following [DM94], we introduce a sequence $(\varepsilon_i)_{i\in\mathbb{N}}$ of i.i.d. Rademacher variables independent of $\bar{\mathbf{X}}$. Then a wild bootstrapped version of T_m may be expressed as $\sum_{i,j\in\{1,\dots,N\},\ i\neq j} K_m(X_i,X_j)\varepsilon_i^0\varepsilon_j^0\varepsilon_i\varepsilon_j$. We consider in fact the simpler version

$$T_m^{\varepsilon} = \sum_{i,j\in\{1,\dots,N\},\ i\neq j} K_m(X_i,X_j)\varepsilon_i\varepsilon_j, \tag{3.5}$$

which has the same distribution under (H_0).
A general result of [DVJ08] allows to prove (see [10, Proposition 1]) that under (H_0), given $\bar{\mathbf{X}}$, $(\varepsilon_1^0,\dots,\varepsilon_N^0)$ has the same conditional distribution as $(\varepsilon_1,\dots,\varepsilon_N)$. Hence, the above wild bootstrap approach satisfies the conditional invariance property: under (H_0), the conditional distribution of T_m^{ε} given $\bar{\mathbf{X}}$, is equal to the conditional distribution of T_m given $\bar{\mathbf{X}}$.
Given a prescribed level α in $(0,1)$, denoting by $q_m^{(\bar{\mathbf{X}})}(1-\alpha)$ the $(1-\alpha)$ quantile of the conditional distribution of T_m^{ε} given $\bar{\mathbf{X}}$, we therefore consider the test

$$\phi_{m,\alpha} = \mathbb{1}_{\left\{T_m > q_m^{(\bar{\mathbf{X}})}(1-\alpha)\right\}}. \tag{3.6}$$

As explained in Section 3.1, from the conditional invariance property, it is easily deduced that $\phi_{m,\alpha}$ satisfies $(\mathcal{P}_{\text{level},\alpha})$.

Remark that the test $\phi_{m,\alpha}$ is very close to the MMD tests in the density model, now well developed in the statistical learning literature, with for instance [GBR+08, GFHS10, GBR+12, SSGF13]. The main difference lies in the construction of the critical values, here based on an exact bootstrap approach, while Gretton and co-authors use some approximated quantiles, based on concentration inequalities, in theory, and Efron's bootstrap or m out of n bootstrap approaches in practice.

Second kind error rate. Now, with a study of uniform separation rates in view, we state a nonasymptotic condition on the alternative (f_1,f_2) guaranteeing that $P_{f_1,f_2}(\phi_{m,\alpha}=0)\leq\beta$. Under (H_1), given $\bar{\mathbf{X}}$, the bootstrapped test statistic T_m^{ε} is a Rademacher chaos whose quantiles can be upper bounded thanks to results in [dlPG99] or [Lat99], which leads to the following theorem.

Theorem 7 (Fromont, Laurent, Reynaud-Bouret, 2013)**.** *Let α,β in $(0,1)$. Let K_m be a symmetric kernel as in [Projection kernel], [Approximation kernel], or [Reproducing kernel], satisfying* (3.3)*, and $\phi_{m,\alpha}$ be the test defined by* (3.6)*. Assume that $\int_{\mathbb{X}^2} K_m^2(x,x')(f_1+f_2)(x)(f_1+f_2)(x')d\nu(x)d\nu(x')$ is upper bounded by C_m. There exists some constant $\kappa>0$ such that $P_{f_1,f_2}(\phi_{m,\alpha}=0)\leq\beta$, as soon as*

$$\|f_1-f_2\|_2^2 \geq \inf_{r>0}\left\{\left\|(f_1-f_2)-r^{-1}K_m\diamond(f_1-f_2)\right\|_2^2 + \frac{4+\kappa\ln(2/\alpha)}{nr\sqrt{\beta}}\sqrt{C_m}\right\} + \frac{8\,\|f_1+f_2\|_\infty}{\beta n}. \tag{3.7}$$

Notice that when K_m is a projection kernel based on an orthonormal basis $\{\varphi_l,\ l \in \mathcal{L}_m\}$, C_m can be taken as $C_m = C(\|f_1+f_2\|_1, \|f_1+f_2\|_\infty)D_m$ if either $\#\mathcal{L}_m = D_m$ or

$$\sup_{x,x'\in\mathbb{X}} \sum_{l\in\mathcal{L}_m} |\varphi_l(x)\varphi_l(x')| = D_m < +\infty, \text{ and } \int_{\mathbb{X}^2} \left(\sum_{l\in\mathcal{L}_m} |\varphi_l(x)\varphi_l(x')| \right)^2 (f_1+f_2)(x')d\nu(x)d\nu(x') < +\infty. \tag{3.8}$$

When K_m is an approximation kernel based on a kernel function k_m and a vector of bandwidth $h_m = (h_{m,1}, \dots, h_{m,d})$, C_m can be taken as $\|f_1+f_2\|_\infty \|f_1+f_2\|_1 \|k_m\|_2^2 / \prod_{i=1}^d h_{m,i}$. Hence in these two cases, by taking $r=1$ in (3.7), the right hand side of the inequality reproduces a bias-variance decomposition which is known to lead to sharp upper bounds for the uniform separation rates over particular classes of alternatives (see Chapter 1 or [Tsy09] for instance).
When K_m is a reproducing kernel proportional to a projection or approximation kernel, then the normalization factor r^{-1} can be chosen such that $\left\|(f_1-f_2) - r^{-1}K_m \diamond (f_1-f_2)\right\|_2$ is still a bias term. We thus recover for such reproducing kernels, like the Gaussian and Laplacian ones commonly used in statistical learning theory, a classical bias-variance decomposition.

Assume now that $\mathbb{X} = \mathbb{R}^d$, $\int_{\mathbb{X}} f_1(x)d\nu(x) = \int_{\mathbb{X}} f_2(x)d\nu(x) = 1$, and that K_m is a bounded measurable characteristic reproducing kernel such that $K_m(x,x)$ is constant. This last assumption is not unusual since it is satisfied by any normalized or translation-invariant kernel (see [SS02, pages 46-47, 57], or [SGF+10, SFL11] for instance). Moreover, as specified in [SGF+10] for instance, bounded continuous characteristic and translation-invariant reproducing kernels exist, at least in $\mathbb{R}^d$, where Bochner's theorem enables to characterize them. In this case, we prove (see [10, Theorem 2]) that for n large enough, there exists $C(\alpha,\beta)$ such that $P_{f_1,f_2}(\phi_{m,\alpha}=0) \leq \beta$ as soon as

$$\|m_f - m_g\|^2_{\mathcal{H}_{K_m}} \geq \frac{C(\alpha,\beta)}{n}.$$

This result allows to obtain a uniform separation rate, for the weak distance between $f_1 d\nu$ and $f_2 d\nu$ equal to $\|m_{f_1} - m_{f_2}\|_{\mathcal{H}_{K_m}}$, of the same order as the usual parametric separation rate, that is of order $n^{-1/2}$. In this sense, it is comparable to previous results of Giné [Gin75] and Wellner [Wel79], where the same parametric separation rate is obtained, but for other weak distances between distributions. We refer the reader to [SSGF13] for a study of connexions between $\|m_{f_1} - m_{f_2}\|_{\mathcal{H}_{K_m}}$ and several more classical weak distances between distributions.

Monte Carlo approximations. Studying our tests from a nonasymptotic point of view poses the additional question of the exact loss in first and second kind error rates due to the Monte Carlo approximation of $q_m^{(\bar{\mathbf{X}})}(1-\alpha)$, which is used in practice. We address this question in [10].

Let B a fixed number of Monte Carlo iterations, large enough so that $\alpha(B+1) \geq 1$. Given $\bar{\mathbf{X}}$ and therefore N, let us introduce a set $\{\varepsilon^b, b \in \{1,\dots,B\}\}$ of B independent samples $\varepsilon^b = (\varepsilon_1^b, \dots, \varepsilon_N^b)$ of N i.i.d. Rademacher random variables. Let, for b in $\{1,\dots,B\}$, $T_m^{\varepsilon^b} = \sum_{i,j\in\{1,\dots,N\},\ i\neq j} K_m(X_i,X_j)\varepsilon_i^b\varepsilon_j^b$. Denoting by $\left(T_m^{\varepsilon^{(1)}}, \dots, T_m^{\varepsilon^{(B+1)}}\right)$ the order statistic associated with $\left(T_m^{\varepsilon^1}, \dots, T_m^{\varepsilon^B}, T_m\right)$, given $\bar{\mathbf{X}}$, we set

$$q_m^{MC(\bar{\mathbf{X}})}(1-\alpha) = T_m^{\varepsilon^{(\lceil (1-\alpha)(B+1)\rceil)}}.$$

The Monte Carlo single kernel-based test that we consider in the present dissertation is

$$\phi_{m,\alpha}^{MC} = \mathbb{1}_{\left\{T_m > q_m^{MC(\bar{\mathbf{X}})}(1-\alpha)\right\}}. \tag{3.9}$$

Note that this test is slightly different from the one considered in [10, Proposition 3], so that it has a better control of the first kind error rate.
The following lemma due to Romano and Wolf is of fundamental interest when considering Monte Carlo approximations of exact bootstrap approaches, so it seems important to us to recall it.

Lemma 4 (Romano, Wolf, Lemma 1 in [RW05]). *If $V_1, ..., V_{B+1}$ are $B+1$ exchangeable real random variables, then for all u in $[0,1]$,*

$$\mathbb{P}\left(\frac{1}{B+1}\left(1+\sum_{i=1}^{B}\mathbb{1}_{V_i\geq V_{B+1}}\right)\leq u\right)\leq u.$$

Noticing that

$$P_{f_1,f_2}\left(T_m > T_m^{\varepsilon(\lceil(1-\alpha)(B+1)\rceil)}\middle|\bar{\mathbf{X}}\right)\leq \mathbb{P}\left(\sum_{b=1}^{B}\mathbb{1}_{T_m^{\varepsilon b}\geq T_m}\leq \alpha(B+1)-1\middle|\bar{\mathbf{X}}\right),$$

and that given $\bar{\mathbf{X}}$, under (H_0), the variables $\left(T_m^{\varepsilon 1},\ldots,T_m^{\varepsilon B},T_m\right)$ are exchangeable, Lemma 4 can be applied and $\phi_{m,\alpha}^{MC}$ is proved to satisfy $(\mathcal{P}_{\text{level},\alpha})$.
A nonasymptotic condition guaranteeing that $P_{f_1,f_2}(\phi_{m,\alpha}^{MC}=0)\leq\beta$ can also be obtained, using Hoeffding's concentration inequality as in the proof of [10, Proposition 4].

3.2.2 Aggregated tests

Given a collection of kernels $\{K_m,\ m\in\mathcal{M}\}$, we now consider the aggregated test $\bar{\Phi}_\alpha^{bootFLR}$ of (3.2), based on the single tests $\phi_{m,\alpha}$ defined by (3.6), and the above exact wild bootstrap approach.
Note that such an aggregated test may be viewed as a multiple kernel procedure in the spirit of multiple kernel learning which is a current challenging topic in machine learning. Indeed, it allows to consider several kernels, instead of a single one (whose choice or calibration is always a major and thorny question), and to combine them in an automatic data-driven way, here through the individual levels $u_{m,\alpha}^{(\bar{\mathbf{X}})}$ involved in $\bar{\Phi}_\alpha^{bootFLR}$.
An alternative, proposed in [SGF+10], is to consider the test statistic $\sup_{m\in\mathcal{M}}T_m$, and to reject (H_0) when this test statistic is larger than a given critical value, either constructed via concentration inequalities, or classical bootstrap approaches. Such a test, close in spirit to Kolmogorov-Smirnov tests (see [VdVW96] for instance) would however not achieve the same nonasymptotic properties, expressed as oracle type inequalities or minimax adaptivity results, as our aggregated test $\bar{\Phi}_\alpha^{bootFLR}$.

Since the aggregated test $\bar{\Phi}_\alpha^{bootFLR}$ is especially constructed so that it satisfies $(\mathcal{P}_{\text{level},\alpha})$, let us now focus on its nonasymptotic properties from the second kind error rate angle. We first obtained in [10] the following oracle type result.

Theorem 8 (Fromont, Laurent, Reynaud-Bouret, 2013). *Let $\alpha,\beta\in(0,1)$. Let $\{K_m, m\in\mathcal{M}\}$ be a collection of kernels, chosen as in one of the two following cases, and $\{w_m, m\in\mathcal{M}\}$ be a collection of positive weights such that $\sum_{m\in\mathcal{M}}w_m\leq 1$.*

[Multiple projection kernel] Let $\{S_m, m\in\mathcal{M}\}$ be a finite collection of linear subspaces of $\mathbb{L}_2(\mathbb{X},\nu)$, spanned by orthonormal bases denoted by $\{\varphi_l,\ l\in\mathcal{L}_m\}$ respectively. Assume either that S_m has finite dimension D_m or that (3.8) holds. We set, for all m in $\mathcal{M}$, $K_m(x,x')=\sum_{l\in\mathcal{L}_m}\varphi_l(x)\varphi_l(x')$, and we introduce the condition:

$$\|f_1-f_2\|_2^2\geq \inf_{m\in\mathcal{M}}\left\{\|(f_1-f_2)-\Pi_{S_m}(f_1-f_2)\|_2^2+ C(\beta,\|f_1+f_2\|_1,\|f_1+f_2\|_\infty)\frac{(C(\alpha)+\ln(1/w_m))\sqrt{D_m}}{n}\right\}. \quad (3.10)$$

[Multiple approximation kernel] If $\mathbb{X}=\mathbb{R}^d$ and ν is the Lebesgue measure on $\mathbb{R}^d$, let $\{k_{m_1},\ m_1\in\mathcal{M}_1\}$ be a collection of kernel functions such that $\int_{\mathbb{X}}k_{m_1}^2(x)d\nu(x)<\infty$, $k_{m_1}(x)=k_{m_1}(-x)$, and a collection

$\{h_{m_2},\ m_2 \in \mathcal{M}_2\}$ of vectors $h_{m_2} = (h_{m_2,1}, \dots, h_{m_2,d})$ of d positive numbers. We set $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2$, and for all $m = (m_1, m_2)$ in $\mathcal{M}$, $x = (x_1, \dots, x_d)$, $x' = (x'_1, \dots, x'_d)$ in $\mathbb{R}^d$,

$$K_m(x, x') = k_{m_1,h_{m_2}}(x - x') = \frac{1}{\prod_{i=1}^d h_{m_2,i}} k_{m_1}\left(\frac{x_1 - x'_1}{h_{m_2,1}}, \dots, \frac{x_d - x'_d}{h_{m_2,d}}\right).$$

We introduce the following condition:

$$\|f_1 - f_2\|_2^2 \geq \inf_{m=(m_1,m_2)\in\mathcal{M}} \left\{ \left\|(f_1 - f_2) - k_{m_1,h_{m_2}} * (f_1 - f_2)\right\|_2^2 + C\left(\beta, \|f_1 + f_2\|_\infty, \|f_1 + f_2\|_1, \|k_{m_1}\|_2\right) \frac{\left(C(\alpha) + \ln(1/w_m)\right)\sqrt{\prod_{i=1}^d h_{m_2,i}^{-1}}}{n} \right\}. \quad (3.11)$$

Let $\bar{\Phi}_\alpha^{bootFLR}$ be the test given by (3.2), based on the single tests $\phi_{m,\alpha}$ defined in (3.6) associated with the weights w_m, and the above exact wild bootstrap approach. Then $P_{f_1,f_2}\left(\bar{\Phi}_\alpha^{bootFLR} = 0\right) \leq \beta$, if either (3.10) in [Multiple projection kernel], or (3.11) in [Multiple approximation kernel] is satisfied, for some constants $C(\alpha)$, $C(\beta, \|f_1 + f_2\|_\infty, \|f_1 + f_2\|_1)$, and $C(\beta, \|f_1 + f_2\|_\infty, \|f_1 + f_2\|_1, \|k_{m_1}\|_2)$.

Comparing this result with the one obtained in Theorem 7 for the proposed constants C_m in the projection and approximation kernels cases, one can see that considering the aggregated test instead of a single one, allows to obtain the infimum over all m in $\mathcal{M}$ in the right hand side of (3.10) and (3.11) at the price of the additional term $\ln(1/w_m)$. This result can thus be viewed as an oracle type property: indeed, without knowing $(f_1 - f_2)$, one has that the uniform separation rate of the aggregated test is of the same order as the smallest uniform separation rate of the involved single kernel tests, up to factors $\ln(1/w_m)$. It is used to prove that, for well-chosen projection or approximation kernels, the aggregated test has minimax adaptivity properties.

Minimax adaptivity over Besov and weak Besov spaces. Let us here assume that $\mathbb{X} = [0, 1]$ and ν is the Lebesgue measure on $[0, 1]$. Consider, as in Section 1.4.1, the Haar basis $\{\varphi_0,\ \psi_{(j,k)},\ j \in \mathbb{N},\ k \in \{0, \dots, 2^j - 1\}\}$ of $\mathbb{L}_2([0, 1], \nu)$ defined by $\varphi_0(x) = \mathbb{1}_{[0,1]}(x)$, and $\psi_{(j,k)}(x) = 2^{j/2}\psi(2^j x - k)$, with $\psi(x) = \mathbb{1}_{[0,1/2)}(x) - \mathbb{1}_{[1/2,1)}(x)$.
We introduce here slightly different versions of the (weak) Besov bodies introduced in (1.22) and (1.23) in Chapter 1, adapted to the present two-sample problem. For $s, s', R, R' > 0$, let

$$\mathcal{B}^{(2)}_{s,2,\infty}(R) = \Big\{(f_1, f_2) \in \mathcal{F}^2,\ f_1 - f_2 = \alpha_0\varphi_0 + \textstyle\sum_{j\in\mathbb{N}} \sum_{k=0}^{2^j-1} \alpha_{(j,k)}\psi_{(j,k)},\ \alpha_0^2 \leq R^2,\ \forall j \in \mathbb{N},\ \sum_{k=0}^{2^j-1} \alpha_{(j,k)}^2 \leq R^2 2^{-2js}\Big\},$$

and

$$w\mathcal{B}^{(2)}_{s'}(R') = \Big\{(f_1, f_2) \in \mathcal{F}^2,\ f_1 - f_2 = \alpha_0\varphi_0 + \textstyle\sum_{j\in\mathbb{N}} \sum_{k=0}^{2^j-1} \alpha_{(j,k)}\psi_{(j,k)},\ \forall t > 0,\ \alpha_0^2\mathbb{1}_{\alpha_0^2\leq t} + \sum_{j\in\mathbb{N}} \sum_{k=0}^{2^j-1} \alpha_{(j,k)}^2\mathbb{1}_{\alpha_{(j,k)}^2\leq t} \leq R'^2 t^{\frac{2s'}{2s'+1}}\Big\}.$$

For $s, s', R, R' > 0$, and $R'' \geq 2$, some levels α and β in $(0, 1)$ such that $\alpha + \beta \leq 0.59$, and the metric d_2 associated with the norm $\|.\|_2$, $m\mathrm{SR}_{d_2}^{\alpha,\beta}\left(\mathcal{B}^{(2)}_{s,2,\infty}(R) \cap w\mathcal{B}^{(2)}_{s'}(R') \cap \left(\mathbb{L}_\infty(R'')\right)^2\right)$ has the same lower bound as $m\mathrm{SR}_{d_2}^{\alpha,\beta}\left(\mathcal{B}_{s,2,\infty}(R) \cap w\mathcal{B}_{s'}(R') \cap \mathbb{L}_\infty(R'')\right)$ in Theorem 3. This lower bound is directly deduced from the arguments of the proof of Theorem 3 (see [11, Section 5] for more details).

In the spirit of the minimax adaptive tests proposed in [7] (see Section 1.4), we consider the test $\bar{\Phi}_\alpha^{bootFLR}$ of Theorem 8, defined from a collection of projection kernels, based on either nested or nonnested subsets of the Haar basis. Let us introduce:

— $K_0(x,x') = \varphi_0(x)\varphi_0(x')$,
— $K_{(j,k)}(x,x') = \psi_{(j,k)}(x)\psi_{(j,k)}(x')$ for every j in $\mathbb{N}$, k in $\{0,\dots,2^j-1\}$,
— $K_J(x,x') = \varphi_0(x)\varphi_0(x') + \sum_{j\in\{0,\dots,J-1\},\ k\in\{0,\dots,2^j-1\}} \psi_{(j,k)}(x)\psi_{(j,k)}(x')$, for every $J \geq 1$.

For some integer $\bar{J}$ such that $2^{\bar{J}} \geq n^2$, let $\bar{\Phi}^{nested}_{\alpha/2}$ stand for the test $\bar{\Phi}^{bootFLR}_{\alpha/2}$, defined with the collection of kernels $\{K_J,\ J \in \{0,\dots,\bar{J}\}\}$, and with corresponding weights $w_J = 6/(\pi^2(J+1)^2)$ for J in $\{0,\dots,\bar{J}\}$. Let moreover $\bar{\Phi}^{nonnested}_{\alpha/2}$ be the test $\bar{\Phi}^{bootFLR}_{\alpha/2}$ defined with the collection of kernels $\{K_0\} \cup \{K_{(j,k)},\ j \in \mathbb{N},\ k \in \{0,\dots,2^j-1\}\}$, and with corresponding weights $w_0 = 1/2$, $w_{(j,k)} = 3/(2^j(j+1)^2\pi^2)$ for j in $\mathbb{N}$, k in $\{0,\dots,2^j-1\}$.

Assuming that $\ln\ln n \geq 1$, from Theorem 8, we prove in [10, Corollary 1] that for any $s, s', R, R', R'' > 0$, $SR^{\beta}_{d_2}\big(\bar{\Phi}^{nested}_{\alpha/2} \vee \bar{\Phi}^{nonnested}_{\alpha/2}, \mathcal{B}^{(2)}_{s,2,\infty}(R) \cap w\mathcal{B}^{(2)}_{s'}(R') \cap \big(\mathbb{L}_\infty(R'')\big)^2\big)$ is upper bounded by

(i) $C(s,s',R,R',R'',\alpha,\beta)\,(\ln\ln n/n)^{2s/(4s+1)}$ if $s \geq s'/2$,
(ii) $C(s,s',R,R',R'',\alpha,\beta)\,(\ln n/n)^{s'/(2s'+1)}$ if $s < s'/2$.

In the same way as in Section 1.4, let us introduce

$$\overline{\mathcal{F}^{(2)}_1}^{(i)} = \Big\{ \mathcal{B}^{(2)}_{s,2,\infty}(R) \cap w\mathcal{B}^{(2)}_{s'}(R') \cap \big(\mathbb{L}_\infty(R'')\big)^2,\ s \geq (s'/2) \vee (s'/(2s'+1)) \Big\},$$

and

$$\overline{\mathcal{F}^{(2)}_1}^{(ii)} = \Big\{ \mathcal{B}^{(2)}_{s,2,\infty}(R) \cap w\mathcal{B}^{(2)}_{s'}(R') \cap \big(\mathbb{L}_\infty(R'')\big)^2,\ s < s'/2,\ s' > 1/2 \Big\}.$$

Then, for n large enough, $\bar{\Phi}^{nested}_{\alpha/2} \vee \bar{\Phi}^{nonnested}_{\alpha/2}$ is minimax adaptive over $\overline{\mathcal{F}^{(2)}_1}^{(i)} \cup \overline{\mathcal{F}^{(2)}_1}^{(ii)}$, with no price to pay for adaptivity on $\overline{\mathcal{F}^{(2)}_1}^{(ii)}$. Notice that the price we pay here for adaptivity over $\overline{\mathcal{F}^{(2)}_1}^{(i)}$ involves a $(\ln\ln n)$ factor instead of the usual $(\ln\ln n)^{1/2}$ one, due to the control of the bootstrapped quantiles.

Minimax adaptivity over Sobolev spaces. Let us now assume that $\mathbb{X} = \mathbb{R}^d$, ν is the Lebesgue measure on $\mathbb{R}^d$, and introduce for $s > 0$ the class of alternatives based on a d dimensional Sobolev ball:

$$\mathcal{S}^{(2)}_{s,2,d}(R) = \left\{ (f_1,f_2) \in \mathcal{F}^2,\ \int_{\mathbb{R}^d} \|u\|_d^{2s}\, |\widehat{f_1 - f_2}(u)|^2 d\nu(u) \leq (2\pi)^d R^2 \right\},$$

where $\|u\|_d$ denotes the euclidean norm of u and $\widehat{f_1 - f_2}$ denotes the Fourier transform of $f_1 - f_2$: $\widehat{f_1 - f_2}(u) = \int_{\mathbb{R}^d} (f_1 - f_2)(x) e^{i\langle x,u\rangle} d\nu(x)$.
We here consider the test $\bar{\Phi}^{bootFLR}_{\alpha}$, defined as in Theorem 8 from a collection of approximation kernels adapted to an isotropic class of alternatives. Let $\mathcal{M}_1 = \mathbb{N} \setminus \{0\}$ and $\mathcal{M}_2 = \mathbb{N}$. For m_1 in $\mathcal{M}_1$, k_{m_1} denotes a kernel function in $\mathbb{L}_1(\mathbb{R}^d,\nu) \cap \mathbb{L}_2(\mathbb{R}^d,\nu)$ such that $k_{m_1}(x) = k_{m_1}(-x)$. For m_2 in $\mathcal{M}_2$, let $h_{m_2} = (2^{-m_2},\dots,2^{-m_2})$ and for $m = (m_1,m_2)$ in $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2$, $K_m(x,x') = k_{m_1,h_{m_2}}(x - x')$.
Let $\bar{\Phi}^{Iso}_{\alpha}$ be the test $\bar{\Phi}^{bootFLR}_{\alpha}$ defined with the collection of kernels $\{K_m,\ m \in \mathcal{M}\}$ and $w_m = (6/(\pi^2 m_1(m_2+1)))^2$ for $m = (m_1,m_2)$ in $\mathcal{M}$.
From Theorem 8, we deduce in [10, Corollary 2] that if $\ln\ln n \geq 1$, for any $s, R, R', R'' > 0$, then

$$SR^{\beta}_{d_2}\left(\bar{\Phi}^{Iso}_{\alpha}, \mathcal{S}^{(2)}_{s,2,d}(R) \cap \big(\mathbb{L}_1(R')\big)^2 \cap \big(\mathbb{L}_\infty(R'')\big)^2\right) \leq C(s,\alpha,\beta,R,R',R'',d)\left(\frac{\ln\ln n}{n}\right)^{\frac{2s}{4s+d}}.$$

When $d = 1$, the rate $n^{-2s/(4s+d)}$ is known to be the minimax separation rate, and an extra $(\ln\ln n)^{1/2}$ factor the price to pay for adaptivity, in many models over Sobolev classes (see Chapter 1). Hence, the test $\bar{\Phi}^{Iso}_{\alpha}$ can be said to be minimax adaptive over the collection of classes $\mathcal{S}^{(2)}_{s,2,d}(R) \cap (\mathbb{L}_1(R'))^2 \cap (\mathbb{L}_\infty(R''))^2$, for $s, R, R', R'' > 0$, when $d = 1$, with a slightly more important loss of efficiency due to

adaptivity than usual. From the results of [HS01], it can be conjectured that this is also true when $d > 1$.
Notice that the test $\bar{\Phi}_\alpha^{Iso}$ can also be defined with nonintegrable kernels such as Pinsker's kernel or the sinc kernel (see [Tsy09] or [10]), so that it still satisfies the same result.

Minimax adaptivity over anisotropic Nikol'skii-Besov spaces. Let us still assume that $\mathbb{X} = \mathbb{R}^d$ and ν is the Lebesgue measure on $\mathbb{R}^d$. We consider anisotropic classes of alternatives, more precisely classes of alternatives (f_1, f_2) such that the difference $(f_1 - f_2)$ belongs to a Nikol'skii-Besov ball. Let for $s = (s_1, \dots, s_d)$ in $(0, +\infty)^d$, and $R > 0$,

$$\mathcal{N}_{s,2,d}^{(2)}(R) = \Big\{ (f_1, f_2) \in \mathcal{F}^2,\ (f_1 - f_2) \text{ has continuous partial derivatives } D_i^{\lfloor s_i \rfloor} \\ \text{of order } \lfloor s_i \rfloor \text{ w.r.t } u_i, \text{ and } \forall i = 1 \dots d,\ u_1, \dots, u_d, v \in \mathbb{R}, \\ \left\| D_i^{\lfloor s_i \rfloor}(f_1 - f_2)(u_1, \dots, u_i + v, \dots, u_d) - D_i^{\lfloor s_i \rfloor}(f_1 - f_2)(u_1, \dots, u_d) \right\|_2 \leq R|v|^{s_i - \lfloor s_i \rfloor} \Big\}.$$

We introduce another test of the form $\bar{\Phi}_\alpha^{bootFLR}$, defined as in Theorem 8 from a collection of approximation kernels, but adapted here to anisotropic spaces.
Let $\Sigma = (\Sigma_1, \dots, \Sigma_d)$, where each Σ_i is a positive integer, and let k_1 be a kernel function such that for $x = (x_1, \dots, x_d)$ in $\mathbb{R}^d$, $k_1(x) = \prod_{i=1}^d k_{1,i}(x_i)$, with, for every $i = 1 \dots d$ and $j = 1 \dots \Sigma_i$,

— $k_{1,i}(x_i) = k_{1,i}(-x_i)$,
— $k_{1,i} \in \mathbb{L}_1(\mathbb{R}, \lambda) \cap \mathbb{L}_2(\mathbb{R}, \lambda)$,
— $\int_{\mathbb{R}} k_{1,i}(x_i) d\lambda(x_i) = 1$,
— $\int_{\mathbb{R}} |k_{1,i}(x_i)||x_i|^{\Sigma_i} d\lambda(x_i) < +\infty$
— $\int_{\mathbb{R}} k_{1,i}(x_i) x_i^j d\lambda(x_i) = 0$.

We set $\mathcal{M}_1 = \{1\}$ and $\mathcal{M}_2 = \mathbb{N}^d$, and for $m_2 = (m_{2,1}, \dots, m_{2,d})$ in $\mathcal{M}_2$, $h_{m_2,i} = 2^{-m_{2,i}}$. For $m = (m_1, m_2)$ in $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2$, let $K_m(x, x') = k_{m_1, h_{m_2}}(x - x')$. The test $\bar{\Phi}_\alpha^{Aniso}$ is of the form $\bar{\Phi}_\alpha^{bootFLR}$, defined with the collection of kernels $\{K_m,\ m \in \mathcal{M}\}$ and $w_{(1,m_2)} = \prod_{i=1}^d (6/(\pi^2(m_{2,i}+1)^2))$ for m_2 in $\mathcal{M}_2$. From Theorem 8, we deduce in [10, Corollary 3] that if $\ln \ln n \geq 1$, for $s = (s_1, \dots, s_d)$ in $\prod_{i=1}^d (0, \Sigma_i]$, $R, R', R'' > 0$,

$$SR_{d_2}^\beta \left(\bar{\Phi}_\alpha^{Aniso}, \mathcal{N}_{s,2,d}^{(2)}(R) \cap \left(\mathbb{L}_1(R') \right)^2 \cap \left(\mathbb{L}_\infty(R'') \right)^2 \right) \leq C(s, \alpha, \beta, R, R', R'', d) \left(\frac{\ln \ln n}{n} \right)^{\frac{2\bar{s}}{4\bar{s}+1}},$$

with $1/\bar{s} = \sum_{i=1}^d 1/s_i$.
When $d = 1$, from [Ing00], we know that in the density model, for the problem of testing uniformity, the minimax adaptive separation rate over a Nikol'skii class with smoothness parameter s is of order $(\sqrt{\ln \ln n}/n)^{2s/(4s+1)}$. We find here an upper bound similar to this univariate rate, but where s is replaced by $\bar{s}$, which does not seem unusual (see [GL11] in estimation problems) and $\sqrt{\ln \ln n}$ replaced by $(\ln \ln n)$ (due to the control of the bootstrapped quantile as in the two above cases). Notice that the minimax separation rates obtained in [IS11] over anisotropic periodic Sobolev balls, but in the Gaussian white noise model, are of order $(\sqrt{\ln \ln n}/n)^{2\bar{s}/(4\bar{s}+1)}$.

3.2.3 Experimental results

A simulation study has been proposed in [11] and completed in [8] and [9], to evaluate the performance of the above aggregated tests based on either projection or approximation kernels.
We compare these tests with existing conditional tests, that is tests initially devoted to the two-sample problem in the density model, usually referred to as homogeneity tests, used given the number of observed points of the Poisson processes N_1 and N_2. Note that such conditional tests only allow to test the proportionality (and not the equality) of two intensities.

We first consider univariate frameworks, with $\mathbb{X} = [0,1]$ or $\mathbb{X} = \mathbb{R}$, $n = 100$, $\nu = \lambda$ and $\alpha = 0.05$. In these cases, we focus on $\bar{\Phi}_\alpha^{nested}$, $\bar{\Phi}_\alpha^{nonnested}$, and $\bar{\Phi}_\alpha^{Iso}$ defined in the above section. The test $\bar{\Phi}_\alpha^{nested}$ is implemented with $\bar{J} = 6$ and $\bar{\Phi}_\alpha^{nonnested}$ with a finite collection of kernels $\{K_0\} \cup \left\{ K_{(j,k)},\ j \in \{0,\dots,6\},\ k \in \{0,\dots,2^j-1\} \right\}$. The test $\bar{\Phi}_\alpha^{Iso}$ is implemented with either Gaussian, or Epanechnikov kernels, with a collection of bandwidths equal to $\{1/24, 1/16, 1/12, 1/8, 1/4, 1/2\}$ and some weights that are all equal to 1/6.
These tests are compared with a conditional Kolmogorov-Smirnov test and a conditional MMD test, based on a Gaussian kernel with a heuristic choice for the parameter of the kernel, and a critical value obtained from Efron's bootstrap approach as in [GBR+08], whose corresponding matlab code is available on Gretton's web page `http://www.gatsby.ucl.ac.uk/~gretton/software.html`.
The first kind error rates or sizes are estimated by Monte Carlo methods for several forms of intensities, taken as the densities of the uniform distribution on $[0,1]$, a Beta distribution on $[0,1]$, the standard Gaussian distribution on $\mathbb{R}$, and a Laplace distribution on $\mathbb{R}$. The obtained estimated sizes of the tests fluctuate between: 0.042 and 0.053 for the conditional Kolmogorov-Smirnov test, 0.048 and 0.052 for the MMD test, 0.047 and 0.049 for $\bar{\Phi}_\alpha^{nested}$, 0.043 and 0.045 for $\bar{\Phi}_\alpha^{nonnested}$, 0.051 and 0.054 for $\bar{\Phi}_\alpha^{Iso}$ with the Gaussian kernel, 0.05 and 0.55 for $\bar{\Phi}_\alpha^{Iso}$ with the Epanechnikov kernel.
The second kind error rates and powers are estimated for f_1 equal to the above densities, and f_2 equal to alternatives to these densities. The main point that we observe is that when $f_1 - f_2$ is very irregular, our tests perform better, even sometimes much better, than the two other ones, and that it is particularly true for the test $\bar{\Phi}_\alpha^{Iso}$ with the Epanechnikov kernel. The only case where the MMD test clearly outperforms ours involves more regular differences of intensities.

We then consider multivariate frameworks, with $\mathbb{X} = [0,1]^2$ or $\mathbb{X} = \mathbb{R}^2$, $n = 200$, $\nu = \lambda$ and $\alpha = 0.05$. We here focus on $\bar{\Phi}_\alpha^{Iso}$ with Gaussian and Epanechnikov kernels, with a collection of bandwidths still equal to $\{1/24, 1/16, 1/12, 1/8, 1/4, 1/2\}$ and the weights still equal to 1/6. These tests are compared with the conditional MMD test as above, but also with the conditional Cramer test proposed in [BF04], and a variant proposed in [Bah96], both available in the package `cramer` of `R`.
The sizes are estimated by Monte Carlo methods for two different intensities, taken as the densities of the uniform distribution on $[0,1]^2$, and the standard Gaussian distribution on $\mathbb{R}^2$.
The obtained estimated sizes of the tests fluctuate between 0.043 and 0.06 for the MMD test, 0.048 and 0.052 for the Cramer test, 0.046 and 0.052 for the Bahr test, and are around 0.0485 and 0.046 for $\bar{\Phi}_\alpha^{Iso}$ with the Gaussian kernel and the Epanechnikov kernel respectively.
As for the estimated powers, with f_1 equal to the uniform and standard Gaussian densities, and f_2 equal to various alternatives to these densities, we observe the same phenomenon as in the univariate case. The test $\bar{\Phi}_\alpha^{Iso}$ is still more powerful than the other ones when $f_1 - f_2$ is irregular.

In situations where the user does not know whether the difference between the underlying intensities of the Poisson processes are irregular or not, a good compromise would be to aggregate several of the studied tests, for instance the MMD test and $\bar{\Phi}_\alpha^{Iso}$ with the Epanechnikov kernel.

Our tests were used in a real data study for l'INSEE in [8], about the spatial representativeness of services like schools, medical services, pharmacies, shops, restaurants, or banks in the city of Rennes. The main questions were to know whether two different services can be assumed to be identically spatially distributed, and whether the spatial distribution of one particular service is homogeneous with respect to houses, in the whole city or in a restricted area.

3.2.4 Tools and sketch of proof

Theorem 7 is central to the study, since the oracle type inequalities given in Theorem 8 directly follow from it. We give here a sketch of proof for this theorem, which is rather general, and can be applied to other single tests based on exact bootstrap approaches (see the PhD thesis of Mélisande Albert for instance).

Denoting by $q_{m,\alpha}(1-\beta/2)$ the $(1-\beta/2)$ quantile of the conditional quantile $q_m^{(\bar{\mathbf{X}})}(1-\alpha)$, one has

$$\begin{aligned} P_{f_1,f_2}\left(\phi_{m,\alpha}=0\right) &= P_{f_1,f_2}\left(T_m \le q_m^{(\bar{\mathbf{X}})}(1-\alpha)\right) \\ &\le \frac{\beta}{2} + P_{f_1,f_2}\left(T_m \le q_{m,\alpha}(1-\beta/2)\right). \end{aligned}$$

[First step] The first step of the proof is to find an upper bound for $q_{m,\alpha}(1-\beta/2)$.
Given $\bar{\mathbf{X}}$, T_m^{ε} is a homogeneous Rademacher chaos, as defined by de la Peña and Giné [dlPG99]. From [dlPG99, Corollary 3.2.6] and Markov's inequality, we deduce that there exists $\kappa > 0$ such that

$$q_m^{(\bar{\mathbf{X}})}(1-\alpha) \le \kappa \ln(2/\alpha) \left(\sum_{i,j\in\{1,\dots,N\},\ i\neq j} K_m^2(X_i,X_j) \right)^{1/2}.$$

So $q_{m,\alpha}(1-\beta/2)$ is upper bounded by the $(1-\beta/2)$ quantile of $\kappa\ln(2/\alpha)\big(\sum_{i,j\in\{1,\dots,N\},i\neq j} K_m^2(X_i,X_j)\big)^{1/2}$. Using Markov's inequality again and [DVJ08, Lemma 5.4 III] on factorial moments measures, we obtain that

$$q_{m,\alpha}(1-\beta/2) \le \kappa \ln(2/\alpha)\, n\sqrt{\frac{2C_m}{\beta}}.$$

[Second step] The second step consists in deducing from the above control of $q_{m,\alpha}(1-\beta/2)$ and a concentration inequality for T_m, a condition guaranteeing that $P_{f_1,f_2}\left(T_m \le q_{m,\alpha}(1-\beta/2)\right) \le \beta/2$. From the above first step, we deduce on the one hand that

$$P_{f_1,f_2}\left(T_m \le q_{m,\alpha}(1-\beta/2)\right) \le P_{f_1,f_2}\left(T_m \le \kappa \ln(2/\alpha)\, n\sqrt{\frac{2C_m}{\beta}}\right).$$

Now, on the other hand, from Markov's inequality, it is obvious that

$$P_{f_1,f_2}\left(T_m \le \mathbb{E}_{f_1,f_2}[T_m] - \sqrt{\frac{2\mathrm{Var}_{f_1,f_2}[T_m]}{\beta}}\right) \le \beta/2.$$

Hence, a condition guaranteeing that $P_{f_1,f_2}\left(T_m \le q_{m,\alpha}(1-\beta/2)\right)$ is less than $\beta/2$, and therefore $P_{f_1,f_2}\left(\phi_{m,\alpha}=0\right) \le \beta$, can be expressed as

$$\kappa \ln(2/\alpha)\, n\sqrt{\frac{2C_m}{\beta}} \le \mathbb{E}_{f_1,f_2}[T_m] - \sqrt{\frac{2\mathrm{Var}_{f_2,f_2}[T_m]}{\beta}}.$$

As $\mathbb{E}_{f_1,f_2}[T_m] = n^2\langle K_m \diamond (f_1-f_2), f_1-f_2\rangle_2$ and (by [DVJ08, Lemma 5.4 III] again)

$$\mathrm{Var}_{f_2,f_2}[T_m] \le 4n^3 \|K_m \diamond (f_1-f_2)\|_2^2 \|f_1+f_2\|_\infty + 2nC_m,$$

this condition can be replaced by

$$n^2\langle K_m \diamond (f_1-f_2), f_1-f_2\rangle_2 \ge \kappa \ln(2/\alpha)\, n\sqrt{\frac{2C_m}{\beta}} + \sqrt{\frac{8n^3 \|K_m \diamond (f_1-f_2)\|_2^2 \|f_1+f_2\|_\infty + 4nC_m}{\beta}}.$$

[Third step] Basic inequalities are finally used to give to the above condition, the form which appears in Theorem 7.

Concentration inequalities are therefore the key tools of our study. Here we use Markov's inequality and the exponential inequality of [dlPG99] to control the quantile of the Rademacher chaos. Using more precise inequalities such as the concentration inequality for U-statistics of order 2 of [HRB03] instead of Markov's inequality, and the inequality of [Lat99] instead of the one of [dlPG99], would slightly improve the obtained results, but would not change the order of the final uniform separation rates for the aggregated tests.

3.3 Kernel methods in density and regression models

In this section, we focus on two-sample problems studied in [9], with Béatrice Laurent, Matthieu Lerasle, and Patricia Reynaud-Bouret. We consider the following density and heteroscedastic regression models.

$\mathcal{M}^{(2)}_{\text{density}}$	$\mathbf{X} = (\mathbf{X}^1, \mathbf{X}^2)$ is a pair of independent sets $\mathbf{X}^1 = \{X^1_1, \dots, X^1_{N_1}\}$ and $\mathbf{X}^2 = \{X^2_1, \dots, X^2_{N_2}\}$ (N_1 and N_2 are fixed positive integers) of independent random variables, observed on a measurable space $\mathbb{X}$, with respective densities f_1 and f_2, with respect to a nonatomic σ-finite measure ν on $\mathbb{X}$.

The measure ν may typically be the Lebesgue measure λ when $\mathbb{X}$ is a measurable subset of $\mathbb{R}^d$.
Note that this density model corresponds to a conditional Poisson process model $\mathcal{M}^{(2)}_{\text{Poisson}}$, given $\#\mathbf{X}^1 = N_1$ and $\#\mathbf{X}^2 = N_2$. Conversely, a Poisson process model can be viewed as a density model, but with random sizes N_1 and N_2, with Poisson distributions.
This parallel may be useful, from both theoretical and practical points of view. We have already seen that a Poissonization trick may be used to study Efron's bootstrap approach, in order to construct i.i.d weights close to the multinomial weights of Efron's bootstrap resampling plan. It also allows to easily simulate Poisson processes of $\mathcal{M}^{(2)}_{\text{Poisson}}$, by a two step simulation procedure. The first step consists in simulating the Poisson variables N_1 and N_2, and the second step in simulating i.i.d. samples $\left(X^1_1, \dots, X^1_{N_1}\right)$ and $\left(X^2_1, \dots, X^2_{N_2}\right)$ with densities $f_1/\int_{\mathbb{X}} f_1 d\nu$ and $f_2/\int_{\mathbb{X}} f_2 d\nu$ with respect to ν. This procedure was used in the simulation study described in Section 3.2.3.

$\mathcal{M}^{(2)}_{\text{regression}}$	$\mathbf{X} = (\mathbf{X}^1, \mathbf{X}^2)$ is a pair of independent sets $\mathbf{X}^1 = \{X^1_1, \dots, X^1_{N_1}\}$ and $\mathbf{X}^2 = \{X^2_1, \dots, X^2_{N_2}\}$ (N_1 and N_2 are fixed positive integers) of independent random variables, such that: for every $i \in \{1, \dots, N_1\}$ $X^1_i = (Y^1_i, Z^1_i)$, with $Z^1_i = f_1(Y^1_i) + \sigma(Y^1_i)\xi^1_i$, and for every $i \in \{1, \dots, N_2\}$, $X^2_i = (Y^2_i, Z^2_i)$, with $Z^2_i = f_2(Y^2_i) + \sigma(Y^2_i)\xi^2_i$. The Y^1_i's and Y^2_i's are observed in a measurable space $\mathbb{Y}$, with known distribution P_Y, and Z^1_i and Z^2_i take their values in a measurable subset of $\mathbb{R}$. The couples (Y^1_i, ξ^1_i) and (Y^2_i, ξ^2_i) are identically distributed, and $\mathbb{E}[\xi^1_i | Y^1_i] = 0$, $\mathbb{E}[(\xi^1_i)^2 | Y^1_i] = 1$.

We assume that f_1 and f_2 both belong to $\mathcal{F} = \mathbb{L}_1(\mathbb{X}, \nu) \cap \mathbb{L}_\infty(\mathbb{X}) \subset \mathbb{L}_2(\mathbb{X}, \nu)$ in the density model, to $\mathcal{F} = \mathbb{L}_1(\mathbb{Y}, P_Y) \cap \mathbb{L}_\infty(\mathbb{Y}) \subset \mathbb{L}_2(\mathbb{Y}, P_Y)$ in the regression model. $\mathbb{L}_2(\mathbb{X}, \nu)$ and $\mathbb{L}_2(\mathbb{Y}, P_Y)$ are endowed with their classical norm $\|.\|_2$ and scalar product $\langle ., . \rangle_2$ as in Section 3.1.

We consider the problem of testing, from the observation of $\mathbf{X} = (\mathbf{X}^1, \mathbf{X}^2)$,

$$(H_0) \quad f_1 = f_2 \quad \text{against} \quad (H_1) \quad f_1 \neq f_2.$$

Since the variance function σ^2 is the same for both signals in the regression model, the corresponding two-sample problem amounts to the problem of testing the equality of densities for the samples $(X^1_1, \dots, X^2_{N_1})$ and $(X^2_1, \dots, X^2_{N_2})$. The two-sample problems in the above two models are therefore equivalent, and this is the reason why they are presented together.
The notation $\bar{\mathbf{X}} = \{X_1, \dots, X_N\}$ still stands for the pooled set $\mathbf{X}^1 \cup \mathbf{X}^2$, with cardinality $N = N_1 + N_2$, with $X_i = (Y_i, Z_i)$ in the regression model.

3.3.1 Kernel-based test statistics

In order to shorten the mathematical expressions of the present section, let us fix here a few other notations. We set

$$a_{N_1,N_2} = \left(\frac{1}{N_1(N_1-1)} - c_{N_1,N_2}\right)^{1/2} \text{ and } b_{N_1,N_2} = -a_{N_2,N_1} = -\left(\frac{1}{N_2(N_2-1)} - c_{N_1,N_2}\right)^{1/2},$$

where $c_{N_1,N_2} = 1/\left(N_1N_2(N_1+N_2-2)\right)$. Then, for every i in $\{1,\dots,N\}$, let ε_i^0 be a mark defined by $\varepsilon_i^0 = a_{N_1,N_2}$ if $X_i \in \mathbf{X}^1$ and $\varepsilon_i^0 = b_{N_1,N_2}$ if $X_i \in \mathbf{X}^2$.
Let K_m be a symmetric kernel as in *[Projection kernel]*, *[Approximation kernel]*, or *[Reproducing kernel]*, and satisfying (3.3), replacing $\mathbb{Y}$ by $\mathbb{X}$ in the density model and ν by P_Y in the regression model.
As above, a single test statistic of $(H_{0,m})$ against $(H_{1,m})$ and therefore of (H_0) against (H_1) is obtained by constructing an unbiased estimator of $\langle K_m \diamond (f_1-f_2), f_1-f_2\rangle_2^2$.
Inspired by the above expression of T_m, we consider:

$$\dot{T}_m = \sum_{i,j\in\{1,\dots,N\},\ i\neq j} K_m(X_i,X_j)\left(\varepsilon_i^0\varepsilon_j^0 + c_{N_1,N_2}\right), \tag{3.12}$$

in the density model, or

$$\dot{T}_m = \sum_{i,j\in\{1,\dots,N\},\ i\neq j} Z_iZ_jK_m(Y_i,Y_j)\left(\varepsilon_i^0\varepsilon_j^0 + c_{N_1,N_2}\right),$$

in the regression model. In [9, Proposition 1], $\dot{T}_m$ is proved to be an unbiased estimator of $\langle K_m \diamond (f_1-f_2), f_1-f_2\rangle_2$ in both models $\mathcal{M}^{(2)}_{\text{density}}$ and $\mathcal{M}^{(2)}_{\text{regression}}$, and therefore be a reasonable test statistic.

3.3.2 The permutation approach

It is well-known that the permutation approach is particularly adapted to the two-sample problem in the density problem, and that it defines an exact bootstrap approach. Hence, we introduce a random permutation Π_N uniformly distributed on the group $\mathfrak{S}_N$ of permutations of the set $\{1,\dots,N\}$, and we set

$$\varepsilon_i = \begin{cases} a_{N_1,N_2} & \text{if } \Pi_N(i) \in \{1,\dots,N_1\}, \\ b_{N_1,N_2} & \text{if } \Pi_N(i) \in \{N_1+1,\dots,N\}. \end{cases}$$

We then define:

$$\dot{T}_m^{\varepsilon} = \sum_{i,j\in\{1,\dots,N\},\ i\neq j} K_m(X_i,X_j)\left(\varepsilon_i\varepsilon_j + c_{N_1,N_2}\right),$$

in the density model, or

$$\dot{T}_m^{\varepsilon} = \sum_{i,j\in\{1,\dots,N\},\ i\neq j} Z_iZ_jK_m(Y_i,Y_j)\left(\varepsilon_i\varepsilon_j + c_{N_1,N_2}\right),$$

in the regression model.

One can check in both models that under (H_0), the conditional distribution of the permuted statistic $\dot{T}_m^{\varepsilon}$ given $\bar{\mathbf{X}}$, is equal to the conditional distribution of $\dot{T}_m$ given $\bar{\mathbf{X}}$, that is the permutation approach described here satisfies the conditional invariance property defined in Section 3.1.
Hence, given a prescribed level α in $(0,1)$, denote by $\dot{q}_m^{(\bar{\mathbf{X}})}(1-\alpha)$ the $(1-\alpha)$ quantile of the conditional distribution of $\dot{T}_m^{\varepsilon}$ given $\bar{\mathbf{X}}$, and consider the test

$$\dot{\phi}_{m,\alpha} = \mathbb{1}_{\left\{\dot{T}_m > \dot{q}_m^{(\bar{\mathbf{X}})}(1-\alpha)\right\}}. \tag{3.13}$$

As explained in Section 3.1, from the conditional invariance property, it is easily deduced that the test $\dot{\phi}_{m,\alpha}$ satisfies $(\mathcal{P}_{\text{level},\alpha})$.
Note that this test can in fact be expressed in a simpler manner, removing the term c_{N_1,N_2} in the initial test statistic $\dot{T}_m$ as well as in the permuted one $\dot{T}_m^{\varepsilon}$.

3.3.3 Kernel-based tests with Monte Carlo approximation

The issue of the Monte Carlo approximation of $\dot{q}_m^{(\bar{\mathbf{X}})}(1-\alpha)$ can be solved in the same way as in Section 3.2 as concerns the first kind error rate control. Indeed, considering a fixed number B of Monte Carlo iterations, large enough so that $\alpha(B+1) \geq 1$, let us introduce an i.i.d. sample of B independent random permutations $\{\Pi_N^b,\ b \in \{1,\dots,B\}\}$ uniformly distributed on the group $\mathfrak{S}_N$ of permutations of the set $\{1,\dots,N\}$, independent of $\bar{\mathbf{X}}$. Let us set:

$$\varepsilon_i^b = \begin{cases} a_{N_1,N_2} & \text{if } \Pi_N^b(i) \in \{1,\dots,N_1\}, \\ b_{N_1,N_2} & \text{if } \Pi_N^b(i) \in \{N_1+1,\dots,N\}, \end{cases}$$

and for b in $\{1,\dots,B\}$,

$$\dot{T}_m^{\varepsilon^b} = \sum_{i,j\in\{1,\dots,N\},\ i\neq j} K_m(X_i,X_j)\left(\varepsilon_i^b\varepsilon_j^b + c_{N_1,N_2}\right),$$

in the density model, or

$$\dot{T}_m^{\varepsilon^b} = \sum_{i,j\in\{1,\dots,N\},\ i\neq j} Z_iZ_jK_m(Y_i,Y_j)\left(\varepsilon_i^b\varepsilon_j^b + c_{N_1,N_2}\right),$$

in the regression model.
Denoting by $\left(\dot{T}_m^{\varepsilon^{(1)}},\dots,\dot{T}_m^{\varepsilon^{(B+1)}}\right)$ the order statistic associated with $\left(\dot{T}_m^{\varepsilon^1},\dots,\dot{T}_m^{\varepsilon^B},\dot{T}_m\right)$, given $\bar{\mathbf{X}}$, we set

$$\dot{q}_m^{MC(\bar{\mathbf{X}})}(1-\alpha) = \dot{T}_m^{\varepsilon^{(\lceil (1-\alpha)(B+1)\rceil)}}.$$

Considering the Monte Carlo single kernel-based test defined by

$$\dot{\phi}_{m,\alpha}^{MC} = \mathbb{1}_{\left\{\dot{T}_m > \dot{q}_m^{MC(\bar{\mathbf{X}})}(1-\alpha)\right\}}, \tag{3.14}$$

then Romano and Wolf's lemma (see Lemma 4 above) allows to prove that $\dot{\phi}_{m,\alpha}^{MC}$ satisfies $(\mathcal{P}_{\text{level},\alpha})$.

3.3.4 Aggregated tests

Of course, these single tests are not intended to be used as such, but in an aggregated test. Instead of a single kernel K_m, that the user would have to choose, we therefore consider a collection of such kernels $\{K_m,\ m \in \mathcal{M}\}$, and the aggregated test $\bar{\Phi}_\alpha^{bootFLR}$ given by (3.2), based on the single tests $\dot{\phi}_{m,\alpha}$ defined in (3.13), and the above permutation approach. This aggregated test still satisfies $(\mathcal{P}_{\text{level},\alpha})$. The study of its second kind error related properties is the object of a current work.

3.4 Nearest Neighbors methods in the density model

This section is devoted to a current work with Christine Tuleau-Malot [16], on k-Nearest Neighbors based tests for the two-sample problem in the density model $\mathcal{M}_{\text{density}}^{(2)}$.
We first consider the multivariate two-sample problem where $\mathbb{X} = \mathbb{R}^d$, and ν is the Lebesgue measure on $\mathbb{X}$, focusing on the tests proposed by Schilling [Sch86] and Henze [Hen88]. Hence, we consider single tests with a test statistic based on a k-Nearest Neighbors method, with a fixed k, and a critical value constructed via a permutation approach. The crucial, but left open in these papers, question when dealing with such k-Nearest Neighbors tests, is the one of the choice of k.
In the spirit of the above aggregated tests based on kernel methods, we propose new testing procedures which overcome this question. Instead of considering a particular number k of nearest neighbors, we consider a whole collection of possible values for k, and the corresponding collection of tests, which allows then to construct an aggregated test. This testing procedure is furthermore adapted to handle the two-sample problem for functional data.

3.4.1 k-Nearest Neighbors tests

We keep the same notation as in sections 3.1 and 3.3. In particular, $\bar{\mathbf{X}} = \{ X_1, \ldots, X_N \}$ still stands for the pooled set $\mathbf{X}^1 \cup \mathbf{X}^2$, with cardinality $N = N_1 + N_2$. Here $\mathbb{X} = \mathbb{R}^d$ and ν is the Lebesgue measure. Consider a fixed but arbitrary norm $\|.\|$ on $\mathbb{R}^d$. The random vector X_j is called the rth nearest neighbor to X_i if $\|X_l - X_i\| < \|X_j - X_i\|$ for exactly $r-1$ values of l ($1 \leq l \leq N$). It is denoted by $N_r(X_i)$, and linked with the indicator variable $I_i(r)$ defined by

$$I_i(r) = \begin{cases} 1 & \text{if } X_i \text{ and } N_r(X_i) \text{ are both elements of either } \mathbf{X}^1 \text{ or } \mathbf{X}^2, \\ 0 & \text{otherwise.} \end{cases}$$

Ties are neglected since they occur with probability zero. When ties however occur in practice, for instance because of limited resolution in measurement scales or rounding, they can be ranked as neighbors at random or in the same order as their indices (like in Section 2.3.1), without changing the validity of the results.
The k-Nearest Neighbors (k-NN) test statistic introduced by Schilling [Sch86], and then further studied by Henze [Hen88] is defined for k in $\{1, \ldots, N\}$ by

$$\ddot{T}_k = \sum_{i=1}^{N} \sum_{r=1}^{k} I_i(r),$$

and represents the total number of k-Nearest Neighbors type coincidences.
Notice that this test statistic can be related to the test statistic $\dot{T}_m$, defined in (3.12) from a kernel K_m. Indeed, if for every i, k in $\{1, \ldots, N\}$, ε_i^0 stands for a mark equal to 1 if X_i is an element of $\mathbf{X}^1$, -1 if it is an element of $\mathbf{X}^2$, and $\mathcal{N}_k(X_i)$ for the set $\{ N_r(X_i),\ 1 \leq r \leq k \}$ of k-Nearest Neighbors to X_i, $\ddot{T}_k$ can be written as

$$\ddot{T}_k = \sum_{i,j \in \{1,\ldots,N\},\ i \neq j} \frac{\varepsilon_i^0 \varepsilon_j^0 + 1}{2} \mathbb{1}_{X_j \in \mathcal{N}_k(X_i)}. \tag{3.15}$$

It is therefore very close in spirit to the test statistic $\dot{T}_m$, replacing $K_m(X_i, X_j)$ by $\mathbb{1}_{X_j \in \mathcal{N}_k(X_i)}$.
The k-NN test proposed in [Sch86] and [Hen88] consists in rejecting (H_0) when $\ddot{T}_k$ is larger than a critical value constructed by a permutation approach, which can be described as in Section 3.3. Thus, let us introduce a random permutation Π_N uniformly distributed on the group $\mathfrak{S}_N$ of permutations of the set $\{1, \ldots, N\}$, and set

$$\varepsilon_i = \begin{cases} 1 & \text{if } \Pi_N(i) \in \{1, \ldots, N_1\}, \\ -1 & \text{if } \Pi_N(i) \in \{N_1 + 1, \ldots, N\}. \end{cases}$$

We then define:

$$\ddot{T}_k^{\varepsilon} = \sum_{i,j \in \{1,\ldots,N\},\ i \neq j} \frac{\varepsilon_i \varepsilon_j + 1}{2} \mathbb{1}_{X_j \in \mathcal{N}_k(X_i)},$$

and for α in $(0,1)$, we introduce $\ddot{q}_k^{(\bar{\mathbf{X}})}(1-\alpha)$ the $(1-\alpha)$ quantile of the conditional distribution of $\ddot{T}_k^{\varepsilon}$ given $\bar{\mathbf{X}}$. The test proposed in [Sch86] and [Hen88] can be rewritten as

$$\ddot{\phi}_{k,\alpha} = \mathbb{1}_{\left\{ \ddot{T}_k > \ddot{q}_k^{\bar{\mathbf{X}}}(1-\alpha) \right\}}.$$

Let us now assume that $\mathbf{X} = (\mathbf{X}^1, \mathbf{X}^2)$ is defined from independent sequences $(X_i^1)_{i \geq 1}$ and $(X_i^2)_{i \geq 1}$ of i.i.d. random vectors from the distributions with densities f_1 and f_2 respectively, by $\mathbf{X}^1 = \{X_1^1, \ldots, X_{N_1}^1\}$, and $\mathbf{X}^2 = \{X_1^2, \ldots, X_{N_2}^2\}$.

Henze proves that, if $d_{N_1,N_2} = \left(N_1(N_1-1)+N_2(N_2-1)\right)/\left(N-1\right)$, under (H_0), the conditional distribution of $N^{-1/2}\big(\ddot{T}_k^{\epsilon} - kd_{N_1,N_2}\big)$ given $\bar{\mathbf{X}}$ converges, almost surely in $\big((X_i^1)_{i\geq 1},(X_i^2)_{i\geq 1}\big)$, towards the same Gaussian limit distribution as $N^{-1/2}\big(\ddot{T}_k - kd_{N_1,N_2}\big)$, as $N\to+\infty$, $N_1/N\to\tau\in(0,1)$. Since the common limit distribution is Gaussian, it has a continuous c.d.f., so using Lemma 21.2 in [VdV00] combined with Slutsky's lemma, one can prove that $\ddot{\phi}_{k,\alpha}$ is asymptotically of size α, that is

$$\mathbb{P}_{(H_0)}\left(\ddot{\phi}_{k,\alpha}=1\right)\to_{N\to+\infty,N_1/N\to\tau\in(0,1)}\alpha.$$

Henze also proves that it is consistent against any alternative, that is, under (H_1),

$$P_{f_1,f_2}\left(\ddot{\phi}_{k,\alpha}=1\right)\to_{N\to+\infty,N_1/N\to\tau\in(0,1)}1.$$

As in Section 3.3, since the permutation approach described above for the k-NN test statistic $\ddot{T}_k$ satisfies the conditional invariance property, from a nonasymptotic point of view, $\ddot{\phi}_{k,\alpha}$ satisfies $(\mathcal{P}_{\text{level},\alpha})$. Moreover, this property remains valid when the conditional quantile $\ddot{q}_k^{(\bar{\mathbf{X}})}(1-\alpha)$ is approximated with a Monte Carlo method, thanks to Romano and Wolf's lemma (see Lemma 4).

3.4.2 Aggregation of Nearest Neighbors tests

Let us here consider a collection $\mathcal{K}\subset\{1,\dots,N\}$ of reasonable values for k, and the aggregated test $\bar{\Phi}_{\alpha}^{bootFLR}$ defined in (3.2) from the collection of tests $\left\{\ddot{\phi}_{k,\alpha},\ k\in\mathcal{K}\right\}$.
This test then satisfies $(\mathcal{P}_{\text{level},\alpha})$, and from Henze's consistency result, one can deduce that it is also consistent against any alternative.

In [16], we propose a variant of this aggregated test, which allows to handle the two-sample problem for functional data. It is based on the same ideas as in [BBW05] and [3] described in Chapter 2. In a few words, when the observed random variables belong to a functional space, we consider a collection $\mathcal{K}$ of reasonable numbers k of neighbors, a collection of positive integers $\mathcal{D}$, and the corresponding collection of k-NN tests based on the d first coefficients in the expansions of the X_i's in a complete system of the functional space, for k in $\mathcal{K}$ and d in $\mathcal{D}$. This collection of single tests is then aggregated following the same aggregation scheme as for $\bar{\Phi}_{\alpha}^{bootFLR}$.

3.5 Perspectives

Our short term perspectives on two-sample problems are numerous as many of the above presented works are still in progress. The nonasymptotic study of second kind error properties of the proposed tests in the density and regression models poses the difficulty of the precise control of the permutation based quantiles. The key elements of this study are therefore concentration inequalities on permutations, which would play the role of the concentration inequalities for Rademacher chaos used in the Poisson process model. A part of the PhD thesis of Mélisande Albert is devoted to such concentration inequalities, as they are also needed to study Mélisande Albert's independence tests from the nonasymptotic point of view. The issue is not completely solved yet, as her concentration inequalities can only be used with a particular form of test statistics.

The introduction of general kernels in test statistics, like for instance reproducing kernels which were initially dedicated to statistical learning tasks, offers numerous perspectives, which are well beyond the scope of two-sample problems. On the one hand, it allows, by giving an original point of view on distance-based tests (see [SSGF13] for instance), to think about many new test statistics for classical two-sample problems. On the other hand, it also allows to handle two-sample problems in complex models, where the data are not necessary assumed to be in finite dimensional spaces, for instance, micro-arrays data and graphs models.

Chapter 4

Bootstrap and permutation tests of independence

4.1 Introduction

This chapter is devoted to a piece of work with Mélisande Albert during her PhD thesis, Yann Bouret and Patricia Reynaud-Bouret, which is motivated by a dependence detection issue in neuroscience, and which has lead to two articles: [12] (with the supplement [13]) and [15] respectively dealing with the theoretical and practical aspects of the issue.
It is rather atypical in the present dissertation, in the sense that the introduced tests are mainly studied from an asymptotic point of view, except when permutation approaches are considered, and that they are single tests, not resulting from the aggregation principle.
In fact, these single tests are included in a multiple test, which enables to select time windows where a particular dependency structure can be detected. This topic is developed in Chapter 5.

The study of correlations between variables is a key point in data analysis, which places the question of testing whether two real valued random variables or random vectors are independent among the main topics of the statistical literature. From the historical, and much used in practice, Pearson's chi-square test of independence (see [Pea00, Pea11]) to the modern test of [GG10] using kernel methods in the line of statistical learning, many nonparametric tests have been proposed. Of particular interest are the tests based on permutation or bootstrap approaches. Two families of such permutation or bootstrap independence tests may be distinguished at least: the whole family of rank tests including the tests of Hotelling and Pabst [HP36], Kendall [Ken38], Wolfowitz [Wol42] or Hoeffding [Hoe48b] on the one hand, the family of Kolmogorov-Smirnov type tests, like Blum, Kiefer, and Rosenblatt's [BKR61], Romano's [Rom89] or Van der Vaart and Wellner's [VdVW96] ones on the other hand. Given some prescribed first kind error level α in $(0, 1)$, these tests are all proved to be asymptotically of size α. The tests based on permutation are known to be in addition exactly (nonasymptotically meaning) of level α, that is to satisfy (2), for any sample size. Furthermore, some of these tests are proved to be consistent against many alternatives, such as Hoeffding's [Hoe48b] one and the family of Kolmogorov-Smirnov type tests (except the permutation test described in [VdVW96]).
Detecting dependence is also a fundamental old issue in the neuroscientific literature (see e.g., [GP69]). The neuroscience problem we are interested in consists in testing whether two spike trains simultaneously recorded on two different neurons, during n independent trials as described in [GDA10], are independent. A spike train is a set of time occurrences of action potentials for one neuron, the spikes being the time occurrences themselves, commonly accepted as some of the main components of the brain activity (see [Sin93]).
In practice, the real recordings of spike trains are discretized in time, so they belong to finite dimensional spaces. However, due to the record resolution, the dimension of these spaces is so huge (from ten

thousand up to one million) that it is neither realistic nor reasonable to model such recordings of spike trains by finite dimensional variables, and to apply usual independence tests. Several methods, such as the classical Unitary Events method (see [GDA10] and references therein), use a dimension reduction method (the binning data pre-processing), which unfortunately involve a significant loss of information. Modeling the recordings of spike trains by point processes and using, constructing if needed, independence tests specifically dedicated to such point processes thus appear as a more realistic and reasonable solution.

More precisely, let us introduce a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and the set $\mathcal{X}$ of possible values for finite point processes observed on an interval $\mathbb{X}$ of $\mathbb{R}$, that is the set of the countable subsets of $\mathbb{X}$. The set $\mathcal{X}$ is endowed with a metric $d_{\mathcal{X}}$ (see (4.13) below), issued from the Skorokhod topology, that makes it separable, thus defining accordingly borelian sets on $\mathcal{X}$ and by extension on $\mathcal{X}^2$ through the product metric. A pair $X = (X^1, X^2)$ of finite point processes defined on $(\Omega, \mathcal{A}, \mathbb{P})$ and observed on $\mathbb{X}$, has joint distribution P, with marginals P^1 and P^2 if $P(\mathcal{B}) = \mathbb{P}(X \in \mathcal{B})$, $P^1(\mathcal{B}^1) = \mathbb{P}(X^1 \in \mathcal{B}^1)$, and $P^2(\mathcal{B}^2) = \mathbb{P}(X^2 \in \mathcal{B}^2)$, for all Borelian set $\mathcal{B}$ of $\mathcal{X}^2$, and Borelian sets $\mathcal{B}^1$, $\mathcal{B}^2$ of $\mathcal{X}$.
With these definitions, we can now consider the following point processes model.

$\mathcal{M}^{(2)}_{\text{point proc.}}$	$\mathbf{X} = \mathbf{X}_n = (X_1, \dots, X_n)$, where $\big(X_i = (X_i^1, X_i^2)\big)_{i \geq 1}$ is a sequence of i.i.d. pairs of finite point processes defined on $(\Omega, \mathcal{A}, \mathbb{P})$, observed on $\mathbb{X} = [0, 1]$, with joint distribution P, with marginals P_1 and P_2.

Typically, in the neuroscience problem described above, $\mathbf{X}$ models pairs of rescaled spike trains issued from two distinct neurons, simultaneously recorded during n trials. Those trials are conducted on living animals that are repeatedly subject to the same stimulus or that are repeatedly executing the same task, and separated by resting periods (more details about the experimental conditions can be found in Chapter 5 and [15]). In these conditions, it is commonly admitted that the n trials are i.i.d. and that the model $\mathcal{M}^{(2)}_{\text{point proc.}}$ is actually realistic.

From the observation of $\mathbf{X}$, we aim at testing (H_0) X_1^1 and X_1^2 are independent against (H_1) X_1^1 and X_1^2 are not independent, which can also be written as

$$(H_0)\ P = P^1 \otimes P^2 \quad \text{against} \quad (H_1)\ P \neq P^1 \otimes P^2.$$

In our neuroscience problem, this amounts to testing whether the two neurons from which the pairs of spike trains are issued are independent or not.

Notice that asymptotic tests of independence had already been introduced in [TMRGRB14] in the parametric model of homogeneous Poisson processes. Such a parametric framework is necessarily restrictive and even possibly inappropriate since the very existence of any precise underlying distribution for the point processes modeling spike train data is subject to broad debate (see [PC09, RBRGTM14]). To construct nonparametric tests of independence for point processes is therefore of utmost importance in this neuroscience context.
Based on these considerations, particular bootstrap methods under the name of trial-shuffling have been proposed in [PG03, PDG03] for binned data with relatively small dimension, but without proper mathematical justification. Besides the loss of information the binning data pre-processing involves, it appears that the test statistics chosen in these papers do not lead to tests of asymptotic prescribed size as shown and explained in [15]. We propose in [12] new nonparametric tests of independence, in the spirit of the Kolmogorov-Smirnov type tests of [Rom89] and [VdVW96], based on U-statistics and whose critical values are obtained via bootstrap or permutation approaches.
But whereas the tests of [Rom89] and [VdVW96] are based on particular U-statistics for i.i.d. real valued random variables or vectors, ours are based on general U-statistics for i.i.d. pairs of point processes. To our knowledge, there is no other work on the bootstrap or permutation of such general U-statistics for i.i.d. pairs of point processes.

Though the proofs of our results are rather close to the ones in [LN09] for the bootstrap, and inspired by Romano's [Rom87, Rom89] work and Hoeffding's [Hoe52] precursor results on the permutation, an additional difficulty thus lies in the nature of the mathematical objects we handle here, that is point processes and their associated point measures which are random measures.
Furthermore, whereas the convergence of the conditional distribution of the permuted test statistic given the observed sample $\mathbf{X}$, towards the limit distribution of the original test statistic under (H_0), when (H_0) is actually satisfied, can be considered as a rather usual result, the asymptotic behavior of the permuted test statistic under (H_1) is rarely studied, even in more classical settings than point processes. Our result in Theorem 9 solves this question in the present framework. It can thus be viewed as the beginning of an answer to a problem stated as open question in [VdVW96, page 371].
Besides these asymptotic aspects, we focus on the nonasymptotic properties of the permutation approach, in particular when Monte Carlo methods are used to approximate the chosen critical values or quantiles. It has been acknowledged that when both bootstrap and permutation approaches are available, permutation should be preferred, since the corresponding tests are guaranteed to be of the prescribed level. Details are given in Section 4.4. Nevertheless, we keep investigating both approaches together, as bootstrap methods - through trial-shuffling - are the usual references in neuroscience.

4.2 General description of the tests

4.2.1 From neuroscience interpretations to general test statistics

The main dependence feature that neuroscientists expect to detect between two neurons corresponds to synchronization in time, referred to as coincidences [GDA10]. More precisely, neuroscientists aim at assessing if such coincidences occur significantly, that is more than what may be due to chance. They speak in this case of a detected synchrony.
Different kinds of coincidence count functions have been introduced in the neuroscientific literature, such as the binned coincidence count function or its shifted version introduced in [Grü96] and [GDG$^+$99]. We here particularly focus on the notion of coincidence count between two point processes X^1 and X^2 with delay δ ($\delta > 0$) defined in [TMRGRB14] by

$$\varphi_\delta^{coinc}(X^1, X^2) = \int_{[0,1]^2} \mathbb{1}_{|u-v|\leq\delta} dN_{X^1}(u) dN_{X^2}(v) = \sum_{u\in X^1, v\in X^2} \mathbb{1}_{|u-v|\leq\delta}, \tag{4.1}$$

where dN is defined as in (1.20).
In the parametric homogeneous Poisson framework of [TMRGRB14], for every i in $\{1,\dots,n\}$, the expectation of $\varphi_\delta^{coinc}\left(X_i^1, X_i^2\right)$ has a simple expression as a function of δ and the intensities λ_1 and λ_2 of X_i^1 and X_i^2. Since λ_1 and λ_2 can be easily estimated, an estimator of this expectation can thus be obtained using the plug-in principle, and subtracted from $\varphi_\delta^{coinc}\left(X_i^1, X_i^2\right)$ to lead to a test statistic, based on

$$C(\mathbf{X}_n) = \sum_{i=1}^{n} \varphi_\delta^{coinc}\left(X_i^1, X_i^2\right), \tag{4.2}$$

with an asymptotic standard Gaussian distribution under (H_0).
In the present nonparametric framework where as few assumptions as possible on P have to be made, such a centering plug-in tool is not available. We use instead a self-centering trick, which amounts, combined with a rescaling step, to considering the statistic

$$\frac{1}{n(n-1)} \sum_{i\neq i'\in\{1,\dots,n\}} \left(\varphi_\delta^{coinc}\left(X_i^1, X_i^2\right) - \varphi_\delta^{coinc}\left(X_i^1, X_{i'}^2\right)\right). \tag{4.3}$$

As we did not want to restrict our study to the particular synchrony detection problem, noticing that (4.3) can be written as a U-statistic of the i.i.d. sample $\mathbf{X}_n = (X_1, \dots, X_n)$ with a symmetric kernel, as

defined by Hoeffding [Hoe48a], we in fact investigate more general independence test statistics. These test statistics are based on U-statistics of the form

$$U_{n,h}(\mathbf{X}_n) = \frac{1}{n(n-1)} \sum_{i \neq i' \in \{1,\dots,n\}} h\left(X_i, X_{i'}\right), \tag{4.4}$$

where $h : (\mathcal{X}^2)^2 \to \mathbb{R}$ is a symmetric kernel such that:

$(\mathcal{A}_{Cent})$ | For all $n \geq 2$, $U_{n,h}(\mathbf{X}_n)$ is zero mean under (H_0), that is, for X and X' i.i.d. with distribution $P^1 \otimes P^2$, $\mathbb{E}\left[h\left(X, X'\right)\right] = 0$.

The particular case where for $x = (x^1, x^2)$, $y = (y^1, y^2)$,

$$h(x,y) = h_{\varphi_\delta^{coinc}}(x,y) := \frac{1}{2}\left(\varphi_\delta^{coinc}\left(x^1,x^2\right) + \varphi_\delta^{coinc}\left(y^1,y^2\right) - \varphi_\delta^{coinc}\left(x^1,y^2\right) - \varphi_\delta^{coinc}\left(y^1,x^2\right)\right), \tag{4.5}$$

for which $U_{n,h_{\varphi_\delta^{coinc}}}(\mathbf{X}_n)$ is equal to the statistic (4.3) is called the *Coincidence case*.

The extended case where

$$h(x,y) = h_\varphi(x,y) := \frac{1}{2}\left(\varphi\left(x^1,x^2\right) + \varphi\left(y^1,y^2\right) - \varphi\left(x^1,y^2\right) - \varphi\left(y^1,x^2\right)\right), \tag{4.6}$$

for some given integrable function φ, is called the *Linear case*.
In these cases, note that $(\mathcal{A}_{Cent})$ is straightforwardly satisfied and that the statistic $U_{n,h_\varphi}(\mathbf{X}_n)$ is an unbiased estimator of

$$\int\int \varphi\left(x^1,x^2\right)\left(dP(x^1,x^2) - dP^1(x^1)dP^2(x^2)\right),$$

without any distribution assumption on the underlying point processes.
If the X_i's were pairs of finite dimensional variables with continuous distributions w.r.t. the Lebesgue measure, this statistic would be closely related to Kolmogorov-Smirnov type tests of independence. For instance, the test statistics of Blum, Kiefer, and Rosenblatt [BKR61], Romano [Rom89], Van der Vaart and Wellner [VdVW96] are equivalent to

$$\sqrt{n} \sup_{v^1 \in \mathcal{V}^1, v^2 \in \mathcal{V}^2} \left| U_{n,h_{\varphi_{(v^1,v^2)}}}(\mathbf{X}_n) \right|,$$

where, respectively:
- $\mathcal{V}^1 = \mathcal{V}^2 = \mathbb{R}$, $\varphi_{(v^1,v^2)}(x^1,x^2) = \mathbb{1}_{]-\infty,v^1]}(x^1)\mathbb{1}_{]-\infty,v^2]}(x^2)$;
- $\mathcal{V}^1$ and $\mathcal{V}^2$ are countable VC classes of subsets of $\mathbb{R}^d$, $\varphi_{(v^1,v^2)}(x^1,x^2) = \mathbb{1}_{v^1}(x^1)\mathbb{1}_{v^2}(x^2)$;
- $\mathcal{V}^1$ and $\mathcal{V}^2$ are well-chosen classes of real-valued functions, $\varphi_{(v^1,v^2)}(x^1,x^2) = v^1(x^1)v^2(x^2)$.

Besides $(\mathcal{A}_{Cent})$, we also assume that $U_{n,h}(\mathbf{X}_n)$ is nondegenerate under (H_0).

$(\mathcal{A}_{nondeg})$ | For all $n \geq 2$, $U_{n,h}(\mathbf{X}_n)$ is nondegenerate under (H_0), that is, for X and X' i.i.d. with distribution $P^1 \otimes P^2$, $\mathrm{Var}\left[\mathbb{E}\left[h(X,X')|X\right]\right] \neq 0$.

Finally, we also need the following moment assumption, which guarantees that the variance of $U_{n,h}(\mathbf{X}_n)$ exists.

$(\mathcal{A}_{Mmt})$ | For X and X' i.i.d. with distribution P, $\mathbb{E}\left[h^2\left(X,X'\right)\right] < +\infty$.

A detailed discussion about these assumptions, and their relevance in practice, is provided in [12].

4.2.2 A first basic asymptotic test

The asymptotic setup and additional notation. As explained in the introduction above, the point of view that we adopt here is mainly asymptotic, so we consider several kinds of convergence. In particular, using bootstrap and permutation approaches entails taking into account random conditional distributions, given the sample $\mathbf{X}_n$. It is therefore worth clearly presenting the notation used in the following to state our results as concisely as possible.

— For any functional $Z : (\mathcal{X}^2)^n \to \mathbb{R}$, $\mathcal{L}(Z, Q)$ denotes the distribution of $Z(\mathbf{Y}_n)$, where $\mathbf{Y}_n$ is an i.i.d. sample from the distribution Q on $\mathcal{X}^2$. In particular, the distribution of $Z(\mathbf{X}_n)$ under (H_0) is denoted by $\mathcal{L}\left(Z, P^1 \otimes P^2\right)$.

— If the distribution $Q = Q(W)$ depends on a random variable W, $\mathcal{L}(Z, Q|\, W)$ is the conditional distribution of $Z(\mathbf{Y}_n)$, $\mathbf{Y}_n$ being an i.i.d. sample from the distribution $Q = Q(W)$, given W.

— " Q-a.s. in $(X_i)_i$" at the end of a statement means that the statement only depends on the sequence $(X_i)_i$, where the X_i's are i.i.d. with distribution Q, and that there exists an event $\mathcal{C}$ only depending on $(X_i)_i$ such that $\mathbb{P}(\mathcal{C}) = 1$, on which the statement is true. Here, Q is usually equal to P.

— "$Q_n \underset{n\to+\infty}{\Longrightarrow} Q$" means that the sequence of distributions $(Q_n)_n$ converges towards Q in the weak sense, that is for any real valued, continuous and bounded function g, $\int g(z)dQ_n(z) \to_{n\to+\infty} \int g(z)dQ(z)$.

Following the historical paper by Bickel and Freedman [BF81], the closeness between two distributions on $\mathbb{R}$ is here measured via the $\mathbb{L}_2$-Wasserstein's metric d_2 (also called Mallows' metric), equivalent to both weak convergence and convergence of second-order moments.
As often, the real Gaussian distributions have a key role in the limit theorems that we obtain: $\mathcal{N}(\nu, \sigma^2)$ stands for the real Gaussian distribution with mean ν and variance σ^2, F_{ν,σ^2} for its c.d.f. and F^{-1}_{ν,σ^2} for its quantile function.

The central limit theorem for U-statistics of i.i.d. pairs of independent point processes. From the results of Rubin and Vitale [RV80] generalizing Hoeffding's [Hoe48a] decomposition of non-degenerate U-statistics to the case where the X_i's are nonnecessarily real valued random vectors, a central limit theorem for $U_{n,h}(\mathbf{X}_n)$ under (H_0) can be easily derived.
Assume that h satisfies $(\mathcal{A}_{Cent})$, $(\mathcal{A}_{nondeg})$, and $(\mathcal{A}_{Mmt})$, and let

$$\sigma^2_{h,P^1\otimes P^2} = 4\mathrm{Var}\left[\mathbb{E}\left[h\left(X, X'\right)|X\right]\right] \text{ for } X \text{ and } X' \text{ i.i.d. with distribution } P^1 \otimes P^2. \tag{4.7}$$

Then,

$$d_2\left(\mathcal{L}\left(\sqrt{n}U_{n,h}, P^1 \otimes P^2\right), \mathcal{N}(0, \sigma^2_{h,P^1\otimes P^2})\right) \underset{n\to+\infty}{\longrightarrow} 0. \tag{4.8}$$

Considering the unbiased estimator of $\sigma^2_{h,P^1\otimes P^2}$ under (H_0) defined by

$$\Sigma^2(\mathbf{X}_n) = \frac{4}{n(n-1)(n-2)} \sum_{i,j,k\in\{1,\dots,n\},\#\{i,j,k\}=3} h(X_i, X_j)h(X_i, X_k),$$

this in particular leads to the following result:

$$\mathcal{L}\left(\sqrt{n}U_{n,h}/\Sigma, P^1 \otimes P^2\right) \underset{n\to+\infty}{\Longrightarrow} \mathcal{N}(0, 1).$$

Given some prescribed first kind error level α in $(0, 1)$, a simple asymptotic test can thus be defined by

$$\phi_{n,h,\alpha} = \mathbb{1}_{\left\{|\sqrt{n}U_{n,h}(\mathbf{X}_n)| > \Sigma(\mathbf{X}_n)F^{-1}_{0,1}(1-\alpha/2)\right\}}. \tag{4.9}$$

This test is asymptotically of size α, that is, it satisfies

$\left(\mathcal{P}^{\infty}_{\text{size},\alpha}\right)$ $\quad\Big|\quad \mathbb{P}\left(\phi_{n,h,\alpha}=1\right) \to_{n\to+\infty} \alpha$ if $P = P^1 \otimes P^2$.

It is also consistent against any reasonable alternative P, that is, considering the set $\mathcal{P}_1 = \left\{ P,\ \int h(x,x')dP(x)dP(x') \neq 0 \right\}$, it satisfies

$\left(\mathcal{P}^{\infty}_{\text{consist.},\mathcal{P}_1}\right)$ $\quad\Big|\quad \mathbb{P}\left(\phi_{n,h,\alpha}=1\right) \to_{n\to+\infty} 1$ if $P \in \mathcal{P}_1$.

Moreover, similar results hold for the corresponding upper and lower-tailed tests, at the difference that consistency is only obtained on:

— $\mathcal{P}_1^+ = \left\{ P,\ \int h(x,x')dP(x)dP(x') > 0 \right\}$ for the upper-tailed test,
— $\mathcal{P}_1^- = \left\{ P,\ \int h(x,x')dP(x)dP(x') < 0 \right\}$ for the lower-tailed test.

However, such a purely asymptotic test may of course suffer from a lack of power when the sample size n is small or even moderate, which is typically the case for the application in neuroscience we are interested in for biological reasons (from few tens up to few hundreds at best). So following the works of Romano [Rom89] or Van der Vaart and Wellner [VdVW96], we turn to classical bootstrap and permutation approaches, which are known to generally outperform such a simple asymptotic test, especially for small sample sizes.

More precisely, we introduce some tests of the same form as the above asymptotic tests:

$$\begin{cases} \phi^+_{n,h,\alpha} &= \mathbb{1}_{\left\{\sqrt{n}U_{n,h}(\mathbf{X}_n) > c^+_{n,h,\alpha}(\mathbf{X}_n)\right\}} & \text{(upper-tailed test)}, \\ \phi^-_{n,h,\alpha} &= \mathbb{1}_{\left\{\sqrt{n}U_{n,h}(\mathbf{X}_n) < c^-_{n,h,\alpha}(\mathbf{X}_n)\right\}} & \text{(lower-tailed test)}, \\ \phi^{+/-}_{n,h,\alpha} &= \phi^+_{n,h,\alpha/2} \vee \phi^-_{n,h,\alpha/2} & \text{(two-tailed test)}, \end{cases} \tag{4.10}$$

but with random critical values $c^+_{n,h,\alpha}(\mathbf{X}_n)$, $c^-_{n,h,\alpha}(\mathbf{X}_n)$ obtained from bootstrap or permutation approaches. Though these approaches are here mainly justified from an asymptotic point of view as well (except for the exact level achievement when considering permutation), the simulation studies presented in [12] and [15] indeed show their efficiency for small sample sizes.
A nonasymptotic theoretical study, by Mélisande Albert, of permutation tests of independence, but for i.i.d. real valued random variables is in progress, and constitutes a large part of her PhD thesis.

Remark that if $(\mathcal{A}_{nondeg})$ does not hold, $\sigma^2_{h,P^1\otimes P^2} = 0$ and $\sqrt{n}U_{n,h}(\mathbf{X}_n)$ tends in probability towards 0. Indeed, degenerate U-statistics of order 2 have a faster rate of convergence than $\sqrt{n}$ (see [AG93] for instance for explicit limit theorems). Hence, in this case, $\sqrt{n}U_{n,h}(\mathbf{X}_n)$ can not be used as a test statistic anymore. As noticed above, the relevance of $(\mathcal{A}_{nondeg})$ is discussed in [12].

4.3 Bootstrap tests of independence

As seen in (4.8), the limit distribution of the statistic $\sqrt{n}U_{n,h}(\mathbf{X}_n)$ under (H_0) is not free from the unknown underlying marginals P^1 and P^2. The main purpose of the classical bootstrap approach is to construct a conditional distribution, which only depends on the observed sample $\mathbf{X}_n$, but which approximates this unknown distribution, for large, but also moderate or small sample sizes. As each $X_i = (X_i^1, X_i^2)$ is $P^1 \otimes P^2$-distributed under (H_0), the first and second coordinates of the elements of $\mathbf{X}_n$ are independently resampled according to the marginal empirical distributions P_n^j given by

$$\text{for } j = 1, 2, \quad P_n^j = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i^j}. \tag{4.11}$$

More precisely, a bootstrap sample from $\mathbf{X}_n$ is denoted by $\mathbf{X}_n^* = \left(X_{n,1}^*, \ldots, X_{n,n}^*\right)$, with $X_{n,i}^* = (X_{n,i}^{*1}, X_{n,i}^{*2})$, and is here defined as an n i.i.d. sample from the distribution $P_n^1 \otimes P_n^2$.

Then, the bootstrap distribution of interest is the conditional distribution of $\sqrt{n}U_{n,h}(\mathbf{X}_n^*)$ given $\mathbf{X}_n$, that is $\mathcal{L}\left(\sqrt{n}U_{n,h}, P_n^1 \otimes P_n^2 | \mathbf{X}_n\right)$ to be compared with the initial distribution of $\sqrt{n}U_{n,h}(\mathbf{X}_n)$ under (H_0), that is $\mathcal{L}\left(\sqrt{n}U_{n,h}, P^1 \otimes P^2\right)$.
In the following, as usual, $\mathbb{E}^*[\cdot]$ stands for the conditional expectation given $\mathbf{X}_n$.

4.3.1 Consistency of the bootstrap approach

Our results of consistency for the global bootstrap approach were obtained under several additional assumptions, that are rather classical in the bootstrap scene. Nevertheless, as the random variables we deal with are not real-valued variables but point processes, these assumptions may be difficult to interpret in the present setup. Their relevance, in theory as well as in practice is therefore also discussed in details in [12].

In addition to assumption $(\mathcal{A}_{Cent})$, we need its following empirical version:

$(\mathcal{A}^*_{Cent})$ | For $x_1 = (x_1^1, x_1^2), \ldots, x_n = (x_n^1, x_n^2)$ in $\mathcal{X}^2$, $\sum_{i_1,i_2,i_1',i_2' \in \{1,\ldots,n\}} h\big((x_{i_1}^1, x_{i_2}^2), (x_{i_1'}^1, x_{i_2'}^2)\big) = 0$.

Notice that this assumption, as well as $(\mathcal{A}_{Cent})$, is fulfilled in the *Linear case* where h is of the form h_φ given by (4.6).
Furthermore, a stronger moment assumption than $(\mathcal{A}_{Mmt})$ is required, namely

$(\mathcal{A}^*_{Mmt})$ | For X_1, X_2, X_3, X_4 i.i.d. with distribution P on $\mathcal{X}^2$, and for i_1, i_2, i_1', i_2' in $\{1,2,3,4\}$, $\mathbb{E}\big[h^2\big((X_{i_1}^1, X_{i_2}^2), (X_{i_1'}^1, X_{i_2'}^2)\big)\big] < +\infty$.

When $(\mathcal{A}^*_{Mmt})$ is satisfied, this implies that
— $(\mathcal{A}_{Mmt})$ is satisfied (taking $i_1 = i_2$, $i_1' = i_2'$, and $i_1' \neq i_1$),
— for $X \sim P$, $\mathbb{E}\big[h^2(X,X)\big] < +\infty$ (taking $i_1 = i_2 = i_1' = i_2'$),
— for X_1, X_2 i.i.d with distribution $P^1 \otimes P^2$, $\mathbb{E}\big[h^2\big(X_1, X_2\big)\big] < +\infty$ (taking i_1, i_2, i_1', i_2' all different).
Finally, a continuity assumption is required on the kernel h: let us describe the topology we use here. The set $\mathcal{X}$ can be embedded in the space $\mathcal{D}$ of càdlàg functions on $[0,1]$ through the identification $N : x \in \mathcal{X} \mapsto N_x \in \mathcal{D}$, where N_x is the counting process associated with x as defined in (1.19). Now consider the uniform Skorokhod topology on $\mathcal{D}$ (see [Bil09]), associated with the metric $d_{\mathcal{D}}$ defined by

$$d_{\mathcal{D}}(f,g) = \inf\left\{ \varepsilon > 0,\ \exists \lambda \in \Lambda,\ \sup_{t\in[0,1]} |\lambda(t) - t| \leq \varepsilon,\ \sup_{t\in[0,1]} |f(\lambda(t)) - g(t)| \leq \varepsilon \right\}, \tag{4.12}$$

where Λ is the set of strictly increasing, continuous mappings of $[0,1]$ onto itself. Thanks to the identification N above, $\mathcal{X}$ can then be endowed with the topology induced by $d_{\mathcal{X}}$ defined on $\mathcal{X}$ by

$$d_{\mathcal{X}}(x,x') = d_{\mathcal{D}}(N(x), N(x')) \quad \text{for every } x,\ x' \text{ in } \mathcal{X}. \tag{4.13}$$

As an illustration, if x and x' are in $\mathcal{X}$, for ε in $(0,1)$, $d_{\mathcal{X}}(x,x') \leq \varepsilon$ implies that x and x' have the same cardinality, and for k in $\{1,\ldots,\#x\}$, the k^{th} point of x is at distance less than ε from the k^{th} point of x'. Since $(\mathcal{D}, d_{\mathcal{D}})$ is a separable metric space, so are $(\mathcal{X}, d_{\mathcal{X}})$, $\left(\mathcal{X}^2, d_{\mathcal{X}^2}\right)$, where $d_{\mathcal{X}^2}$ is the product metric defined from $d_{\mathcal{X}}$ (see [Dud02, p. 32]), and $\left(\mathcal{X}^2 \times \mathcal{X}^2, d\right)$, where d, the product metric defined from $d_{\mathcal{X}^2}$, is given by

$$d\big((x,y),(x',y')\big) = \sup\left\{ \sup_{j=1,2}\left\{ d_{\mathcal{X}}(x^j, x'^j) \right\}, \sup_{j=1,2}\left\{ d_{\mathcal{X}}(y^j, y'^j) \right\} \right\}, \tag{4.14}$$

for every $x = (x^1, x^2)$, $y = (y^1, y^2)$, $x' = (x'^1, x'^2)$, $y' = (y'^1, y'^2)$ in $\mathcal{X}^2$.

The kernel h chosen to define the U-statistic $U_{n,h}(\mathbf{X}_n)$ in (4.4) is here assumed to satisfy:

$(\mathcal{A}_{Cont})$ | There exists a subset $\mathcal{C}$ of $\mathcal{X}^2 \times \mathcal{X}^2$, such that $(P^1 \otimes P^2)^{\otimes 2}(\mathcal{C}) = 1$ and h is continuous on $\mathcal{C}$ for the topology induced by d.

Notice that in the *Linear case* with

$$\varphi(X^1, X^2) = \varphi^w(X^1, X^2) := \int_{[0,1]^2} w(u,v) dN_{X^1}(u) dN_{X^2}(v), \tag{4.15}$$

for some continuous integrable function w, the kernel h_φ defined by (4.6) is proved to satisfy $(\mathcal{A}_{Cont})$. As for the *Coincidence case*, we also prove that the coincidence count kernel $h_{\varphi_\delta^{coinc}}$ defined on $\mathcal{X}^2 \times \mathcal{X}^2$ by (4.1) and (4.6) satisfies $(\mathcal{A}_{Cont})$ for a reasonable choice of δ.

Let us now state the main results we obtained in [12].

Theorem 9 (Albert, Bouret, Fromont, Reynaud-Bouret, 2015)**.** *Under* $(\mathcal{A}_{Cent})$, $(\mathcal{A}^*_{Cent})$, $(\mathcal{A}^*_{Mmt})$ *and* $(\mathcal{A}_{Cont})$,

$$d_2\left(\mathcal{L}\left(\sqrt{n}U_{n,h}, P_n^1 \otimes P_n^2 \middle| \mathbf{X}_n\right), \mathcal{L}\left(\sqrt{n}U_{n,h}, P^1 \otimes P^2\right)\right) \underset{n\to+\infty}{\longrightarrow} 0, \ P\text{-a.s. in } (X_i)_i.$$

Notice first that the above convergence result holds under (H_0) as well as under (H_1).
Then, its proof does not use (4.8), so here the bootstrap distribution $\mathcal{L}\left(\sqrt{n}U_{n,h}, P_n^1 \otimes P_n^2 \middle| \mathbf{X}_n\right)$ is directly compared to $\mathcal{L}\left(\sqrt{n}U_{n,h}, P^1 \otimes P^2\right)$, without arguing that they both converge to the same Gaussian limit distribution, as often done in papers dealing with the bootstrap. In particular, the assumption $(\mathcal{A}_{nondeg})$ is not needed therefore, when the U-statistic $U_{n,h}(\mathbf{X}_n)$ is degenerate, both considered distributions converge to the Dirac mass in 0.
However, (4.8) plays a crucial role to obtain the following corresponding convergence of the bootstrapped c.d.f. and quantiles.

Corollary 3 (Albert, Bouret, Fromont, Reynaud-Bouret, 2015)**.** *Let* $\mathbf{X}_n^{\perp\!\!\!\perp}$ *be an* n *i.i.d. sample from the distribution* $P^1 \otimes P^2$ *on* $\mathcal{X}^2$*. Under* $(\mathcal{A}_{nondeg})$ *and the assumptions of Theorem 9,*

$$\sup_{z\in\mathbb{R}} \left| \mathbb{P}\left(\sqrt{n}U_{n,h}(\mathbf{X}_n^*) \le z \middle| \mathbf{X}_n\right) - \mathbb{P}\left(\sqrt{n}U_{n,h}(\mathbf{X}_n^{\perp\!\!\!\perp}) \le z\right) \right| \underset{n\to+\infty}{\longrightarrow} 0, \ P\text{-a.s. in } (X_i)_i.$$

If moreover, $q_{n,h}^{*(\mathbf{X}_n)}$ *denotes the conditional quantile function of* $\sqrt{n}U_{n,h}(\mathbf{X}_n^*)$ *given* $\mathbf{X}_n$ *and* $q_{n,h}^{\perp\!\!\!\perp}$ *denotes the quantile function of* $\sqrt{n}U_{n,h}(\mathbf{X}_n^{\perp\!\!\!\perp})$*, for every* u *in* $(0,1)$*,*

$$|q_{n,h}^{*(\mathbf{X}_n)}(u) - q_{n,h}^{\perp\!\!\!\perp}(u)| \underset{n\to+\infty}{\longrightarrow} 0, \ P\text{-a.s. in } (X_i)_i.$$

4.3.2 Asymptotic properties of the bootstrap tests

The result of Corollary 3 is the fundamental point to construct our bootstrap tests. Let $\phi_{n,h,\alpha}^{*,+}$, $\phi_{n,h,\alpha}^{*,-}$ and $\phi_{n,h,\alpha}^{*,+/-}$ respectively be the three tests defined by (4.10) with

$$\begin{cases} c_{n,h,\alpha}^{+}(\mathbf{X}_n) &= q_{n,h}^{*(\mathbf{X}_n)}(1-\alpha), \\ c_{n,h,\alpha}^{-}(\mathbf{X}_n) &= q_{n,h}^{*(\mathbf{X}_n)}(\alpha). \end{cases} \tag{4.16}$$

Note that the quantiles $q_{n,h}^{*(\mathbf{X}_n)}(u)$ are random, depending on $\mathbf{X}_n$, and that they may be exactly computed by considering the n^{2n} possible bootstrap samples. The algorithmic complexity of such an exact computation is usually so large that a Monte Carlo approximation, based on resampling from the original $\mathbf{X}_n$, is preferred in practice. This Monte Carlo step is also considered here. So let $(B_n)_{n\ge2}$ be a sequence of possible numbers of Monte Carlo iterations, such that $B_n \to_{n\to+\infty} +\infty$, and for

$n \geq 1$, let $\left(\mathbf{X}_n^{*1}, \ldots, \mathbf{X}_n^{*B_n}\right)$ be B_n independent bootstrap samples from $\mathbf{X}_n$. Set $(U^{*1}, \ldots, U^{*B_n}) = (U_{n,h}(\mathbf{X}_n^{*1}), \ldots, U_{n,h}(\mathbf{X}_n^{*B_n}))$. Introduce the corresponding order statistic $(U^{*(1)}, \ldots, U^{*(B_n)})$, and let $\phi_{n,h,\alpha}^{*MC,+}$, $\phi_{n,h,\alpha}^{*MC,-}$ and $\phi_{n,h,\alpha}^{*MC,+/-}$ respectively be the tests defined by (4.10) with

$$\left\{ \begin{array}{rcl} c_{n,h,\alpha}^{+}(\mathbf{X}_n) & = & \sqrt{n} U^{*(\lceil (1-\alpha) B_n \rceil)}, \\ c_{n,h,\alpha}^{-}(\mathbf{X}_n) & = & \sqrt{n} U^{*(\lfloor \alpha B_n \rfloor + 1)}. \end{array} \right. \tag{4.17}$$

Theorem 10 (Albert, Bouret, Fromont, Reynaud-Bouret, 2015). *If* $(\mathcal{A}_{nondeg})$, $(\mathcal{A}_{Cent})$, $(\mathcal{A}_{Cent}^{*})$, $(\mathcal{A}_{Mmt}^{*})$ *and* $(\mathcal{A}_{Cont})$ *hold, then the tests* $\phi_{n,h,\alpha}^{*,+}$, $\phi_{n,h,\alpha}^{*,-}$ *and* $\phi_{n,h,\alpha}^{*,+/-}$ *satisfy* $\left(\mathcal{P}_{size,\alpha}^{\infty}\right)$ *and respectively* $\left(\mathcal{P}_{consist.,\mathcal{P}_1^+}^{\infty}\right)$, $\left(\mathcal{P}_{consist.,\mathcal{P}_1^-}^{\infty}\right)$, $\left(\mathcal{P}_{consist.,\mathcal{P}_1}^{\infty}\right)$. *The same results hold for* $\phi_{n,h,\alpha}^{*MC,+}$, $\phi_{n,h,\alpha}^{*MC,-}$ *and* $\phi_{n,h,\alpha}^{*MC,+/-}$.

Notice that when φ is equal to φ^w defined by (4.15) with a continuous integrable function w, Theorem 10 means that the corresponding two-tailed tests are consistent against any alternative such that $\beta_w = \int w(u,v) \left(\mathbb{E}\left[dN_{X^1}(u) dN_{X^2}(v)\right] - \mathbb{E}\left[dN_{X^1}(u)\right] \mathbb{E}\left[dN_{X^2}(v)\right]\right) \neq 0$. In [STM15], for a particular function w, a nonasymptotic oracle-type result states that under specific Poisson assumptions, if β_w is larger than an explicit lower bound, then the second kind error rate of the proposed upper-tailed test of interaction is less than a prescribed level β in $(0,1)$. In some sense, Theorem 10 thus generalizes the result of [STM15] to a setup with much less reductive assumptions on the underlying stochastic models, but in an asymptotic way, and without any evaluation of the separation rate.

4.3.3 Sketch of proof

We give below a short sketch of proof of Theorem 9, which follows similar arguments to the ones of [BF81] for the bootstrap of the mean, or to [DM94] and [LN09] for the bootstrap of U-statistics. The main novel point here consists in using the topologies induced by the metrics $d_{\mathcal{D}}$, $d_{\mathcal{X}}$ and d defined by (4.12), (4.13), (4.14) and the properties of the separable Skorokhod metric space $(\mathcal{X}, d_{\mathcal{X}})$, where weak convergence of sample probability distributions is available (see [Var58]).

Recall that $(\Omega, \mathcal{A}, \mathbb{P})$ is the probability space on which all the X_i's are defined, so Ω represents the randomness in the original sequence $(X_i)_i$ and a given ω in Ω represents a given realization of $(X_i)_i$.

[First step] The first step of the proof consists in constructing, for almost all ω in Ω, a sequence of random variables $(\bar{Y}_{n,\omega,a}^{*})_{n \geq 1}$ such that for every $n \geq 1$, $\bar{Y}_{n,\omega,a}^{*} \sim P_{n,\omega}^{1} \otimes P_{n,\omega}^{2}$, where $P_{n,\omega}^{j} = n^{-1} \sum_{i=1}^{n} \delta_{X_i^j(\omega)}$ is the jth marginal empirical measure corresponding to the realization $\mathbf{X}_n(\omega)$, a random variable $\bar{Y}_{\omega,a} \sim P^1 \otimes P^2$, and $\{(\bar{Y}_{n,\omega,b}^{*})_{n \geq 1}, \bar{Y}_{\omega,b}\}$ an independent copy of $\{(\bar{Y}_{n,\omega,a}^{*})_{n \geq 1}, \bar{Y}_{\omega,a}\}$, both defined on some probability space $(\Omega'_\omega, \mathcal{A}'_\omega, \mathbb{P}'_\omega)$ depending on ω such that

$$\mathbb{E}'_\omega \left[\left(h\left(\bar{Y}_{n,\omega,a}^{*}, \bar{Y}_{n,\omega,b}^{*}\right) - h\left(\bar{Y}_{\omega,a}, \bar{Y}_{\omega,b}\right) \right)^2 \right] \underset{n \to +\infty}{\longrightarrow} 0, \tag{4.18}$$

where $\mathbb{E}'_\omega$ denotes the expectation corresponding to $\mathbb{P}'_\omega$.

From the strong law of large numbers for U-statistics due to Hoeffding [Hoe61], we deduce that there exists $\Omega_1 \subset \Omega$ such that $\mathbb{P}(\Omega_1) = 1$ and for every ω in Ω_1,

$$\frac{1}{n^4} \sum_{i,j,k,l=1}^{n} h^2\left(\left(X_i^1(\omega), X_j^2(\omega)\right), \left(X_k^1(\omega), X_l^2(\omega)\right)\right) \underset{n \to +\infty}{\longrightarrow} \mathbb{E}\left[h^2\left(\left(X_1^1, X_2^2\right), \left(X_3^1, X_4^2\right)\right)\right]. \tag{4.19}$$

Since $(\mathcal{X}, d_{\mathcal{X}})$ defined by (4.13) is separable, Theorem 3 in [Var58] can be applied, so there exists $\Omega_2 \subset \Omega$ such that $\mathbb{P}(\Omega_2) = 1$ and for every ω in Ω_2,

$$P_{n,\omega}^{1} \otimes P_{n,\omega}^{2} \underset{n \to +\infty}{\Longrightarrow} P^1 \otimes P^2.$$

Fix ω in $\Omega_0 = \Omega_1 \cap \Omega_2$. Following the proof of Skorokhod's representation theorem in [Dud02, Theorem 11.7.2, p. 415], since $(\mathcal{X}^2, d_{\mathcal{X}^2})$ is also a separable space, it is possible to construct

— some probability space $(\Omega'_\omega, \mathcal{A}'_\omega, \mathbb{P}'_\omega)$,

— some random variables $\bar{Y}^*_{n,\omega,a} : \Omega'_\omega \to \mathcal{X}^2$, $\bar{Y}^*_{n,\omega,b} : \Omega'_\omega \to \mathcal{X}^2$ with distribution $P^1_{n,\omega} \otimes P^2_{n,\omega}$,

— $\bar{Y}_{\omega,a} : \Omega'_\omega \to \mathcal{X}^2$, $\bar{Y}_{\omega,b} : \Omega'_\omega \to \mathcal{X}^2$ with distribution $P^1 \otimes P^2$,

such that $\{(\bar{Y}^*_{n,\omega,a})_{n\geq 1}, \bar{Y}_{\omega,a}\}$ and $\{(\bar{Y}^*_{n,\omega,b})_{n\geq 1}, \bar{Y}_{\omega,b}\}$ are independent, and w.r.t. the metric d, under $(\mathcal{A}_{Cont})$,

$$\mathbb{P}'_\omega\text{-a.s.},\ h\left(\bar{Y}^*_{n,\omega,a}, \bar{Y}^*_{n,\omega,b}\right) \to_{n\to+\infty} h\left(\bar{Y}_{\omega,a}, \bar{Y}_{\omega,b}\right).$$

As $\mathbb{P}'_\omega$-a.s. convergence implies convergence in probability, to obtain (4.18), we only need to prove that the sequence $(h^2(\bar{Y}^*_{n,\omega,a}, \bar{Y}^*_{n,\omega,b}))_{n\geq 1}$ is uniformly integrable. This is done just noting that (4.19) is equivalent to

$$\begin{aligned}\mathbb{E}'_\omega\left[h^2\left(\bar{Y}^*_{n,\omega,a}, \bar{Y}^*_{n,\omega,b}\right)\right] &= \frac{1}{n^4}\sum_{i,j,k,l=1}^{n} h^2\left(\left(X^1_i(\omega), X^2_j(\omega)\right), \left(X^1_k(\omega), X^2_l(\omega)\right)\right) \\ &\underset{n\to+\infty}{\longrightarrow} \mathbb{E}\left[h^2\left(\left(X^1_1, X^2_2\right), \left(X^1_3, X^2_4\right)\right)\right] = \mathbb{E}'_\omega\left[h^2\left(\bar{Y}_{\omega,a}, \bar{Y}_{\omega,b}\right)\right].\end{aligned}$$

(4.18) is thus obtained for any ω in Ω_0, with $\mathbb{P}(\Omega_0) = 1$.

[Second step] We prove that for all $n \geq 2$,

$$\begin{aligned}&d_2\Big(\mathcal{L}\left(\sqrt{n}U_{n,h}, P^1_n \otimes P^2_n | \mathbf{X}_n\right), \mathcal{L}\left(\sqrt{n}U_{n,h}, P^1 \otimes P^2\right)\Big) \\ &\qquad \leq \kappa \inf_{\substack{\left(Y^*_{n,a}, Y_a\right), \left(Y^*_{n,b}, Y_b\right)\ i.i.d., \\ Y^*_{n,a}, Y^*_{n,b} \sim P^1_n \otimes P^2_n,\ Y_a, Y_b \sim P^1 \otimes P^2}} \mathbb{E}^*\left[\left(h\left(Y^*_{n,a}, Y^*_{n,b}\right) - h\left(Y_a, Y_b\right)\right)^2\right].\end{aligned}$$

Since

$$\begin{aligned}&\inf_{\substack{\left(Y^*_{n,a}, Y_a\right), \left(Y^*_{n,b}, Y_b\right)\ i.i.d., \\ Y^*_{n,a}, Y^*_{n,b} \sim P^1_{n,\omega} \otimes P^2_{n,\omega},\ Y_a, Y_b \sim P^1 \otimes P^2}} \mathbb{E}^*\left[\left(h\left(Y^*_{n,a}, Y^*_{n,b}\right) - h\left(Y_a, Y_b\right)\right)^2\right](\omega) \\ &\qquad \leq \quad \mathbb{E}'_\omega\left[\left(h\left(\bar{Y}^*_{n,\omega,a}, \bar{Y}^*_{n,\omega,b}\right) - h\left(\bar{Y}_{\omega,a}, \bar{Y}_{\omega,b}\right)\right)^2\right],\end{aligned}$$

(4.18) allows to conclude.

4.4 Permutation tests of independence

The permutation approach we consider consists in randomly permuting the second coordinates of the observed pairs of point processes. More precisely, given a random permutation Π_n, uniformly distributed on the set $\mathfrak{S}_n$ of permutations of the set $\{1, \ldots, n\}$, and independent of $\mathbf{X}_n$, a permuted sample from $\mathbf{X}_n$ is defined by $\mathbf{X}^{\Pi_n}_n = (X^{\Pi_n}_1, \ldots, X^{\Pi_n}_n)$ with $X^{\Pi_n}_i = (X^1_i, X^2_{\Pi_n(i)})$.

Let $P^\star_n$ be the conditional distribution of $\mathbf{X}^{\Pi_n}_n$ given $\mathbf{X}_n$. Like for the bootstrap, the idea of the present permutation principle is to mimic the distribution of the test statistic under (H_0), so that permutation tests, defined through permutation-based critical values, can be introduced. More precisely, if for all kernel h, all $n \geq 2$, $q^{\star(\mathbf{X}_n)}_{n,h}$ denotes the quantile function of $\mathcal{L}\left(\sqrt{n}U_{n,h}, P^\star_n | \mathbf{X}_n\right)$, we introduce the tests $\phi^{\star,+}_{n,h,\alpha}$, $\phi^{\star,-}_{n,h,\alpha}$ and $\phi^{\star,+/-}_{n,h,\alpha}$ defined by (4.10) with

$$\left\{\begin{array}{rcl} c^+_{n,h,\alpha}(\mathbf{X}_n) &=& q^{\star(\mathbf{X}_n)}_{n,h}(1-\alpha), \\ c^-_{n,h,\alpha}(\mathbf{X}_n) &=& q^{\star(\mathbf{X}_n)}_{n,h}(\alpha). \end{array}\right. \tag{4.20}$$

Like for the bootstrap approach, even if an exact computation of the quantiles $q_{n,h}^{\star(\mathbf{X}_n)}(u)$ is possible by sorting the $n!$ values of $\{U_{n,h}(\mathbf{X}_n^{\pi_n})\}_{\pi_n\in\mathfrak{S}_n}$, for algorithmic reasons, a Monte Carlo approximation is used in practice. So, let $(B_n)_{n\geq 2}$ be a sequence of possible numbers of Monte Carlo iterations, such that $B_n \to_{n\to+\infty} +\infty$. For $n \geq 1$, let $\left(\Pi_n^1, \ldots, \Pi_n^{B_n}\right)$ be a sample of B_n i.i.d. random permutations uniformly distributed on $\mathfrak{S}_n$. The order statistic associated with $\left(U_{n,h}\left(\mathbf{X}_n^{\Pi_n^1}\right), \ldots, U_{n,h}\left(\mathbf{X}_n^{\Pi_n^{B_n}}\right), U_{n,h}(\mathbf{X}_n)\right)$ is denoted by $\left(U^{\star(1)}, \ldots, U^{\star(B_n+1)}\right)$. Our Monte Carlo permutation tests $\phi_{n,h,\alpha}^{\star MC,+}$, $\phi_{n,h,\alpha}^{\star MC,-}$ and $\phi_{n,h,\alpha}^{\star MC,+/-}$ are then defined by (4.10), with

$$\left\{\begin{array}{rcl} c_{n,h,\alpha}^{+}(\mathbf{X}_n) & = & \sqrt{n}U^{\star(\lceil(1-\alpha)(B_n+1)\rceil)}, \\ c_{n,h,\alpha}^{-}(\mathbf{X}_n) & = & \sqrt{n}U^{\star(\lfloor\alpha(B_n+1)\rfloor+1)}. \end{array}\right. \tag{4.21}$$

4.4.1 Asymptotic properties in the *Linear case*

We prove in [12] that the conditional distribution $\mathcal{L}\left(\sqrt{n}U_{n,h}, P_n^\star|\mathbf{X}_n\right)$ is asymptotically close to the distribution $\mathcal{L}\left(\sqrt{n}U_{n,h}, P^1\otimes P^2\right)$. Following the statements of our bootstrap results in Section 4.3, we still express (see (4.23)) the closeness in distributions between the permuted and original statistics in terms of Wasserstein's metric, and this even under (H_1). This result is therefore one of the newest results of our work, whose scope is thus beyond the only generalization to the point processes setting. In particular, as it allows to understand the behavior, under (H_0) as well as under (H_1), of the permuted test statistic, it can be viewed as a step toward a solution for the open question of [VdVW96, p. 371], as explained in the introduction. However, the result is only obtained in the *Linear case*, where h is of the form h_φ for some integrable function φ, as defined in (4.6), under $(\mathcal{A}_{nondeg})$ and the following moment assumption.

$(\mathcal{A}_{\varphi,Mmt})$ | For $X = \left(X^1, X^2\right)$ with distribution P or $P^1\otimes P^2$, $\mathbb{E}\left[\varphi^4\left(X^1, X^2\right)\right] < \infty$.

Recall that in the *Linear case*, $(\mathcal{A}_{Cent})$ is always satisfied.

Theorem 11 (Albert, Bouret, Fromont, Reynaud-Bouret, 2015)**.** *In the* Linear case*, under* $(\mathcal{A}_{nondeg})$ *and* $(\mathcal{A}_{\varphi,Mmt})$*,*

$$d_2\left(\mathcal{L}\left(\sqrt{n}U_{n,h_\varphi}, P_n^\star|\mathbf{X}_n\right), \mathcal{N}\left(0, \sigma^2_{h_\varphi,P^1\otimes P^2}\right)\right) \xrightarrow[n\to+\infty]{\mathbb{P}} 0, \tag{4.22}$$

where $\xrightarrow{\mathbb{P}}$ *stands for the usual convergence in* $\mathbb{P}$*-probability.*

We then deduce from (4.8) that, in the conditions of Theorem 11,

$$d_2\left(\mathcal{L}\left(\sqrt{n}U_{n,h_\varphi}, P_n^\star|\mathbf{X}_n\right), \mathcal{L}\left(\sqrt{n}U_{n,h_\varphi}, P^1\otimes P^2\right)\right) \xrightarrow[n\to+\infty]{\mathbb{P}} 0, \tag{4.23}$$

and for every u in $(0,1)$,

$$q_{n,h_\varphi}^{\star(\mathbf{X}_n)}(u) \xrightarrow[n\to+\infty]{\mathbb{P}} F^{-1}_{0,\sigma^2_{h_\varphi,P^1\otimes P^2}}(u).$$

These results nearly have the same role as Theorem 9 and Corollary 3, as one can see in Theorem 12. However, note that unlike the bootstrap approach, the conditional distribution of the permuted test statistic is not here directly compared to the initial distribution of the test statistic under (H_0), but to its Gaussian limit distribution. As a consequence, the consistency result can only hold under the nondegeneracy assumption $(\mathcal{A}_{nondeg})$. Moreover, the convergence occurs here in probability and not almost surely, but note that no continuity assumption for the kernel h_φ is used anymore.
Finally, let us remark that the moment assumption, which is here stronger than the one used for the bootstrap could perhaps be replaced by slighter Lindeberg conditions as in the combinatorial central limit theorems for permutation (see [Háj61, Mot56, HC78]).

Theorem 12 (Albert, Bouret, Fromont, Reynaud-Bouret, 2015). *In the Linear case, under* $(\mathcal{A}_{nondeg})$ *and* $(\mathcal{A}_{\varphi,Mmt})$, *the tests* $\phi^{\star,+}_{n,h_\varphi,\alpha}$, $\phi^{\star,-}_{n,h_\varphi,\alpha}$ *and* $\phi^{\star,+/-}_{n,h_\varphi,\alpha}$ *satisfy* $\left(\mathcal{P}^{\infty}_{size,\alpha}\right)$ *and respectively* $\left(\mathcal{P}^{\infty}_{consist.,\mathcal{P}_1^+}\right)$, $\left(\mathcal{P}^{\infty}_{consist.,\mathcal{P}_1^-}\right)$, $\left(\mathcal{P}^{\infty}_{consist.,\mathcal{P}_1}\right)$. *The same results hold for the tests* $\phi^{\star MC,+}_{n,h_\varphi,\alpha}$, $\phi^{\star MC,-}_{n,h_\varphi,\alpha}$ *and* $\phi^{\star MC,+/-}_{n,h_\varphi,\alpha}$.

At this stage, since the permutation independence tests satisfy the same asymptotic properties as the bootstrap ones, but with much more computation difficulties to prove them, one might wonder whether the introduction of the permutation tests is of genuine interest.
The main known advantage of the permutation approach lies in its nonasymptotic properties, which are satisfied, whatever the symmetric kernel h, so not only in the *Linear case*.

4.4.2 General nonasymptotic properties

As far as we know, the nonasymptotic properties of the original permutation tests are known from the fundamental paper by Hoeffding [Hoe52]. They are based on an *invariance property*, which can be expressed in the present context as follows.

For every (deterministic) element π_n of the group of permutations $\mathfrak{S}_n$, if $P = P^1 \otimes P^2$, that is under (H_0), then $\mathbf{X}_n^{\pi_n}$ has the same distribution as the original sample $\mathbf{X}_n$.

This property justifies that the permutation tests $\phi^{\star,+}_{n,h,\alpha}$, $\phi^{\star,-}_{n,h,\alpha}$ and $\phi^{\star,+/-}_{n,h,\alpha}$ are of level α.
Indeed, focusing for instance on the upper-tailed test $\phi^{\star,+}_{n,h,\alpha}$, if $P = P^1 \otimes P^2$, following Hoeffding's arguments, from the invariance property, one has

$$\begin{aligned}
\mathbb{P}\left(\phi^{+}_{n,h,\alpha} = 1\right) &= \mathbb{P}\left(\sqrt{n}U_{n,h}\left(\mathbf{X}_n\right) > q^{\star(\mathbf{X}_n)}_{n,h}(1-\alpha)\right) \\
&= \frac{1}{n!}\sum_{\pi_n\in\mathfrak{S}_n}\mathbb{P}\left(\sqrt{n}U_{n,h}\left(\mathbf{X}_n^{\pi_n}\right) > q^{\star(\mathbf{X}_n^{\pi_n})}_{n,h}\left(1-\alpha\right)\right) \\
&= \frac{1}{n!}\sum_{\pi_n\in\mathfrak{S}_n}\mathbb{P}\left(\sqrt{n}U_{n,h}\left(\mathbf{X}_n^{\Pi_n}\right) > q^{\star(\mathbf{X}_n^{\Pi_n})}_{n,h}\left(1-\alpha\right)\middle|\Pi_n = \pi_n\right) \\
&= \mathbb{P}\left(\sqrt{n}U_{n,h}\left(\mathbf{X}_n^{\Pi_n}\right) > q^{\star(\mathbf{X}_n^{\Pi_n})}_{n,h}\left(1-\alpha\right)\right) \\
&= \mathbb{E}\left[\mathbb{P}\left(\sqrt{n}U_{n,h}\left(\mathbf{X}_n^{\Pi_n}\right) > q^{\star(\mathbf{X}_n^{\Pi_n})}_{n,h}\left(1-\alpha\right)\middle|\mathbf{X}_n\right)\right] \\
&= \mathbb{E}\left[\frac{1}{n!}\sum_{\pi_n\in\mathfrak{S}_n}\mathbb{1}_{\sqrt{n}U_{n,h}(\mathbf{X}_n^{\pi_n})>q^{\star(\mathbf{X}_n^{\pi_n})}_{n,h}(1-\alpha)}\right].
\end{aligned}$$

But for all x in $(\mathcal{X}^2)^n$, $q^{\star(x^{\pi_n})}_{n,h}(1-\alpha)$ is the $\lceil(1-\alpha)n!\rceil$th ordered value of $\{\sqrt{n}U_{n,h}\left(x^{\pi_n}\right),\ \pi_n\in\mathfrak{S}_n\}$, therefore,

$$\sum_{\pi_n\in\mathfrak{S}_n}\mathbb{1}_{\sqrt{n}U_{n,h}(x^{\pi_n})>q^{\star(x^{\pi_n})}_{n,h}(1-\alpha)} \leq n!\alpha,$$

which allows to conclude that the test $\phi^{+}_{n,h,\alpha}$ is of level α.
This historical proof, which can be found in [Hoe52], can of course be reproduced for many testing problems, and other groups of transformations than $\mathfrak{S}_n$. For instance, it can be used in the testing problems of [DR06, ABR10] where the random transformations are defined from Rademacher variables.

Now, from a more recent result of Romano and Wolf, namely [RW05, Lemma 1] recalled in Lemma 4, one can furthermore prove that for every $n \geq 2$, the tests $\phi^{\star MC,+}_{n,h,\alpha}$, $\phi^{\star MC,-}_{n,h,\alpha}$ and $\phi^{\star MC,+/-}_{n,h,\alpha}$ are also of level α, and we can state the following result.

Proposition 6 (Albert, Bouret, Fromont, Reynaud-Bouret, 2015)**.** *For every fixed sample size n, any test $\phi_{n,h,\alpha}$ equal to $\phi^{\star,+}_{n,h,\alpha}$, $\phi^{\star,-}_{n,h,\alpha}$, $\phi^{\star,+/-}_{n,h,\alpha}$, $\phi^{\star MC,+}_{n,h,\alpha}$, $\phi^{\star MC,-}_{n,h,\alpha}$, or $\phi^{\star MC,+/-}_{n,h,\alpha}$ satisfies*

$$(\mathcal{P}_{level,\alpha}) \qquad \Big| \ \mathbb{P}\left(\phi_{n,h,\alpha}=1\right) \leq \alpha \text{ if } P = P^1 \otimes P^2.$$

Note that these nonasymptotic results have no counterpart for the bootstrap approaches in general, except when a particular exact wild bootstrap approach, based on Rademacher variables, is used (see [DR06], [ABR10], [10] and Chapter 3). In a regression framework, [ABR10] further gives a nonasymptotic control of the first kind error rate for other kinds of bootstrap tests, based on concentration inequalities in the spirit of [6], but such concentration inequalities are not accessible in any model and for any test statistic.

4.4.3 Sketch of proof

We give below a sketch of proof of Theorem 11.

Let d_{BL} denote the bounded Lipschitz metric for the weak convergence (see [Dud02]). For any variable Z_n depending on $\mathbf{X}_n$ and Π_n, $\mathcal{L}\left(Z_n|\mathbf{X}_n\right)$ denotes the conditional distribution of Z_n given $\mathbf{X}_n$.
[First step] The first step of the proof consists in decomposing $\sqrt{n}U_{n,h_\varphi}\left(\mathbf{X}_n^{\Pi_n}\right)$ as

$$\sqrt{n}U_{n,h_\varphi}\left(\mathbf{X}_n^{\Pi_n}\right) = \frac{n}{n-1}\left(M_n^{\Pi_n}\left(\mathbf{X}_n\right) + R_n^{\Pi_n}\left(\mathbf{X}_n\right) - T_n\left(\mathbf{X}_n\right)\right),$$

where $M_n^{\Pi_n}\left(\mathbf{X}_n\right) = n^{-1/2}\sum_{i\neq j}\mathbb{1}_{\Pi_n(i)=j}C_{i,j}$, $R_n^{\Pi_n}\left(\mathbf{X}_n\right) = n^{-1/2}\sum_{i=1}^n\left(\mathbb{1}_{\Pi_n(i)=i}-1/n\right)C_{i,i}$, $T_n\left(\mathbf{X}_n\right) = n^{-3/2}\sum_{i\neq j}C_{i,j}$, with $C_{i,j} = \varphi(X_i^1, X_j^2) - \mathbb{E}\left[\varphi(X_i^1, X^2)\middle|X_i^1\right] - \mathbb{E}\left[\varphi(X^1, X_j^2)\middle|X_j^2\right] + \mathbb{E}\left[\varphi(X_1^1, X_2^2)\right]$, $X = (X^1, X^2)$ being P-distributed and independent of $(X_i)_i$.
From Markov's inequality, we prove that

$$\mathbb{E}\left[\left|R_n^{\Pi_n}\left(\mathbf{X}_n\right)\right|\middle|\mathbf{X}_n\right] \underset{n\to+\infty}{\overset{\mathbb{P}}{\longrightarrow}} 0 \text{ and } T_n\left(\mathbf{X}_n\right) \underset{n\to+\infty}{\overset{\mathbb{P}}{\longrightarrow}} 0,$$

which leads to

$$d_{BL}\left(\mathcal{L}\left(\sqrt{n}U_{n,h_\varphi}\left(\mathbf{X}_n^{\Pi_n}\right)\middle|\mathbf{X}_n\right), \mathcal{L}\left(\frac{n}{n-1}M_n^{\Pi_n}\left(\mathbf{X}_n\right)\middle|\mathbf{X}_n\right)\right) \underset{n\to+\infty}{\overset{\mathbb{P}}{\longrightarrow}} 0.$$

[Second step] The second, and most difficult, step of the proof consists in proving that

$$d_{BL}\left(\mathcal{L}\left(M_n^{\Pi_n}\left(\mathbf{X}_n\right)\middle|\mathbf{X}_n\right), \mathcal{N}\left(0, \sigma^2_{h_\varphi, P^1\otimes P^2}\right)\right) \underset{n\to+\infty}{\overset{\mathbb{P}}{\longrightarrow}} 0. \tag{4.24}$$

First note that $M_n^{\Pi_n}\left(\mathbf{X}_n\right) = \sum_{i=1}^n Y_{n,i}$, with $Y_{n,i} = n^{-1/2}\sum_{j=1}^{i-1}\left(\mathbb{1}_{\Pi_n(i)=j}C_{i,j} + \mathbb{1}_{\Pi_n(j)=i}C_{j,i}\right)$.
Now, let Π_n' be a random permutation uniformly distributed on $\mathfrak{S}_n$, independent of Π_n and $\mathbf{X}_n$, and define $Y'_{n,i}$ and $M_n^{\Pi_n'}\left(\mathbf{X}_n\right)$ by replacing Π_n by Π_n' in the definitions of $Y_{n,i}$ and $M_n^{\Pi_n}\left(\mathbf{X}_n\right)$. Fix a, b in $\mathbb{R}$. Setting $\mathcal{F}_{n,i} = \sigma\left(\Pi_n, \Pi_n', X_1, X_2, \ldots, X_i\right)$ for $2 \leq i \leq n$, we prove by technical computations that $(aY_{n,i} + bY'_{n,i}, \mathcal{F}_{n,i})_{2\leq i\leq n}$ is a martingale difference array. From a central limit theorem for such martingale difference arrays, we obtain that

$$\mathcal{L}\left(aM_n^{\Pi_n}\left(\mathbf{X}_n\right) + bM_n^{\Pi_n'}\left(\mathbf{X}_n\right)\right) \underset{n\to+\infty}{\Longrightarrow} \mathcal{N}\left(0, \left(a^2+b^2\right)\sigma^2_{h_\varphi, P^1\otimes P^2}\right),$$

which, according to the Cramér-Wold device, gives that for every t in $\mathbb{R}$,

$$\begin{cases} \mathbb{P}\left(M_n^{\Pi_n}\left(\mathbf{X}_n\right) \leq t\right) \underset{n\to+\infty}{\longrightarrow} F_{0,\sigma^2_{h_\varphi, P^1\otimes P^2}}(t), \\ \mathbb{P}\left(M_n^{\Pi_n}\left(\mathbf{X}_n\right) \leq t, M_n^{\Pi_n'}\left(\mathbf{X}_n\right) \leq t\right) \underset{n\to+\infty}{\longrightarrow} F^2_{0,\sigma^2_{h_\varphi, P^1\otimes P^2}}(t). \end{cases}$$

Using Chebychev's inequality, with the fact (see [Dud02, Theorem 9.2.1] for instance) that in a separable metric space, convergence in probability is metrizable, and therefore is equivalent to almost sure convergence of a subsequence of any initial subsequence, we prove that this leads to (4.24). Hence

$$d_{BL}\left(\mathcal{L}\left(\sqrt{n}U_{n,h_\varphi}\left(\mathbf{X}_n^{\Pi_n}\right)\middle|\mathbf{X}_n\right),\mathcal{N}\left(0,\sigma^2_{h_\varphi,P^1\otimes P^2}\right)\right)\xrightarrow[n\to+\infty]{\mathbb{P}}0.$$

[Third step] We finally derive via direct computations and the strong law of large numbers for U-statistics of Hoeffding [Hoe61], the convergence of the conditional second order moments:

$$\mathbb{E}\left[\left(\sqrt{n}U_{n,h_\varphi}\left(\mathbf{X}_n^{\Pi_n}\right)\right)^2\middle|\mathbf{X}_n\right]\xrightarrow[n\to+\infty]{a.s.}\sigma^2_{h_\varphi,P^1\otimes P^2},$$

which ends the proof.

4.5 Experimental results

The above bootstrap and permutation tests have been applied on simulated data as well as real data from neuroscience experiments. We only tackle in this short section the simulation study, as the real application is developed in the first part of Chapter 5 devoted to multiple tests. Indeed, in this application, the tests are used on several time windows, that cover the whole observation time interval, simultaneously, and therefore integrated in a multiple testing procedure.

On the one hand, samples from several distributions $P = P^1 \otimes P^2$ have been simulated in order to estimate the size, that is, the first kind error rate, of the tests. On the other hand, alternatives were chosen, so that they are either pairs of Poisson processes, or more realistic point processes in the neuroscience context (for instance Hawkes processes) with particular dependence structures, in order to estimate the power of the tests.
All the obtained results confirm that the permutation test should be preferred, as, in all the experiments, it shows an exact control of the size, and, compared to other tests that also guarantee such a control (which is not the case for the bootstrap test), it is also the most powerful one.
The test based on the trial-shuffling approach from [GDA10], which is the reference distribution-free method for neuroscientists, is clearly too conservative. We explain and illustrate in details in [15] that this conservative behavior is due to a centering defect: the bootstrap approach is applied on a statistic close to $C(\mathbf{X}_n)$ (see (4.2)), which is not properly centered. The present bootstrap tests were initially devoted to correcting this defect.
When we introduced the permutation tests however, we also observed that a test based on the permutation approach with the statistic $C(\mathbf{X}_n)$ does not suffer from the same defect. In fact, as our centering correction term is invariant under permutation, this test is equivalent to ours.

4.6 Perspectives

An immediate perspective would be to study an aggregated test based on the above single independence tests, typically defined from a collection of tests based on $h_{\varphi_\delta^{coinc}}$, for several values of δ. However, in the present context, this issue has only theoretical interest. Neuroscientists are in fact especially interested in the value of δ which leads to a rejection of the independence hypothesis, as it provides the delay of interaction between neurons.
The nonasymptotic study of the permutation based independence tests constitutes a large part of the PhD thesis of Mélisande Albert, where appropriate concentration tools on permutation are developed. This work is still in progress and in particular, new concentration results are expected to handle very general forms of independence test statistics, for instance based on kernels in the spirit of [GG10].
Another path would be to develop a new exact bootstrap approach, different from the permutation one, whose nonasymptotic theoretical study would be facilitated, at least for some forms of U-statistics.

Chapter 5

Multiple tests

5.1 Introduction

Let $\mathbf{X}$ be an observed random variable, defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Given a possible set $\mathcal{P}$ of distributions P for $\mathbf{X}$, we recall that a hypothesis is defined through a subset of $\mathcal{P}$. In a classical single testing problem, two hypotheses are considered: the null hypothesis (H_0), which is viewed as the favorite one, and expressed from a subset $\mathcal{P}_0$ as

$$(H_0) \quad P \in \mathcal{P}_0,$$

and an alternative (H_1), expressed from a subset $\mathcal{P}_1 \subset \mathcal{P} \setminus \mathcal{P}_0$ as

$$(H_1) \quad P \in \mathcal{P}_1.$$

In a multiple testing problem, a whole collection of null hypotheses is considered. For sake of simplicity, following the terminology of Goeman and Solari [GS10], these hypotheses are confused with their associated subset of $\mathcal{P}$. An hypothesis H is therefore defined as a subset of $\mathcal{P}$. It is said to be true under P if P belongs to H, and false under P otherwise.
Given a finite collection $\mathcal{H}$ of such hypotheses, the aim is simultaneously testing H against $\mathcal{P} \setminus H$, for every H in $\mathcal{H}$, that is, simultaneously testing "H is true under P" against "H is false under P", or equivalently "$P \in H$" against "$P \notin H$", for every H in $\mathcal{H}$.
We introduce the set of true hypotheses under P, given by

$$\mathcal{T}(P) = \{H \in \mathcal{H},\ P \in H\},$$

and the set of false hypotheses under P, given by

$$\mathcal{F}(P) = \mathcal{H} \setminus \mathcal{T}(P) = \{H \in \mathcal{H},\ P \notin H\}.$$

A multiple testing procedure or a multiple test is a statistic given by a collection of rejected hypotheses $\mathcal{R} \subset \mathcal{H}$, only depending on the observed random variable $\mathbf{X}$, whose goal is to infer the set $\mathcal{F}(P)$.
It is usually constructed from single tests of "$P \in H$" against "$P \notin H$", for all H in $\mathcal{H}$, that are mostly defined through p-values p_H, for H in $\mathcal{H}$. For instance, given a collection of p-values $\{p_H,\ H \in \mathcal{H}\}$ such that for all P in H,

$$\forall u \in (0,1), \quad P(p_H \leq u) \leq u,$$

given a prescribed error rate level α in $(0,1)$, the historical Bonferroni multiple test (see e.g., [Sim86]) is defined by $\{H \in \mathcal{H},\ p_H \leq \alpha/\#\mathcal{H}\}$.
As seen in Lemma 1, to go back and forth the expression of the considered single tests via test statistics and critical values and their expression via p-values, it is however more convenient to consider the following version of the Bonferroni multiple test:

$$\mathcal{R}_{Bonf} = \{H \in \mathcal{H},\ p_H < \alpha/\#\mathcal{H}\}, \tag{5.1}$$

which is generalized to the weighted Bonferroni multiple test:

$$\mathcal{R}_{wBonf} = \{ H \in \mathcal{H},\ p_H < w_H \alpha \}, \tag{5.2}$$

$\{ w_H,\ H \in \mathcal{H} \}$ being a collection of positive weights such that $\sum_{H \in \mathcal{H}} w_H \leq 1$.

The Bonferroni multiple tests have been specifically constructed so that they have a Family-Wise Error Rate, and so a weak Family-Wise Error Rate, defined as follows, both controlled by the prescribed level α.

Definition 7 ((Weak) Family-Wise Error Rate)**.** The weak Family-Wise Error Rate of a multiple test $\mathcal{R}$ is defined by:

$$w\mathrm{FWER}(\mathcal{R}) = \sup_{P,\ \mathcal{T}(P)=\mathcal{H}} P(\mathcal{R} \cap \mathcal{T}(P) \neq \emptyset),$$

and the (strong) Family-Wise Error Rate of $\mathcal{R}$ by:

$$\mathrm{FWER}(\mathcal{R}) = \sup_{P \in \mathcal{P}} P(\mathcal{R} \cap \mathcal{T}(P) \neq \emptyset).$$

But controlling the above FWER may be too stringent and not needed in some applications. Many other first kind error-related criteria for multiple tests have thus been introduced in the statistical literature, generalizing or relaxing the FWER, defined above as the maximal probability of one or more false discoveries (true null hypotheses that are rejected). Among them, the Per-Family Error Rate (PFER) suggested by Spjøtvoll [Spj72] corresponds to the average number of false discoveries, while the $k-$FWER introduced by Hommel and Hoffman [HH88] and further studied by Korn et al. [KTMS04], Lehmann and Romano [LR05], Romano and Shaikh [RS06] or Romano and Wolf [RW07, RW10], is the probability of k or more false discoveries. Like Genovese and Wasserman [GW04], several of these authors also focused on the False Discovery Proportion (FDP), whose expected value is the very popular False Discovery Rate (FDR) introduced by Benjamini and Hochberg [BH95].

Definition 8 (False Discovery Rate)**.** The False Discovery Rate of a multiple test $\mathcal{R}$ is defined by:

$$\mathrm{FDR}_P(\mathcal{R}) = \mathbb{E}_P\left[\frac{\#(\mathcal{T}(P) \cap \mathcal{R})}{1 \vee \#\mathcal{R}}\right] = \begin{cases} \mathbb{E}_P\left[\frac{\#(\mathcal{T}(P)\cap\mathcal{R})}{\#\mathcal{R}}\right] & \text{if } \mathcal{R} \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

Notice that for every multiple test $\mathcal{R}$, $\sup_{P,\ \mathcal{T}(P)=\mathcal{H}} \mathrm{FDR}_P(\mathcal{R}) = w\mathrm{FWER}(\mathcal{R})$.
So, if $\sup_{P,\ \mathcal{T}(P)=\mathcal{H}} \mathrm{FDR}_P(\mathcal{R}) \leq \alpha$ then $\mathcal{R}$ has a wFWER controlled by α. Furthermore, a multiple test such that $\mathrm{FWER}(\mathcal{R}) \leq \alpha$ satisfies $\sup_{P \in \mathcal{P}} \mathrm{FDR}_P(\mathcal{R}) \leq \alpha$.
Benjamini and Hochberg introduce in [BH95] their famous multiple test defined as follows. Considering the ordered p-values $p^{(1)} \leq \ldots \leq p^{(\#\mathcal{H})}$ of $\{p_H,\ H \in \mathcal{H}\}$, and denoting the corresponding hypotheses by $H^{(1)}, \ldots, H^{(\#\mathcal{H})}$, Benjamini and Hochberg's procedure is given by

$$\mathcal{R}_{BH} = \left\{ H^{(1)}, \ldots, H^{(k)} \right\} \text{ with } k = \max\left\{ i,\ p^{(i)} \leq i\alpha/\#\mathcal{H} \right\}.$$

If the single test statistics or the p-values $\{p_H,\ H \in \mathcal{H}\}$ are independent, then

$$\sup_{P \in \mathcal{P}} \mathrm{FDR}_P(\mathcal{R}_{BH}) \leq \alpha.$$

Benjamin and Yekutieli [BY01] proved that the procedure still controls the FDR_P for every P in $\mathcal{P}$ such that the single test statistics or the p-values (corresponding to the true hypotheses) satisfy a positive dependency property, namely the PRDS property. They propose in other cases of dependency a modified procedure with $\alpha / \left(\sum_{i=1}^{\#\mathcal{H}} 1/i \right)$ instead of α so that the FDR_P is always controlled by α.

The present chapter is devoted to two very different topics in the multiple testing scene: an applied one about a neuroscience issue, and a theoretical one about the definition of new second kind error related criteria for multiple tests.

5.2 A distribution free Unitary Events method in neuroscience

Following the work presented in Chapter 4, and focusing on the neuroscience application which has motivated it, we introduce a new distribution free Unitary Events (UE) method, named Permutation UE, in a paper with Mélisande Albert, Yann Bouret, and Patricia Reynaud-Bouret under revision in a neuroscience journal [15].
Before precisely describing this method, which consists in a multiple testing procedure involving the permutation independence tests based on delayed coincidence count, as defined in Section 4.4, we briefly come back to the experimental protocol and the problem at hand.

5.2.1 Experimental design, data and testing problem

As explained in Chapter 4, the eventual time dependence either between cerebral areas or between neurons, and in particular the synchrony phenomenon, has been and is still vastly debated and investigated as a potential element of the neuronal code (see [Sin93] for instance). To detect this phenomenon at the microscopic level, multi-micro-electrodes are used to record the nearby electrical activity. After pretreatment, the time occurrences of action potentials (spikes) for several neurons are available. One of the first steps of analysis is then to assess whether two simultaneously recorded spike trains, corresponding to two different neurons, are dependent or not.
The data used here were partially published in previous experimental studies [RGDG00, GR03, RGM06] and also treated in [TMRGRB14]. These data were collected on a 5-year-old male Rhesus monkey who was trained to perform a delayed multidirectional pointing task. Concretely, the animal sat in a primate chair in front of a vertical panel on which touch-sensitive light-emitting diodes, the targets, were mounted. After the preparatory signal (PS) consisting in the illumination of one of the targets in green, a response signal (RS) illuminated the target that the monkey had to touch in red. Data recorded from several micro-electrodes were amplified and band-pass filtered, and using a window discriminator, spike trains from only one single neuron per electrode were then isolated (see [15] for instance for more details about the experimental design). Neuronal data (the spike trains) together with behavior data (such as the reaction time or the movement time of the monkey) were stored on a PC for off-line analysis with a time resolution of 10kHz.
In the present study, only trials where the response signal (RS) occurred at 1.7s, from the pair of neurons 13 are considered, as they were already treated in [TMRGRB14].

The trials were recorded on a time interval $[0, T]$ with $T = 2.2$s. From these data, we aim at detecting precise locations of dependence periods between the two neurons of the pair 13.
To this end, we consider a collection $\mathcal{W}$ of small time windows, that are potentially overlapping intervals covering the whole interval $[0, T]$. The problem at hand then becomes a multiple testing problem of the collection of hypotheses $\mathcal{H} = \{ H_W,\ W \in \mathcal{W} \}$, where each H_W can be expressed as "the two neurons of the pair 13 are independent on W".

On each window W in $\mathcal{W}$, the data restricted to W are assumed to be the observation of a sample $\mathbf{X} = \mathbf{X}_n = (X_1, \dots, X_n)$ whose distribution follows the model $\mathcal{M}^{(2)}_{\text{point proc.}}$ of Chapter 4, just replacing the interval of observation $[0, 1]$ by W, that is

$\mathcal{M}^{(2)}_{\text{point proc.}}$	$\mathbf{X} = \mathbf{X}_n = (X_1, \dots, X_n)$, where $\big(X_i = (X_i^1, X_i^2)\big)_{i \geq 1}$ is a sequence of i.i.d. pairs of finite point processes defined on $(\Omega, \mathcal{A}, \mathbb{P})$, observed on $\mathbb{X} = W$, with joint distribution P, with marginals P_1 and P_2.

Among the most popular methods used to detect dependence periods between two neurons, is the Unitary Events (UE) method of Grün and collaborators [Grü96, GDA10], which has been applied in the last decade on a large amount of real data (see for instance [KRPA$^+$09] and references therein).

From the initial method, substantial upgrades have been developed, like in particular the distribution free methods based on a bootstrap approach named trial-shuffling (see [PG03, PDG03]). We show in [15] that the test statistic used in these methods to detect dependence on a given window is not correctly centered, which makes the trial-shuffling approach inappropriate. The bootstrap and permutation tests proposed in [12], and presented in Chapter 4, both overcome this difficulty. As explained in Chapter 4, the permutation tests should nevertheless be preferred as they guarantee a strict control of the first kind error rate, and are still at least as powerful as the other reasonable tests. Hence, we use these permutation tests, and integrate them in a multiple testing procedure.

5.2.2 Single permutation independence tests

Let us here focus on a single window W in $\mathcal{W}$, and consider the problem of testing the null hypothesis H_W, from data modeled by $\mathcal{M}^{(2)}_{\text{point proc.}}$ above.
In [12], we propose new tests that suit the dependence feature that has to be detected in the present neuroscience issue, based on the test statistic

$$\sqrt{n}U_{n,h_{\varphi_\delta^{coinc}}}(\mathbf{X}_n) = \frac{\sqrt{n}}{n(n-1)} \sum_{i\neq i'\in\{1,\dots,n\}} \left(\varphi_\delta^{coinc}\left(X_i^1, X_i^2\right) - \varphi_\delta^{coinc}\left(X_i^1, X_{i'}^2\right)\right),$$

with $\varphi_\delta^{coinc}(X^1, X^2) = \int_{W^2} \mathbb{1}_{|u-v|\leq\delta} dN_{X^1}(u) dN_{X^2}(v)$ (see (4.4) and (4.1)). A detailed algorithm to compute $\varphi_\delta^{coinc}(X^1, X^2)$, with a study of its complexity, are provided in [15].
The corresponding critical values are constructed from a classical bootstrap or permutation approach, and approximated by a Monte Carlo method to lead to the tests $\phi^{*MC,+}_{n,h_{\varphi_\delta^{coinc}},\alpha}$, $\phi^{*MC,-}_{n,h_{\varphi_\delta^{coinc}},\alpha}$, $\phi^{*MC,+/-}_{n,h_{\varphi_\delta^{coinc}},\alpha}$, $\phi^{\star MC,+}_{n,h_{\varphi_\delta^{coinc}},\alpha}$, $\phi^{\star MC,-}_{n,h_{\varphi_\delta^{coinc}},\alpha}$ and $\phi^{\star MC,+/-}_{n,h_{\varphi_\delta^{coinc}},\alpha}$ defined by (4.10), (4.17) and (4.21).
Notice that when such resampling approaches are used, the normalization factor in the test statistic, that is $\sqrt{n}/\left(n(n-1)\right)$, can be removed provided that it is also removed in the bootstrapped and permuted test statistics used to compute the critical values.
Furthermore, it has to be underlined that $U_{n,h_{\varphi_\delta^{coinc}}}(\mathbf{X}_n)$ can also be written as

$$U_{n,h_{\varphi_\delta^{coinc}}}(\mathbf{X}_n) = \frac{1}{n-1}C(\mathbf{X}_n) - \frac{1}{n(n-1)}\sum_{i,i'=1}^{n} \varphi_\delta^{coinc}(X_i^1, X_{i'}^2),$$

where $C(\mathbf{X}_n)$ is defined as in (4.2) by $C(\mathbf{X}_n) = \sum_{i=1}^n \varphi_\delta^{coinc}\left(X_i^1, X_i^2\right)$.
As the term $\sum_{i,i'=1}^n \varphi_\delta^{coinc}(X_i^1, X_{i'}^2)$ is invariant by permutation, that is, for any permutation π_n in $\mathfrak{S}_n$, $\sum_{i,i'=1}^n \varphi_\delta^{coinc}\left(X_i^1, X_{\pi_n(i')}^2\right) = \sum_{i,i'=1}^n \varphi_\delta^{coinc}\left(X_i^1, X_{i'}^2\right)$, $C(\mathbf{X}_n)$ can be used as test statistic in the permutation tests, as well as $\sqrt{n}U_{n,h_{\varphi_\delta^{coinc}}}(\mathbf{X}_n)$.
Such a "rough" test statistic was already used in the trial-shuffling methods. However, in this case, as in our classical bootstrap approach, it is important to see that the term $\sum_{i,i'=1}^n \varphi_\delta^{coinc}(X_i^1, X_{i'}^2)$ is not resampling-invariant anymore, which makes the use of $C(\mathbf{X}_n)$ as test statistic irrelevant.

As explained above, we here choose to use the permutation tests, and since neuroscientists expect to assess whether coincidences occur either more or less than what may be due to chance, we in particular focus on the upper and lower-tailed tests $\phi^{\star MC,+}_{n,h_{\varphi_\delta^{coinc}},\alpha}$ and $\phi^{\star MC,-}_{n,h_{\varphi_\delta^{coinc}},\alpha}$.
For practical convenience, these two tests are in fact expressed through their associated p-values. So, let B be a given number of Monte Carlo simulations, and $(\Pi_n^1, \dots, \Pi_n^B)$ be a sample of B i.i.d. random permutations uniformly distributed on $\mathfrak{S}_n$. Set for every b in $\{1, \dots, n\}$, $C^{\star b} = C\left(\mathbf{X}_n^{\Pi_n^b}\right)$, where $\mathbf{X}_n^{\Pi_n^b} = \left(X_1^{\Pi_n^b}, \dots, X_n^{\Pi_n^b}\right)$, with $X_i^{\Pi_n^b} = \left(X_i^1, X_{\Pi_n^b(i)}^2\right)$. The p-values associated with the tests $\phi^{\star MC,+}_{n,h_{\varphi_\delta^{coinc}},\alpha}$

and $\phi^{\star MC,-}_{n,h_{\varphi^{coinc}_{\delta}},\alpha}$ are respectively defined by

$$\begin{cases} p_W^{\star MC,+} &= \frac{1}{B+1}\left(1+\sum_{b=1}^{B}\mathbb{1}_{C^{\star b}\geq C(\mathbf{X}_n)}\right), \\ p_W^{\star MC,-} &= \frac{1}{B+1}\left(1+\sum_{b=1}^{B}\mathbb{1}_{C^{\star b}\leq C(\mathbf{X}_n)}\right). \end{cases}$$

From [RW05, Lemma 1] (see Lemma 4), the tests that reject H_W when $p_W^{\star MC,+}$ or $p_W^{\star MC,-}$ is less than α in $(0,1)$ are proved to be both of level α.

5.2.3 Permutation UE method

Let us now consider the whole collection of windows $\mathcal{W}$. Dealing with the multiple testing problem described above, we propose a multiple test named the *Permutation UE method*, which consists in integrating the above single permutation tests based on the p-values $p_W^{\star MC,+}$ and/or $p_W^{\star MC,-}$ for W in $\mathcal{W}$, in Benjamini and Hochberg's procedure.
Notice that when we only consider the upper-tailed tests that is the p-values $p_W^{\star MC,+}$ (or similarly only the lower-tailed tests), when the windows in the collection are disjoint, and when the considered point processes in the model $\mathcal{M}^{(2)}_{\text{point proc.}}$ are (nonnecessarily homogeneous) Poisson processes, the p-values are independent. Therefore, the original procedure of Benjamini and Hochberg is proved to control the FDR_P for every P in $\mathcal{P}$. The correction of Benjamini and Yekutieli can be used in other cases.
We give below the complete algorithm that enables to assess if the coincidence count is significantly too large or too small on each window, which is a useful information for neuroscientists.

Permutation UE algorithm

Fix real numbers $\delta > 0$ and q in $(0, 0.5)$ and an integer B larger than 2.

- Do in parallel for all window $W = [a, b]$ in $\mathcal{W}$:
  - * Extract the points of the X_i^1's and X_i^2's in $[a, b]$.
  - * For all (i, j) in $\{1, ..., n\}^2$, compute $a_{i,j} = \varphi_{\delta}^{coinc}\left(X_i^1, X_j^2\right)$ over $[a, b]$ by the *delayed coincidence count algorithm* (see [15])
  - * Draw at random B i.i.d. permutations Π_n^b, $1 \leq b \leq B$, and compute $C^{\star b} = \sum_i a_{i,\Pi_n^b(i)}$.
  - * Compute also $C^{obs} = \sum_i a_{i,i}$.
  - * Return $p_W^+ = \frac{1}{B+1}\left(1+\sum_{b=1}^{B}\mathbb{1}_{C^{\star b}\geq C^{obs}}\right)$ and $p_W^- = \frac{1}{B+1}\left(1+\sum_{b=1}^{B}\mathbb{1}_{C^{\star b}\leq C^{obs}}\right)$.
- Perform the procedure of [BH95] on the set of the above $2\#\mathcal{W}$ p-values:
  - * Sort the p-values $p^{(1)} \leq ... \leq p^{(2\#\mathcal{W})}$.
  - * Find $k = \max\{i,\ p^{(i)} \leq iq/(2\#\mathcal{W})\}$.
  - * Return all the (W, ε_W)'s, for which W is associated with one of the p-values $p^{(i)}$ for $i \leq k$, with $\varepsilon_W = 1$ if $p_W^+ \leq p^{(k)}$, so the coincidence count is significantly too large on W, and $\varepsilon_W = -1$ if $p_W^- \leq p^{(k)}$, so the coincidence count is significantly too small on W.

The code has been parallelized in `C++` and interfaced with `R`. The full corresponding `R`-package is a work in progress but the program is available at: `https://github.com/ybouret/neuro-stat`.

The experimental results on the chosen real data, with $\delta = 0.02$s, $q = 0.05$ and $B = 10000$ are presented in Figure 5.1. Our Permutation UE method (P) is compared with the MTGAUE method (MTGAUE) of [TMRGRB14], devoted to Poisson processes, and the trial-shuffling method (TSC) of [PG03].

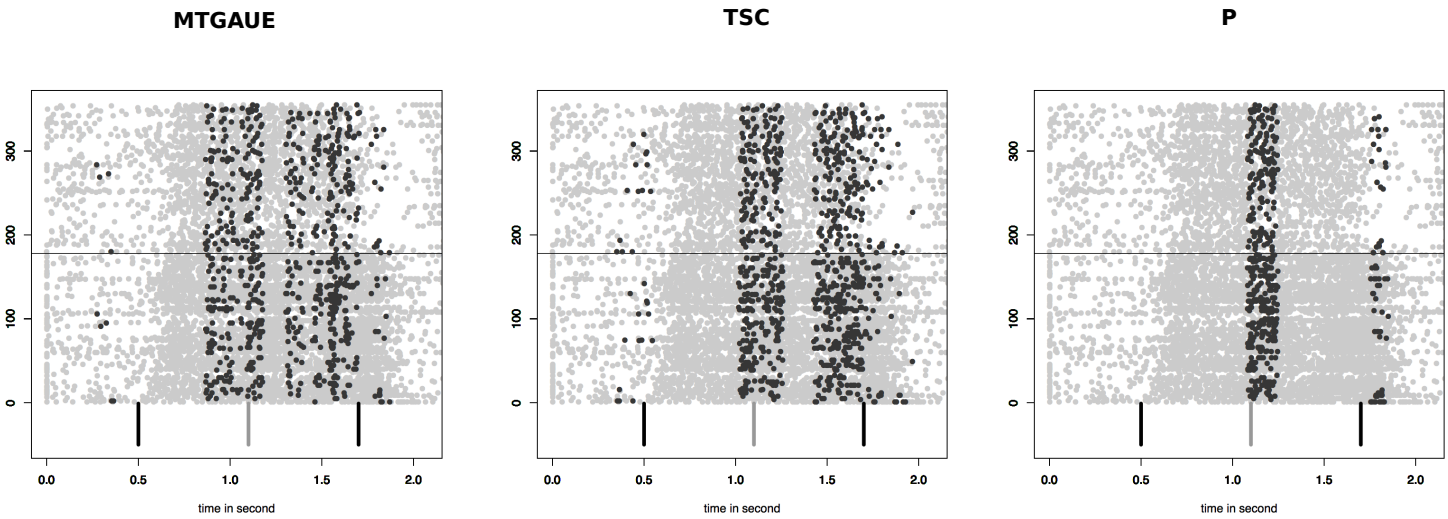

Figure 5.1 – Raster plots for the pair of neurons 13. In black the Unitary Events where the coincidence count is significantly too large for the three methods (MTGAUE, TSC and P). No interval was detected for a significantly too small coincidence count. Signs on bottom correspond to behavioral events. The first black vertical bar corresponds to the preparatory signal (PS), the gray vertical bar to the expected signal (ES), the second black vertical bar to the response signal (RS).

Permutation UE method detects less windows than both (MTGAUE) and (TSC) methods, but the detected windows are still in adequation with the experimental or behavioral events. The simulation study performed in [15] let us think that the extra detections of both (MTGAUE) and (TSC) may be false discoveries, since both methods do not control the FDR as well as the Permutation UE method.

5.3 Family-Wise Separation Rates for multiple testing

The details of the work described in this section can be found in [14]. It is the result of a collaboration with Matthieu Lerasle and Patricia Reynaud-Bouret, which is currently continuing with Nicolas Verzelen.

As stated in the introduction, many first kind error-related criteria have been introduced in the multiple testing literature, such as the wFWER, the FWER or more generally the $k-$FWER, the PFER, the FDP, or the popular FDR used in Section 5.2. By contrast, very few articles deal with the optimality of multiple tests in terms of second kind error. The articles by Lehmann, Romano, and Shaffer [LRS05], and by Romano, Shaikh and Wolf [RSW11] both give maximin type optimality results, but each with a different notion of maximin optimality. While Romano, Shaikh, and Wolf [RSW11] consider the minimum probability of rejecting at least one hypothesis when the hypotheses are not all true simultaneously, Lehmann, Romano, and Shaffer [LRS05] consider the minimum probability of rejecting one or more false hypotheses when at least one hypothesis deviates from the truth at a given degree.
We propose here new second kind error-related criteria to evaluate multiple tests whose FWER is controlled by a prescribed level α in $(0,1)$, inspired by the nonasymptotic minimax theory for nonparametric tests of a single null hypothesis introduced by Baraud [Bar02].
The literature on minimax and minimax adaptive testing is huge (see e.g., Chapter 1 and Chapter 3), and provides a now well-known and convenient framework to study the theoretical performance of nonparametric tests of single null hypotheses. Our purpose here is to provide such a framework in the multiple testing context.
Most of minimax adaptive tests of a single null hypothesis (H_0) are based on the aggregation of a collection of minimax tests for different null hypotheses, all related to (H_0). We first investigate the

parallel that can be drawn between such aggregated tests, and some classical single-step or step-down multiple testing procedures. From this parallel, we define the criterion of *weak Family-Wise Separation Rate*, denoted by wFWSR, which extends the notion of uniform separation rate for tests of a single null hypothesis to the multiple testing context, and its stronger counterpart: the (strong) *Family-Wise Separation Rate*, denoted by FWSR. This last criterion is in fact the key point to lay the foundations of a minimax theory for multiple tests whose FWER is controlled by a prescribed level α. The FWSR and its corresponding benchmark, the *minimax Family-Wise Separation Rate*, are thus new tools to evaluate the second kind error performance of a multiple test. Considering simple multiple testing problems in Gaussian regression frameworks, we prove for instance that in some cases, the FWSR of all the Bonferroni, Holm and min-p procedures are optimal from this minimax point of view, whereas in other cases, the Bonferroni procedure is clearly sub-optimal. Beyond the evaluation of a multiple test itself, the minimax Family-Wise Separation Rate can also be viewed as an indicator of the difficulty or complexity of the considered testing problem. In particular, we exhibit general conditions on the considered hypotheses, which guarantee that the minimax Family-Wise Separation Rate for multiple tests is lower bounded by the classical minimax Separation Rate for single tests, thus formalizing the intuition that multiple testing is more difficult than single testing. Through our illustrations in Gaussian regression frameworks, we furthermore prove that when these general conditions are not satisfied, the minimax Family-Wise Separation Rate for multiple tests may be smaller than the classical minimax Separation Rate for single tests, which may suggest, looking at things superficially, that multiple testing may be easier than single testing in some cases. This apparent counter-intuitive result in fact leads to a deeper analysis of the introduced criteria, and to a further reflection about the basic nature of a multiple testing problem, focusing on its fundamental differences with single testing problems. The emphasis is here placed on the importance attached, in a multiple testing problem, to each individual tested hypotheses, contrary to an aggregation-based single testing problem where only a single null hypothesis contained in all the tested hypotheses has to be taken into account.

5.3.1 Parallel between aggregated tests and multiple tests

In the following, for any subset $\mathcal{G}$ of $\mathcal{H}$, $\cap\mathcal{G}$ is an abbreviation for $\cap_{H\in\mathcal{G}}H$, e.g., $\cap\mathcal{H} = \cap_{H\in\mathcal{H}}H$.
Regarding Lemma 1, as explained above, in order to more conveniently draw a parallel between aggregated tests, that are usually expressed through test statistics T_H and corresponding càdlàg quantiles $F^{-1}_{H,-}(1-\alpha)$, and multiple tests that are usually expressed through p-values $p_H = 1 - F_{H,-}(T_H)$, we focus all along this section on single tests on the form: $\mathbb{1}_{\{T_H > F^{-1}_{H,-}(1-\alpha)\}} = \mathbb{1}_{\{p_H<\alpha\}}$. In particular, for sake of simplicity, when we refer in the sequel to well-known procedures such as Bonferroni or Holm's ones, we in fact refer to the versions of these procedures written via single tests of this form, though the original ones are in fact written with single tests of the form $\mathbb{1}_{\{p_H\leq\alpha\}}$.

Multiple tests controlling the Family-Wise Error Rate. Considering the FWER as first kind error rate evaluation criterion, one main concern is to construct a multiple test $\mathcal{R}$ such that

$$\mathrm{FWER}\,(\mathcal{R}\,) \leq \alpha, \tag{5.3}$$

for a given prescribed level α in $(0,1)$, which obviously also implies that $w\mathrm{FWER}\,(\mathcal{R}\,) \leq \alpha$.
A large number of multiple tests satisfying (5.3) have been constructed, among them the historical procedures of Bonferroni $\mathcal{R}_{Bonf}$ or $\mathcal{R}_{wBonf}$ described above, and of Holm [Sim86, Hol79], and the more recent min-p type procedures (see [DVDL07] for instance). Many of these procedures can be described via the general sequential rejection scheme proposed by Goeman and Solari [GS10], which consists in iteratively rejecting hypotheses through an application $\mathcal{N}$ from the set of all subsets of $\mathcal{H}$ to itself, as follows.

1. Start with $\mathcal{R}_0 = \emptyset$.
2. For any $n \geq 0$, build $\mathcal{R}_{n+1} = \mathcal{R}_n \cup \mathcal{N}(\mathcal{R}_n)$.
3. Define $\mathcal{R} = \lim_{n\to\infty} \mathcal{R}_n$.

For any prescribed α in $(0,1)$, Goeman and Solari [GS10, Theorem 1] proved that sequential rejective procedures satisfy (5.3), as soon as the two conditions below are true:

$$\forall \mathcal{S} \subset \mathcal{S}' \subset \mathcal{H},\ \mathcal{N}(\mathcal{S}) \subset \mathcal{S}' \cup \mathcal{N}(\mathcal{S}'), \tag{5.4}$$

$$\forall P \in \mathcal{P},\ P\left(\mathcal{N}(\mathcal{F}(P)) \subset \mathcal{F}(P)\right) \geq 1-\alpha. \tag{5.5}$$

Let us focus on a generic example, the min-p procedure, assuming that a collection of p-values p_H, for H in $\mathcal{H}$, is given, as in the definition of $\mathcal{R}_{Bonf}$ (5.1) and $\mathcal{R}_{wBonf}$ (5.2) above.
For any subset $\mathcal{G}$ of $\mathcal{H}$, and any α in $(0,1)$, let $c_{mp,\mathcal{G},\alpha}$ be a nonincreasing function of $\mathcal{G}$ such that

$$\forall P \in \cap\mathcal{G},\ P\left(\min_{H\in\mathcal{G}} p_H < c_{mp,\mathcal{G},\alpha}\right) \leq \alpha.$$

Then the min-p procedure is defined as a sequential rejective procedure with $\mathcal{N}$ equal to

$$\mathcal{N}_{mp} : \mathcal{S} \mapsto \left\{ H \in \mathcal{H} \setminus \mathcal{S},\ p_H < c_{mp,\mathcal{H}\setminus\mathcal{S},\alpha} \right\}.$$

As it satisfies (5.4) and (5.5), by [GS10, Theorem 1], the min-p procedure has a FWER controlled by α. It is always possible to use $c_{mp,\mathcal{G},\alpha} = \alpha/\#\mathcal{G}$: the obtained multiple test is due to Holm [Hol79], so we denote it by $\mathcal{R}_{Holm}$ and the corresponding application by $\mathcal{N}_{Holm}$.
Remark that the first step of this procedure in fact corresponds to $\mathcal{R}_{Bonf}$.
If the distribution of $\min_{H\in\mathcal{G}} p_H$ (with càglàd c.d.f. $F_{\mathcal{G},-}$) does not depend on P in $\cap\mathcal{G}$ and is known, one can now take $c_{mp,\mathcal{G},\alpha} = F^{-1}_{\mathcal{G},-}(\alpha)$. The resulting rejection set is then denoted by $\mathcal{R}_{mp}$. Note that this multiple testing procedure is less conservative than $\mathcal{R}_{Holm}$, that is, $\mathcal{R}_{Holm} \subset \mathcal{R}_{mp}$. If $F_{\mathcal{G},-}$ is unknown, the quantiles may be replaced by random quantiles, depending on $\mathbf{X}$, based on permutation or bootstrap approaches [RW05, RW07, RW10], at the possible price of an asymptotic control of the FWER instead of an exact control. Finally, as for the Bonferroni procedure, the min-p procedures may also be extended to weighted min-p procedures by defining

$$\mathcal{N}_{wmp} : \mathcal{S} \mapsto \left\{ H \in \mathcal{H} \setminus \mathcal{S},\ p_H < w_H\ c_{wmp,\mathcal{H}\setminus\mathcal{S},\alpha} \right\},$$

where $(w_H)_{H\in\mathcal{H}}$ is still a family of positive weights satisfying $\sum_{H\in\mathcal{H}} w_H \leq 1$, and where $c_{wmp,\mathcal{G},\alpha}$ satisfies for any α in $(0,1)$,

$$\forall P \in \cap\mathcal{G},\ P\left(\min_{H\in\mathcal{G}} w_H^{-1} p_H < c_{wmp,\mathcal{G},\alpha}\right) \leq \alpha.$$

When the distribution of $\min_{H\in\mathcal{G}} w_H^{-1} p_H$ (with càglàd c.d.f. $F_{w,\mathcal{G},-}$) does not depend on P in $\cap\mathcal{G}$ and is known, one can take $c_{wmp,\mathcal{G},\alpha} = F^{-1}_{w,\mathcal{G},-}(\alpha)$, which defines rejection sets denoted by $\mathcal{R}_{wmp}$. Note that these last procedures are very close to the balanced procedure of Romano and Wolf [RW10].

Aggregated tests controlling the first kind error rate. Let us now consider the problem of testing a single null hypothesis $(H_0)\ P \in \mathcal{P}_0$ against $(H_1)\ P \in \mathcal{P} \setminus \mathcal{P}_0$, and recall, in the present notation, the principle of aggregated tests described in Section 1.1.2.
Let $\mathcal{H}$ be a collection of hypotheses, chosen such that $\mathcal{P}_0 \subset \cap\mathcal{H}$. For each hypothesis H in the collection $\mathcal{H}$, an individual test ϕ_H of the null hypothesis H against the alternative $\mathcal{P} \setminus H$ is constructed. The obtained collection of tests is here denoted by $\Phi_{\mathcal{H}} = \{\phi_H,\ H \in \mathcal{H}\}$.

Then, the corresponding aggregated test $\bar{\Phi}_{\mathcal{H}}$ consists in rejecting (H_0) if at least one H in $\mathcal{H}$ is rejected with ϕ_H, that is

$$\bar{\Phi}_{\mathcal{H}} = \sup_{H\in\mathcal{H}} \phi_H. \tag{5.6}$$

The first kind error rate of an aggregated test $\bar{\Phi}_{\mathcal{H}}$ of the single null hypothesis (H_0) is defined as in (1) by

$$\mathrm{ER}_1\left(\bar{\Phi}_{\mathcal{H}}, \mathcal{P}_0\right) = \sup_{P\in\mathcal{P}_0} P\left(\bar{\Phi}_{\mathcal{H}} = 1\right) = \sup_{P\in\mathcal{P}_0} P\left(\sup_{H\in\mathcal{H}} \phi_H = 1\right).$$

Following the Neyman-Pearson principle, this criterion should be controlled by a prescribed level α in $(0,1)$. For any hypothesis H of the collection $\mathcal{H}$, the individual test ϕ_H is usually defined from a test statistic T_H, whose distribution does not depend on P provided that P belongs to $\mathcal{P}_0$. Respectively denoting by $F_{H,-}$ and $F_{H,-}^{-1}$ the càglàd c.d.f. and càdlàg quantile function of this distribution under (H_0), ϕ_H is then defined as $\mathbb{1}_{\left\{T_H > F_{H,-}^{-1}(1-u_{H,\alpha})\right\}}$, where $u_{H,\alpha}$ is chosen so that the aggregated test is actually of level α, that is

$$\mathrm{ER}_1\left(\bar{\Phi}_{\mathcal{H}}, \mathcal{P}_0\right) \leq \alpha.$$

The most obvious choice for $u_{H,\alpha}$ is a Bonferroni-type choice $u_{H,\alpha} = \alpha/\#\mathcal{H}$. This leads to the Bonferroni-type aggregated test $\bar{\Phi}_{\mathcal{H}}^{Bonf}$ based on the collection

$$\Phi_{\mathcal{H}}^{Bonf} = \left\{ \phi_H^{Bonf} = \mathbb{1}_{\left\{T_H > F_{H,-}^{-1}(1-\alpha/\#\mathcal{H})\right\}},\ H\in\mathcal{H} \right\}.$$

A weighted Bonferroni-type choice $u_{H,\alpha} = w_H\alpha$ can also be considered where $(w_H)_{H\in\mathcal{H}}$ is a family of positive weights such that $\sum_{H\in\mathcal{H}} w_H \leq 1$. This leads to the weighted Bonferroni-type aggregated test $\bar{\Phi}_{\mathcal{H}}^{wBonf}$ based on the collection

$$\Phi_{\mathcal{H}}^{wBonf} = \left\{ \phi_H^{wBonf} = \mathbb{1}_{\left\{T_H > F_{H,-}^{-1}(1-w_H\alpha)\right\}},\ H\in\mathcal{H} \right\}.$$

A less conservative choice in practice and still guaranteeing a level α consists in taking

$$u_{H,\alpha} = w_H \sup\left\{ u,\ \sup_{P\in H_0} P\left(\exists H\in\mathcal{H},\ T_H > F_{H,-}^{-1}(1-w_H u)\right) \leq \alpha \right\}.$$

This leads, when $w_H = 1/\#\mathcal{H}$, to the aggregated test $\bar{\Phi}_{\mathcal{H}}^{BHL}$ proposed by Baraud, Huet, and Laurent [BHL03], based on the collection

$$\Phi_{\mathcal{H}}^{BHL} = \left\{ \phi_H^{BHL} = \mathbb{1}_{\left\{T_H > F_{H,-}^{-1}(1-u_\alpha)\right\}},\ H\in\mathcal{H} \right\},$$

with

$$u_\alpha = \sup\left\{ u,\ \sup_{P\in H_0} P\left(\exists H\in\mathcal{H},\ T_H > F_{H,-}^{-1}(1-u)\right) \leq \alpha \right\},$$

and, in the general case, to the aggregated test $\bar{\Phi}_{\mathcal{H}}^{FLR}$ proposed in [7], based on the collection

$$\Phi_{\mathcal{H}}^{FLR} = \left\{ \phi_H^{FLR} = \mathbb{1}_{\left\{T_H > F_{H,-}^{-1}(1-u_{H,\alpha})\right\}},\ H\in\mathcal{H} \right\}.$$

Correspondences between multiple and aggregated tests. Let us here present the main correspondences that can be underlined between multiple tests and aggregated tests. To do so, we always assume that a finite collection of hypotheses $\mathcal{H}$ and a single null hypothesis (H_0) $P\in\mathcal{P}_0$, with $\mathcal{P}_0 \subset \cap\mathcal{H}$, are given.

From any collection $\Phi_{\mathcal{H}} = \{\phi_H,\ H \in \mathcal{H}\}$ of tests ϕ_H of the single hypothesis H defining an aggregated test, a multiple test of $\mathcal{H}$ is constructed as

$$\mathcal{R}\left(\Phi_{\mathcal{H}}\right) = \left\{ H \in \mathcal{H},\ \phi_H = 1 \right\}.$$

Conversely, from any multiple test $\mathcal{R}$ of $\mathcal{H}$, we construct

$$\bar{\Phi}(\mathcal{R}) = \mathbb{1}_{\{\mathcal{R} \neq \emptyset\}},$$

which can be seen as an aggregated test of the single null hypothesis (H_0).

First notice that

$$w\mathrm{FWER}\left(\mathcal{R}\left(\Phi_{\mathcal{H}}\right)\right) = \mathrm{ER}_1\left(\bar{\Phi}_{\mathcal{H}}, \cap\mathcal{H}\right),$$

and conversely

$$w\mathrm{FWER}\left(\mathcal{R}\right) = \mathrm{ER}_1\left(\bar{\Phi}\left(\mathcal{R}\right), \cap\mathcal{H}\right). \tag{5.7}$$

Since $\mathcal{P}_0 \subset \cap\mathcal{H}$, $w\mathrm{FWER}\left(\mathcal{R}\left(\Phi_{\mathcal{H}}\right)\right) \geq \mathrm{ER}_1\left(\bar{\Phi}_{\mathcal{H}}, \mathcal{P}_0\right)$ and $w\mathrm{FWER}\left(\mathcal{R}\right) \geq \mathrm{ER}_1\left(\bar{\Phi}\left(\mathcal{R}\right), \mathcal{P}_0\right)$. Except when $\mathcal{P}_0 = \cap\mathcal{H}$, controlling $w\mathrm{FWER}\left(\mathcal{R}\left(\Phi_{\mathcal{H}}\right)\right)$ or $w\mathrm{FWER}\left(\mathcal{R}\right)$ is thus more difficult than controlling $\mathrm{ER}_1\left(\bar{\Phi}_{\mathcal{H}}, \mathcal{P}_0\right)$ or $\mathrm{ER}_1\left(\bar{\Phi}\left(\mathcal{R}\right), \mathcal{P}_0\right)$ respectively.

Next, assume that for every H in $\mathcal{H}$, a test statistic T_H, whose distribution does not depend on P provided that P belongs to H, is given and denote by p_H its corresponding p-value, as defined by Lemma 1. We prove in [14, Proposition 2] that for such a collection of p-values $\{p_H,\ H \in \mathcal{H}\}$, $\mathcal{R}(\Phi_{\mathcal{H}}^{Bonf}) = \mathcal{R}_{Bonf}$, and $\bar{\Phi}_{\mathcal{H}}^{Bonf} = \bar{\Phi}(\mathcal{R}_{Bonf}) = \bar{\Phi}(\mathcal{R}_{Holm})$. If additionally the distribution of $\min_{H\in\mathcal{H}} w_H^{-1} p_H$ does not depend on P provided that P belongs to $\cap\mathcal{H}$, then $\mathcal{N}_{wmp}(\emptyset) = \mathcal{R}(\Phi_{\mathcal{H}}^{FLR})$ and $\bar{\Phi}_{\mathcal{H}}^{FLR} = \bar{\Phi}(\mathcal{R}_{wmp})$.

Notice that the assumptions needed to establish [14, Proposition 2] are quite strong, and that few frameworks satisfy them. Among these frameworks, we will focus on classical Gaussian regression ones to illustrate our results.

5.3.2 From uniform Separation Rates to Family-Wise Separation Rates

Let d be a distance on $\mathcal{P}$, and for any P in $\mathcal{P}$, any subset $\mathcal{Q}$ of $\mathcal{P}$, let $d\left(P, \mathcal{Q}\right) = \inf_{Q\in\mathcal{Q}} d(P, Q)$.

Uniform separation rates for aggregated tests. As seen in Chapter 1 and Chapter 3, uniform separation rates are second kind error-related quality criteria of a test of

$$(H_0) \quad P \in \mathcal{P}_0 \subset \mathcal{P} \quad \text{against} \quad (H_1) \quad P \in \mathcal{P} \setminus \mathcal{P}_0.$$

Because $\mathcal{P}$ is in general too large to define separation rates over the whole set $\mathcal{P}$ properly, particularly in nonparametric frameworks, these quantities are first defined on a subset $\mathcal{Q}$ of $\mathcal{P}$. The question of adaptivity with respect to $\mathcal{Q}$ can then be treated. More precisely, let us recall the following definitions due to Baraud [Bar02], which can be viewed as nonasymptotic versions of Ingster's definitions [Ing93].

Definition 9 (Uniform and minimax separation rate)**.** Let α and β be fixed error rates levels in $(0, 1)$, and let $\bar{\Phi}$ be a level α test of a null hypothesis (H_0) $P \in \mathcal{P}_0 \subset \mathcal{P}$. For a subset $\mathcal{Q}$ of $\mathcal{P}$, the uniform separation rate of $\bar{\Phi}$ over $\mathcal{Q}$ with prescribed second kind error rate β is defined by

$$\mathrm{SR}_d^{\beta}\left(\bar{\Phi}, \mathcal{Q}, \mathcal{P}_0\right) = \inf\left\{ r > 0,\ \sup_{P\in\mathcal{Q},\ d(P,\mathcal{P}_0)\geq r} P(\bar{\Phi} = 0) \leq \beta \right\}.$$

Note that this definition holds for any null hypothesis, that is any subset $\mathcal{P}_0$ of $\mathcal{P}$, and in particular for $\cap\mathcal{H}$. Hence when $\mathcal{P}_0 \subset \cap\mathcal{H}$, $\mathrm{SR}_d^{\beta}\left(\bar{\Phi}, \mathcal{Q}, \mathcal{P}_0\right) \geq \mathrm{SR}_d^{\beta}\left(\bar{\Phi}, \mathcal{Q}, \cap\mathcal{H}\right)$.

The corresponding minimax separation rate over $\mathcal{Q}$ with prescribed error rates α and β is defined as

$$m\mathrm{SR}_d^{\alpha,\beta}\left(\mathcal{Q}, \mathcal{P}_0\right) = \inf_{\bar{\Phi},\ \mathrm{ER}_1\left(\bar{\Phi}, \mathcal{P}_0\right)\leq\alpha} \mathrm{SR}_d^{\beta}\left(\bar{\Phi}, \mathcal{Q}, \mathcal{P}_0\right),$$

where the infimum is taken over all possible level α tests.

Definition 10 (Minimax (adaptive) test)**.** Let $\overline{\mathcal{Q}}$ be a collection of classes of probability distributions $\mathcal{Q} \subset \mathcal{P}$. A level α test $\bar{\Phi}$ of the null hypothesis (H_0) $P \in \mathcal{P}_0 \subset \mathcal{P}$ is said to be minimax over a class $\mathcal{Q}$ of the collection $\overline{\mathcal{Q}}$ if $\mathrm{SR}_d^\beta\left(\bar{\Phi}, \mathcal{Q}, \mathcal{P}_0\right)$ achieves $m\mathrm{SR}_d^{\alpha,\beta}(\mathcal{Q}, \mathcal{P}_0)$, possibly up to a multiplicative constant depending on α and β. It is said to be minimax adaptive over $\overline{\mathcal{Q}}$ if $\mathrm{SR}_d^\beta\left(\bar{\Phi}, \mathcal{Q}, \mathcal{P}_0\right)$ achieves, or nearly achieves, $m\mathrm{SR}_d^{\alpha,\beta}(\mathcal{Q}, \mathcal{P}_0)$, for all the classes $\mathcal{Q}$ in $\overline{\mathcal{Q}}$ simultaneously, without knowing in advance to which class of the collection the distribution P may belong.

Family-Wise Separation Rates for multiple tests. Following the idea of the definition of the weak Family-Wise Error Rate wFWER of $\mathcal{R}$, which is in fact equal to the first kind error rate of $\bar{\Phi}(\mathcal{R})$ for the null hypothesis $\cap\mathcal{H}$ (see (5.7)), a natural idea would be to define a notion of weak Family-Wise Separation Rate as

$$\mathrm{SR}_d^\beta\left(\bar{\Phi}(\mathcal{R}), \mathcal{Q}, \cap\mathcal{H}\right) = \inf\left\{ r > 0, \sup_{P\in\mathcal{Q},\ d(P,\cap\mathcal{H})\geq r} P(\mathcal{R} = \emptyset) \leq \beta \right\}.$$

However, in this second kind error criterion, only alternatives which deviate from the intersection $\cap\mathcal{H}$ with a certain distance are taken into account. Considering such definition would thus amount to confusing multiple tests with their corresponding aggregated tests, seeing all the tested hypotheses as only intermediate hypotheses to an ultimate one: $\cap\mathcal{H}$. This would depart from the multiple testing philosophy, where each tested hypothesis has its own significance and has to be taken into account by itself. In order to address this requirement, instead of alternatives P in $\mathcal{Q}$ such that "$d(P, \cap\mathcal{H}) \geq r$" (for $r > 0$), are considered alternatives P in $\mathcal{Q}$ such that "$\exists H \in \mathcal{H},\ d(P, H) \geq r$".
So, we introduce the set of false hypotheses under P at least at distance r from P, that is $\mathcal{F}_r(P) = \{H \in \mathcal{H},\ d(P, H) \geq r\}$, which enables us to introduce the following notion of weak Family-Wise Separation Rate for a multiple test.

Definition 11 (Weak Family-Wise Separation Rate)**.** Let α and β be fixed error rates levels in $(0, 1)$, and let $\mathcal{R}$ be a multiple test of $\mathcal{H}$, whose FWER is controlled by α. For any subset $\mathcal{Q}$ of $\mathcal{P}$, the weak Family-Wise Separation Rate of $\mathcal{R}$ over $\mathcal{Q}$ with prescribed second kind error rate β is defined by

$$\begin{aligned} w\mathrm{FWSR}_d^\beta(\mathcal{R}, \mathcal{Q}) &= \inf\left\{ r > 0, \sup_{P\in\mathcal{Q},\ \mathcal{F}_r(P)\neq\emptyset} P(\mathcal{R} = \emptyset) \leq \beta \right\} \\ &= \inf\left\{ r > 0, \inf_{P\in\mathcal{Q},\ \mathcal{F}_r(P)\neq\emptyset} P(\mathcal{R} \neq \emptyset) \geq 1 - \beta \right\}. \end{aligned}$$

Note that the quantity $\inf_{P\in\mathcal{Q},\ \mathcal{F}_r(P)\neq\emptyset} P(\mathcal{R} \neq \emptyset)$ involved in the above definition is clearly related to $\beta_{\#\mathcal{H},1}(\alpha, r) = \inf_{P\in\mathcal{Q},\ \mathcal{F}_r(P)\neq\emptyset} P(\mathcal{R} \cap \mathcal{F}(P) \neq \emptyset)$ in [LRS05] which is expected to be maximized in the maximin optimality criterion.
Furthermore, it is easy to see that $w\mathrm{FWSR}_d^\beta(\mathcal{R}, \mathcal{Q}) \leq \mathrm{SR}_d^\beta\left(\bar{\Phi}(\mathcal{R}), \mathcal{Q}, \cap\mathcal{H}\right)$, with an equality if the collection of hypotheses $\mathcal{H}$ and the distance d satisfy

$$\forall r > 0, \quad \mathcal{F}_r(P) \neq \emptyset \quad \text{if and only if} \quad d(P, \cap\mathcal{H}) \geq r. \tag{5.8}$$

In particular, if the collection of hypotheses $\mathcal{H}$ is closed (under intersection), then condition (5.8) is always satisfied. We state in [14] the following more general and useful result.

Proposition 7 (Fromont, Lerasle, Reynaud-Bouret, 2015)**.** *Let d be a distance on $\mathcal{P}$, and $\mathcal{Q}$ be a subset of $\mathcal{P}$. If there exists some distance d' on $\mathcal{P}$ such that:*

$$\forall P \in \mathcal{Q},\ \forall r > 0, \quad \mathcal{F}_r(P) \neq \emptyset \quad \textit{if and only if} \quad d'(P, \cap\mathcal{H}) \geq r, \tag{5.9}$$

then for every β in $(0, 1)$, for every multiple test $\mathcal{R}$ of $\mathcal{H}$,

$$w\mathrm{FWSR}_d^\beta(\mathcal{R}, \mathcal{Q}) = \mathrm{SR}_{d'}^\beta\left(\bar{\Phi}(\mathcal{R}), \mathcal{Q}, \cap\mathcal{H}\right).$$

We now introduce the stronger notion of Family-Wise Separation Rate.

Definition 12 (Family-Wise Separation Rate)**.** Let α and β be fixed error rates levels in $(0,1)$, and let $\mathcal{R}$ be a multiple test of $\mathcal{H}$, whose FWER is controlled by α. For any subset $\mathcal{Q}$ of $\mathcal{P}$, the Family-Wise Separation Rate of $\mathcal{R}$ over $\mathcal{Q}$ with prescribed second kind error rate β is defined by

$$\begin{aligned}\mathrm{FWSR}_d^{\beta}(\mathcal{R},\mathcal{Q}) &= \inf\left\{ r>0,\ \sup_{P\in\mathcal{Q}} P\left(\mathcal{F}_r(P)\cap(\mathcal{H}\setminus\mathcal{R})\neq\emptyset\right)\leq\beta\right\}\\ &= \inf\left\{ r>0,\ \inf_{P\in\mathcal{Q}} P\left(\mathcal{F}_r(P)\subset\mathcal{R}\right)\geq 1-\beta\right\}.\end{aligned}$$

Note that $w\mathrm{FWSR}_d^{\beta}(\mathcal{R},\mathcal{Q})\leq \mathrm{FWSR}_d^{\beta}(\mathcal{R},\mathcal{Q})$.
Let us now introduce the corresponding minimax approach for multiple tests.

Definition 13 (Minimax (adaptive) multiple test)**.** Let α and β be fixed error rates levels in $(0,1)$, and $\mathcal{Q}$ be a subset of $\mathcal{P}$.
The minimax Family-Wise Separation Rate over $\mathcal{Q}$ with prescribed FWER α and prescribed second kind error rate β is defined by

$$m\mathrm{FWSR}_d^{\alpha,\beta}(\mathcal{Q}) = \inf_{\mathcal{R},\ \mathrm{FWER}(\mathcal{R})\leq\alpha} \mathrm{FWSR}_d^{\beta}(\mathcal{R},\mathcal{Q}),$$

where the infimum is taken over all possible multiple tests with a FWER controlled by α.
A multiple test $\mathcal{R}$, whose FWER is controlled by α, is then said to be minimax over $\mathcal{Q}$ if $\mathrm{FWSR}_d^{\beta}(\mathcal{R},\mathcal{Q})$ achieves $m\mathrm{FWSR}_d^{\alpha,\beta}(\mathcal{Q})$, possibly up to a multiplicative constant depending on α and β.
Finally, it is said to be minimax adaptive over a collection $\overline{\mathcal{Q}}$ of classes $\mathcal{Q}$ if $\mathrm{FWSR}_d^{\beta}(\mathcal{R},\mathcal{Q})$ achieves, or nearly achieves, $m\mathrm{FWSR}_d^{\alpha,\beta}(\mathcal{Q})$, for all the classes $\mathcal{Q}$ in $\overline{\mathcal{Q}}$ simultaneously, without knowing in advance to which class the distribution P may belong.

It is worth to underline that when the collection $\mathcal{H}$ is reduced to a single hypothesis or subset $\mathcal{P}_0$ of $\mathcal{P}$, for any subset $\mathcal{Q}$ of $\mathcal{P}$,

$$m\mathrm{FWSR}_d^{\alpha,\beta}(\mathcal{Q}) = m\mathrm{SR}_d^{\alpha,\beta}(\mathcal{Q},\mathcal{P}_0).$$

In this sense, the present minimax approach for multiple tests can be viewed as a generalization of the classical minimax theory for single hypothesis tests.

Links between minimax Separation Rates and minimax Family-Wise Separation Rates. Even when $\mathcal{H}$ is not reduced to a single hypothesis or subset $\mathcal{P}_0$ of $\mathcal{P}$, both theories also have, under particular conditions, close links that are established in the following result.

Theorem 13 (Fromont, Lerasle, Reynaud-Bouret, 2015)**.** *Let d be a distance on $\mathcal{P}$, and $\mathcal{Q}$ be a subset of $\mathcal{P}$. If there exists some distance d' on $\mathcal{P}$ satisfying* (5.9)*, then for every α,β in* $(0,1)$*,*

$$m\mathrm{FWSR}_d^{\alpha,\beta}(\mathcal{Q}) \geq m\mathrm{SR}_{d'}^{\alpha,\beta}(\mathcal{Q},\cap\mathcal{H}). \tag{5.10}$$

This result thus provides lower bounds for the minimax Family-Wise Separation Rates over some classes $\mathcal{Q}$ from the existing literature on classical minimax testing. As a particular case, if (5.8) holds, then for any subset $\mathcal{Q}$ of $\mathcal{P}$ and α,β in $(0,1)$,

$$m\mathrm{FWSR}_d^{\alpha,\beta}(\mathcal{Q}) \geq m\mathrm{SR}_d^{\alpha,\beta}(\mathcal{Q},\cap\mathcal{H}),$$

which formalizes the natural idea that testing multiple hypotheses is more difficult than testing a single hypothesis.

5.3.3 Illustrations in Gaussian regression frameworks

Independent Gaussian regression model. The observed random variable $\mathbf{X}$ is here assumed to be a random Gaussian vector according to the following model.

$\mathcal{M}^{(1)}_{\text{ind. reg.}}$	$\mathbf{X} = \mathbf{X}_n = (X_1, \ldots, X_n)'$ has a distribution $P = P_f$ defined by: $X_i = f_i + \sigma\varepsilon_i$ for i in $\{1, \ldots, n\}$, $f = (f_1, \ldots, f_n)'$ being an unknown real vector, the ε_i's being independent standard Gaussian random variables, and σ being a known positive real number.

Two different collections of hypotheses, that is two different collections of subsets of $\mathcal{P} = \{P_f,\ f = (f_1, \ldots, f_n)' \in \mathbb{R}^n\}$ are considered, that can be defined from the canonical basis $\{e_1, \ldots, e_n\}$ of $\mathbb{R}^n$. The first one is given by $\mathcal{H} = \{H_{S_i},\ i = 1, ..., n\}$, where for every i in $\{1, \ldots, n\}$, $S_i = \text{Vect}(e_i)$ and

$$H_{S_i} = \{P_f,\ f_i = 0\} = \{P_f,\ f \in S_i^{\perp}\}.$$

The second one is given by $\mathcal{H} = \{H_{\bar{S}_i},\ i = 1, ..., n\}$, where for every i in $\{1, \ldots, n\}$, $\bar{S}_i = \text{Vect}(e_1, \ldots, e_i)$, so

$$H_{\bar{S}_i} = \{P_f,\ f_1 = \ldots = f_i = 0\} = \left\{ P_f,\ f \in \bar{S}_i^{\perp} \right\}.$$

Both collections in particular satisfy $\cap\mathcal{H} = \{P_0\} := \mathcal{P}_0$.

We here consider various metrics on $\mathcal{P}$. Let for $g = (g_1, \ldots, g_n)'$ and $f = (f_1, \ldots, f_n)'$ in $\mathbb{R}^n$,

$$d_\infty(P_f, P_g) = \|f - g\|_\infty = \max_{i=1,\ldots,n} |f_i - g_i|, \tag{5.11}$$

and for $s \geq 1$,

$$d_s(P_f, P_g) = \left(\sum_{i=1}^{n} |f_i - g_i|^s \right)^{1/s}. \tag{5.12}$$

As in [Bar02], we investigate the minimax Family-Wise Separation Rates over the classes of alternatives $\mathcal{Q} = \mathcal{P}_k$ defined, for any integer $k \leq n$, by

$$\mathcal{P}_k = \{ P_f,\ |f|_0 \leq k \}, \tag{5.13}$$

where $|f|_0$ is the number of nonzero coefficients in f.
For these classes, one knows (see [Bar02]) in particular that for α and β in $(0, 1)$ such that $\alpha + \beta \leq 0.5$ and $k \geq 1$,

$$m\text{SR}^{\alpha,\beta}_{d_2}(\mathcal{P}_k, \mathcal{P}_0) \geq \sigma \left(k \ln \left(1 + \frac{n}{k^2} \vee \sqrt{\frac{n}{k^2}} \right) \right)^{1/2}, \tag{5.14}$$

and that this lower bound is tight.
Baraud, Huet and Laurent [BHL03] then build aggregated tests that are adaptive over a collection of classes $\mathcal{P}_k$, when σ^2 is not assumed to be known anymore, and Laurent, Loubes, Marteau [LLM12] further study the case of heteroscedasticity. In a preliminary version [LLM], they also prove that for α, β in $(0, 1)$ such that $\alpha + \beta \leq 0.5$,

$$m\text{SR}^{\alpha,\beta}_{d_\infty}(\mathcal{P}_k, \mathcal{P}_0) \geq \sigma\sqrt{\ln(1 + n)}, \tag{5.15}$$

by remarking that

$$m\text{SR}^{\alpha,\beta}_{d_\infty}(\mathcal{P}_k, \mathcal{P}_0) \geq m\text{SR}^{\alpha,\beta}_{d_\infty}(\mathcal{P}_1, \mathcal{P}_0) = m\text{SR}^{\alpha,\beta}_{d_2}(\mathcal{P}_1, \mathcal{P}_0)$$

and using Baraud's lower bound.

Let us firstly focus on the multiple testing problem of $\mathcal{H} = \{H_{S_i},\ i = 1, ..., n\}$. Let k be a fixed integer in $\{1, \dots, n\}$, and s in $[1, \infty]$. Using $d' = d_\infty$ in Theorem 13 leads to

$$m\mathrm{FWSR}_{d_s}^{\alpha,\beta}(\mathcal{P}_k) \geq m\mathrm{SR}_{d_\infty}^{\alpha,\beta}(\mathcal{P}_k, \mathcal{P}_0),$$

and from (5.15), we deduce that

$$m\mathrm{FWSR}_{d_s}^{\alpha,\beta}(\mathcal{P}_k) \geq \sigma\sqrt{\ln(1+n)}. \tag{5.16}$$

Let F be the c.d.f. of a standard Gaussian distribution, and for every i in $\{1, \dots, n\}$, p_i be the p-value associated with the test $\mathbb{1}_{\{|X_i|\sigma^{-1} > F^{-1}(1-\alpha/2)\}}$, given by $p_i = 1 - F\left(|X_i|\sigma^{-1}\right)$ (see Lemma 1). The multiple tests $\mathcal{R}_{Bonf}, \mathcal{R}_{Holm}, \mathcal{R}_{mp}$ and $\mathcal{R}(\Phi_{\mathcal{H}}^{BHL})$ based on these p-values are proved to have a FWSR for the distance d_s (s in $[1, \infty]$) over $\mathcal{P}_k$ upper bounded by

$$\sigma\left(\sqrt{2\ln(n/\alpha)} + \sqrt{2\ln(k/(2\beta))}\right).$$

This proves on the one hand that the minimax Family-Wise Separation Rate over $\mathcal{P}_k$ is of order $\sigma(\ln n)^{1/2}$. By comparison, the minimax Separation Rate $m\mathrm{SR}_{d_2}^{\alpha,\beta}(\mathcal{P}_k, \mathcal{P}_0)$ is of order $\sigma n^{\gamma/2}(\ln n)^{1/2}$ when k is proportional to n^γ for γ in $(0, 1/2)$ (see (5.14)), which is much larger than this minimax Family-Wise Separation Rate. This could let think that, when considering the distance $d = d_2$, performing a multiple testing procedure may be much easier than performing a test of a single hypothesis, which would be completely counter-intuitive. However, making such a comparison consists in comparing quantities that are not comparable at all. When $d = d_2$, the set of alternatives considered in the definition of $w\mathrm{FWSR}_{d_2}^{\beta}(\mathcal{R}, \mathcal{P}_k)$ of any multiple test $\mathcal{R}$ is in fact smaller than the set of alternatives in the definition of $\mathrm{SR}_{d_2}^{\beta}(\bar{\Phi}(\mathcal{R}), \mathcal{P}_k, \mathcal{P}_0)$, but exactly equal to the one in $\mathrm{SR}_{d_\infty}^{\beta}(\bar{\Phi}(\mathcal{R}), \mathcal{P}_k, \mathcal{P}_0)$. This explains why $m\mathrm{FWSR}_{d_2}^{\alpha,\beta}(\mathcal{P}_k)$, and also more generally $m\mathrm{FWSR}_{d_s}^{\alpha,\beta}(\mathcal{P}_k)$ (s in $[1, \infty]$), are in fact of the same order as the minimax Separation Rate $m\mathrm{SR}_{d_\infty}^{\alpha,\beta}(\mathcal{P}_k, \mathcal{P}_0)$ determined in [LLM].
This proves on the other hand that the four considered multiple tests $\mathcal{R}_{Bonf}, \mathcal{R}_{Holm}, \mathcal{R}_{mp}$ and $\mathcal{R}(\Phi_{\mathcal{H}}^{BHL})$ are minimax over the classes $\mathcal{P}_k$ with a Family-Wise Separation Rate of order $\sigma(\ln n)^{1/2}$, up to a multiplicative constant. Since the considered multiple tests do not depend on the value of k, they are moreover minimax adaptive over the whole collection of classes $\mathcal{P}_k$, for $k = 1 \dots n$. Notice that asymptotically, there is here no additional price to pay for adaptivity, phenomenon which is rather rarely observed in minimax adaptive testing problems (see Chapter 1 and Chapter 3 for instance).
Furthermore, notice that when $\mathcal{H}$ is reduced to a single hypothesis H_{S_i}, then $m\mathrm{FWSR}_{d_s}^{\alpha,\beta}(\mathcal{P}_k) = m\mathrm{SR}_{d_s}^{\alpha,\beta}(\mathcal{P}_k, H_{S_i})$, both being of order σ. In this sense, $(\ln n)^{1/2}$ can be viewed as the price to pay for multiplicity.
The study of the present Gaussian framework highlights another interesting point. Baraud's [Bar02] result gives that when $\sqrt{n} \leq k \leq n$, $m\mathrm{SR}_{d_2}^{\alpha,\beta}(\mathcal{P}_k, \mathcal{P}_0)$ is of order $\sigma n^{1/4}$.
As seen above, $\mathcal{R}_{Bonf}$ achieves an optimal $\mathrm{FWSR}_{d_2}^{\beta}$ over the $\mathcal{P}_k$'s. However, we prove in [14, Proposition 8] that its corresponding aggregated test $\bar{\Phi}(\mathcal{R}_{Bonf})$ does not necessarily achieve $m\mathrm{SR}_{d_2}^{\alpha,\beta}(\mathcal{P}_k, \mathcal{P}_0)$ when $\sqrt{n} \leq k \leq n$. Conversely, one can have some aggregated tests that are minimax, whereas the corresponding multiple tests are not.

Let us now secondly focus on the multiple testing problem of $\mathcal{H} = \{H_{\bar{S}_i},\ i = 1, ..., n\}$.
The main point that has to be pointed out here is that this collection of nested hypotheses is closed under intersection, and so (5.8) (or (5.9) with $d' = d$) is satisfied for $d = d_s$, with any s in $[1, \infty]$. From Theorem 13 and (5.14), we deduce in particular, that for α and β in $(0, 1)$ such that $\alpha + \beta \leq 0.5$, for k in $\{1, \dots, n\}$,

$$m\mathrm{FWSR}_{d_2}^{\alpha,\beta}(\mathcal{P}_k) \geq \sigma\left(k\ln\left(1 + \frac{n}{k^2} \vee \sqrt{\frac{n}{k^2}}\right)\right)^{1/2}. \tag{5.17}$$

Considering the above p-values p_i for i in $\{1,\dots,n\}$, we introduce the multiple test:

$$\bar{\mathcal{R}} = \{H_{\bar{S}_i},\ \min_{j\le i} p_j \le \alpha/n\}, \tag{5.18}$$

which is a particular basic case of the variant of the closure method of [MPG76] introduced by Romano and Wolf in [RW05, Algorithm 1 (idealized step-down method)] and [RW05, Theorem 1], when critical values satisfy a monotonicity assumption. Then this test has a FWER controlled by α and for any k in $\{1,\dots,n\}$, β in $(0, 0.5)$,

$$\mathrm{FWSR}^{\beta}_{d_2}\left(\bar{\mathcal{R}}, \mathcal{P}_k\right) \le \sigma\sqrt{k}\left(\sqrt{2\ln(n/\alpha)} + \sqrt{-2\ln(2\beta)}\right).$$

For k proportional to n^γ with $\gamma \in [0, 1/2)$, this upper bound coincides with the lower bound obtained in (5.17), up to some constant. Therefore, in this case, $m\mathrm{FWSR}^{\alpha,\beta}_{d_2}(\mathcal{P}_k)$ is of order $\sigma(n^\gamma \ln n)^{1/2}$, and the multiple test $\bar{\mathcal{R}}$ is minimax adaptive over the considered classes. Notice moreover that there is again here no price to pay for adaptivity.

From these results, we deduce that some classical multiple testing procedures are optimal in the present minimax sense. As some of these procedures, such as the Bonferroni, Holm, and the above basic variant of the closure method introduced by Romano and Wolf [RW05], are in fact a priori not expected to give optimality from a second kind error point of view, this may be a bit disturbing. We guess that the loss in Family-Wise Separation Rates of such basic procedures is hidden in multiplicative constants and that this loss would probably become more visible if the Gaussian vector $\mathbf{X}$ had a strong dependence structure. This consideration motivated the next study, where it is shown that Bonferroni procedures are not always optimal and can be outperformed by optimal min-p procedures in the minimax sense.

Gaussian regression model with strong dependency. As the gap in FWER between one-step procedures such as Bonferroni ones, and step-down procedures such as min-p ones, is usually more perceptible when the considered p-values are dependent, we here follow the same idea, and introduce a somewhat artificial, but nevertheless determinative, dependent Gaussian regression framework. The chosen dependence structure is quite extreme, so that lower bounds for minimax Family-Wise Separation Rates can be easily deduced as in the classical minimax theory for single hypothesis tests.
Let τ be a partition of $\{1,\dots,n\}$. Let the observed random variable be a Gaussian random vector $\mathbf{X} = (X_1,\dots,X_n)'$ distributed as in the following model.

$\mathcal{M}^{(1)}_{\text{dep. reg.}}$	$\mathbf{X} = \mathbf{X}_n = (X_1,\dots,X_n)'$ has a distribution $P = P_{f,\tau}$ defined by: $X_i = f_i + \sigma\varepsilon_t$ for every t in τ and every i in t, $f = (f_1,\dots,f_n)'$ being an unknown real vector, the ε_t's (t in τ) being independent standard Gaussian random variables, and σ being a known positive real number.

We consider the collection of hypotheses $\mathcal{H} = \{H_{S_i},\ i = 1, ..., n\}$, and for any $i = 1\dots n$, the same above p-value p_i.
Let T, k in $\{1,\dots,n\}$ and

$$\mathcal{P}_{k,T} = \{P_{f,\tau},\ f \in \mathbb{R}^n,\ |f|_0 \le k \text{ and } \#\tau = T\}.$$

We prove (see [14, Propositions 10 and 11]) that for α, β in $(0, 1)$, $m\mathrm{FWSR}^{\alpha,\beta}_{d_\infty}(\mathcal{P}_{k,T})$ is of order $\sigma(\ln T)^{1/2}$, and that in particular the min-p procedure $\mathcal{R}_{mp}$ associated with the p-values p_i is minimax adaptive over the collection of classes $\mathcal{P}_{k,T}$.
By contrast, it is also proved (see [14, Proposition 12]) that for n large enough, the Bonferroni procedure $\mathcal{R}_{Bonf}$ based on the same p-values has a $\mathrm{FWSR}^{\beta}_{d_\infty}$ over $\mathcal{P}_{k,T}$ lower bounded by $\sigma(\ln n)^{1/2}$, up to a multiplicative constant, and therefore can not be minimax over $\mathcal{P}_{k,T}$ as soon as $\ln T << \ln n$.

5.4 Perspectives

The purpose of the theoretical work presented in this chapter was to lay some foundations of a minimax theory for multiple testing, and in this sense, it has to be viewed as only a starting point for future studies of multiple tests from the minimax point of view.
Lots of emerging issues remain unsolved, encouraging us to pursue this path.
We have proved that the present theory may legitimate one-step and step-down procedures, such as the Bonferroni, Holm or min-p ones for simple multiple testing problems in a very basic Gaussian regression model, where p-values are clearly independent. Our results, and in particular the lower bounds for the minimax Family-Wise Separation Rates, were obtained using classical tools and results from the existing minimax theory for single hypothesis tests. We then have considered another Gaussian regression model, where p-values are roughly dependent, where the Bonferroni procedure is suboptimal from the minimax point of view, contrary to the min-p procedure which is proved to be minimax adaptive. The present strong dependence structure enables us to use again known results in the classical minimax theory for single hypothesis tests.
Studying some multiple testing problems in other frameworks, typically involving more reasonable dependence structures, will be challenging, all the more as very few works deal with minimax single testing in models suffering from dependency. Considering more complex classes of alternatives than the ones introduced here is also an interesting matter. New approaches and tools to establish lower bounds for the minimax Family-Wise Separation Rates will have to be developed.
All this will probably allow to validate already existing sophisticated multiple tests from the second kind error angle, but will also make necessary the construction of new optimal multiple tests. A paper dealing with these issues is in progress with Nicolas Verzelen.
Questions and problems known to appear in high dimension, which is inherent to many multiple testing problems, will also have to be investigated within the present minimax theory.
Finally, and this is actually closely related to the above question of high dimension, extending the criteria developed here, which are exclusively dedicated to multiple tests controlling the FWER, to multiple tests controlling the False Discovery Rate would be a major progress. It seems to be definitely more difficult, as no parallel between multiple tests controlling the FDR and aggregated tests can be established as clearly as in the present work.

Author's bibliography

[1] M. Fromont. *Quelques problèmes de sélection de modèles : Construction de tests adaptatifs, ajustement de pénalités par des méthodes de bootstrap*. PhD thesis, Université Paris Sud-Paris XI, 2003.

[2] M. Fromont. Model selection by bootstrap penalization for classification. In *Learning Theory: 17th Annual Conference On Learning Theory, COLT 2004*, pages 285–299. Springer, 2004.

[3] M. Fromont and C. Tuleau. Functional classification with margin conditions. In *Learning Theory: 19th Annual Conference On Learning Theory, COLT 2006*, pages 94–108. Springer, 2006.

[4] M. Fromont and C. Lévy-Leduc. Adaptive tests for periodic signal detection with applications to laser vibrometry. *ESAIM: Probability and Statistics*, 10:46–75, 2006.

[5] M. Fromont and B. Laurent. Adaptive goodness-of-fit tests in a density model. *The Annals of Statistics*, 34(2):680–720, 2006.

[6] M. Fromont. Model selection by bootstrap penalization for classification. *Machine Learning*, 66(2-3):165–207, 2007.

[7] M. Fromont, B. Laurent, and P. Reynaud-Bouret. Adaptive tests of homogeneity for a Poisson process. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 47(1):176–213, 2011.

[8] J. Bessac, F. Coquet, J.-M. Floch, and M. Fromont. Non-parametric tests for Poisson processes: studies on spatial representativeness of services. In *Actes des Journées de Méthodologie Statistique de l'INSEE*, 2012.

[9] M. Fromont, B. Laurent, M. Lerasle, and P. Reynaud-Bouret. Kernels based tests with non-asymptotic bootstrap approaches for two-sample problems. In *Journal of Machine Learning Research: Workshop and Conference Proceedings, 25th Annual Conference On Learning Theory, COLT 2012*, volume 23, pages 1–22, 2012.

[10] M. Fromont, B. Laurent, and P. Reynaud-Bouret. The two-sample problem for Poisson processes: Adaptive tests with a non-asymptotic wild bootstrap approach. *The Annals of Statistics*, 41(3):1431–1461, 2013.

[11] M. Fromont, B. Laurent, and P. Reynaud-Bouret. Supplement to « The two-sample problem for Poisson processes: Adaptive tests with a non-asymptotic wild bootstrap approach ». *The Annals of Statistics*, 2013.

[12] M. Albert, Y. Bouret, M. Fromont, and P. Reynaud-Bouret. Bootstrap and permutation tests of independence for point processes. *The Annals of Statistics*, 43(6):2537–2564, 2015.

[13] M. Albert, Y. Bouret, M. Fromont, and P. Reynaud-Bouret. Supplement to « Bootstrap and permutation tests of independence for point processes ». *The Annals of Statistics*, 2015.

[14] M. Fromont, M. Lerasle, and P. Reynaud-Bouret. Family-Wise Separation Rates for multiple testing. *The Annals of Statistics*, 2015. To appear.

[15] M. Albert, Y. Bouret, M. Fromont, and P. Reynaud-Bouret. Surrogate data methods based on a shuffling of the trials for synchrony detection: the centering issue. Submitted manuscript, arXiv:1505.06129, 2015.

[16] M. Fromont and C. Tuleau-Malot. Aggregation of Nearest Neighbors-based tests for the two-sample problem. Submitted manuscript, 2015.

General bibliography

[AB11] S. Arlot and P. L. Bartlett. Margin-adaptive model selection in statistical learning. *Bernoulli*, 17(2):687–713, 2011.

[ABR10] S. Arlot, G. Blanchard, and E. Roquain. Some nonasymptotic results on resampling in high dimension, I: Confidence regions and II: Multiple tests. *The Annals of Statistics*, 38(1):51–82, 83–99, 2010.

[ACV14] E. Arias-Castro and N. Verzelen. Community detection in dense random networks. *The Annals of Statistics*, 42(3):940–969, 2014.

[AF13] J.-M. Azaïs and J.-C. Fort. Remark on the finite-dimensional character of certain results of functional statistics. *Comptes Rendus de l'Académie des Sciences, Mathématiques*, 351(3):139–141, 2013.

[AG92] M. A. Arcones and E. Giné. On the bootstrap of U and V statistics. *The Annals of Statistics*, 20(2):655–674, 1992.

[AG93] M. A. Arcones and E. Giné. Limit theorems for U-processes. *The Annals of Probability*, 21(3):1494–1542, 1993.

[AL15] S. Arlot and M. Lerasle. Choice of V for V-fold Cross-Validation in least-squares density estimation. *The Journal of Machine Learning Research*, 2015. To appear.

[AM09] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *The Journal of Machine Learning Research*, 10:245–279, 2009.

[Arl07] S. Arlot. *Rééchantillonnage et sélection de modèles*. PhD thesis, Université Paris Sud-Paris XI, 2007.

[Arl09] S. Arlot. Model selection by resampling penalization. *Electronic Journal of Statistics*, 3:557–624, 2009.

[AT07] J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

[Ath87] K. B. Athreya. Bootstrap of the mean in the infinite variance case. *The Annals of Statistics*, 15(2):724–731, 1987.

[Bah96] R. Bahr. *Ein neuer test fuer das mehrdimensionale zwei-stichproben-problem*. PhD thesis, University of Hanover, 1996.

[Bar91] A. R. Barron. Complexity regularization with application to artificial neural networks. *Nonparametric functional estimation and related topics*, C 335:561–576, 1991.

[Bar02] Y. Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606, 2002.

[BB95] P. Barbe and P. Bertail. *The weighted bootstrap*. Springer-Verlag New York, 1995.

[BBL02] P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48(1-3):85–113, 2002.

[BBM99] A. R. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.

[BBM05] P. L. Bartlett, O. Bousquet, and S. Mendelson. Localized Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[BBW05] G. Biau, F. Bunea, and M. Wegkamp. Functional classification in Hilbert spaces. *IEEE Trans. Inf. Theory*, 51(6):2163–2172, 2005.

[BCG10] G. Biau, F. Cérou, and A. Guyader. On the rate of convergence of the bagged nearest neighbor estimate. *The Journal of Machine Learning Research*, 11:687–712, 2010.

[BD13] P. J. Brockwell and R. A. Davis. *Time series: theory and methods*. Springer-Verlag New York, 2013.

[BDNN04] M. Bhattacharjee, J. V. Deshpande, and U. V. Naik-Nimbalkar. Unconditional tests of goodness of fit for the intensity of time-truncated nonhomogeneous Poisson processes. *Technometrics*, 46(3):330–338, 2004.

[BEW85] L. J. Bain, M. Engelhardt, and F. T. Wright. Tests for an increasing trend in the intensity of a Poisson process: a power study. *Journal of the American Statistical Association*, 80(390):419–422, 1985.

[BF81] P. J. Bickel and D. A. Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6):1196–1217, 1981.

[BF04] L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206, 2004.

[BH95] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 57(1):289–300, 1995.

[BHL03] Y. Baraud, S. Huet, and B. Laurent. Adaptive tests of linear hypotheses by model selection. *The Annals of Statistics*, 31(1):225–251, 2003.

[BI13] C. Butucea and Yu. I. Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688, 2013.

[Bil09] P. Billingsley. *Convergence of probability measures*. Wiley-Interscience, 2009.

[BKR61] J. R. Blum, J. Kiefer, and M. Rosenblatt. Distribution free tests of independence based on the sample distribution function. *The Annals of Mathematical Statistics*, 32(2):485–498, 1961.

[BLM99] D. Bitouzé, B. Laurent, and P. Massart. A Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 35(6):735–763, 1999.

[BM97] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics (D. Pollard, E. Torgersen, and G. Yang, eds.)*, pages 55–87, 1997.

[BM98] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

[BM03] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

[BMP04] P. L. Bartlett, S. Mendelson, and P. Philips. Local complexities for Empirical Risk Minimization. In *Learning Theory: 17th Annual Conference On Learning Theory, COLT 2004*, pages 270–284. Springer, 2004.

[BMP09] C. Butucea, C. Matias, and C. Pouet. Adaptive goodness-of-fit testing from indirect observations. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 45(2):352–372, 2009.

[BR92] P. J. Bickel and Y. Ritov. Testing for goodness of fit: a new approach. *Nonparametric Statistics and Related Topics*, pages 51–57, 1992.

[Bre83] J. Bretagnolle. Lois limites du bootstrap de certaines fonctionnelles. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 19(3):281–296, 1983.

[BS80] J. M. Bovett and J. G. Saw. On comparing two Poisson intensity functions. *Communications in Statistics. Theory and Methods*, 9(9):943–948, 1980.

[BT06] C. Butucea and K. Tribouley. Nonparametric homogeneity tests. *Journal of Statistical Planning and Inference*, 136(3):597–639, 2006.

[BTA04] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US, 2004.

[But07] C. Butucea. Goodness-of-fit testing and quadratic functional estimation from indirect observations. *The Annals of Statistics*, 35(5):1907–1930, 2007.

[BY01] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[Cas00] G. Castellan. Sélection d'histogrammes à l'aide d'un critère de type Akaike. *Comptes Rendus de l'Académie des Sciences, Mathématiques*, 330(8):729–732, 2000.

[Cas03] G. Castellan. Density estimation via exponential model selection. *Information Theory, IEEE Transactions on Information Theory*, 49(8):2052–2060, 2003.

[CD12] O. Collier and A. Dalalyan. Minimax hypothesis testing for curve registration. *Electronic Journal of Statistics*, 6:1129–1154, 2012.

[CH67] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[CLLM06] I. Castillo, C. Lévy-Leduc, and C. Matias. Exact adaptive estimation of the shape of a periodic function with unknown period corrupted by white noise. *Mathematical Methods of Statistics*, 15(2):146–175, 2006.

[CS93] A. Cohen and H. B. Sackrowitz. Evaluating tests for increasing intensity of a Poisson process. *Technometrics*, 35(4):446–448, 1993.

[DGKL94] L. Devroye, L. Gyorfi, A. Krzyzak, and G. Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics*, 22(3):1371–1385, 1994.

[DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Springer-Verlag New York, 1996.

[DH12] A. Delaigle and P. Hall. Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society, Series B*, 74(2):267–286, 2012.

[DJ98] D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879–921, 1998.

[DK06] S. Dachian and Yu. A. Kutoyants. Hypotheses testing: Poisson versus self-exciting. *Scandinavian Journal of Statistics*, 33(2):391–408, 2006.

[DL93] R. A. DeVore and G. G. Lorentz. *Constructive approximation*. Springer-Verlag Berlin Heidelberg, 1993.

[DL95] L. Devroye and G. Lugosi. Lower bounds in pattern recognition and learning. *Pattern Recognition*, 28:1011–1018, 1995.

[dlPG99] V. H. de la Peña and E. Giné. *Decoupling. From dependence to independence*. Springer-Verlag New York, 1999.

[DM94] H. Dehling and T. Mikosch. Random quadratic forms and the bootstrap for U-statistics. *Journal of Multivariate Analysis*, 51(2):392–413, 1994.

[DMNN99] J. V. Deshpande, M. Mukhopadhyay, and U. V. Naik-Nimbalkar. Testing of two sample proportional intensity assumption for non-homogeneous Poisson processes. *Journal of Statistical Planning and Inference*, 81(2):237–251, 1999.

[DR06] C. Durot and Y. Rozenholc. An adaptive test for zero mean. *Mathematical Methods of Statistics*, 15(1):26–60, 2006.

[Dud02] R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

[DVDL07] S. Dudoit and M. J. Van Der Laan. *Multiple testing procedures with applications to genomics*. Springer-Verlag New York, 2007.

[DVJ08] D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes. Vol. II. General theory and structure*. Springer-Verlag New York, 2008.

[DW76] L. Devroye and T. Wagner. Nonparametric discrimination and density estimation. Technical Report 183, Electronics Research Center, University of Texas, 1976.

[Efr79] B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[Faz07] K. Fazli. Second-order efficient test for inhomogeneous Poisson processes. *Statistical inference for stochastic processes*, 10(2):181–208, 2007.

[FH07] J. Franke and S. Halim. Wild bootstrap tests. *Signal Processing Magazine, IEEE*, 24(4):31–37, 2007.

[Fis35] R. A. Fisher. *The design of experiments*. 1935.

[FK05] K. Fazli and Yu. A. Kutoyants. Two simple hypotheses testing for Poisson process. *Far East Journal of Theoretical Statistics*, 15(2):251, 2005.

[FV06] F. Ferraty and P. Vieu. *Nonparametric functional data analysis. Theory and Practice*. Springer-Verlag New York, 2006.

[GBR+08] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample problem. *The Journal of Machine Learning Research*, 1:1–10, 2008.

[GBR+12] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13:723–773, 2012.

[GDA10] S. Grün, M. Diesmann, and A. M. Aertsen. *Analysis of parallel spike trains*, chapter Unitary Events analysis. Springer Series in Computational Neuroscience, 2010.

[GDG+99] S. Grün, M. Diesmann, F. Grammont, A. Riehle, and A. M. Aertsen. Detecting Unitary Events without discretization of time. *Journal of Neuroscience Methods*, 93:67–79, 1999.

[GFHS10] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems 22 (NIPS 2009)*, pages 673–681, 2010.

[GG10] A. Gretton and L. Györfi. Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11:1391–1423, 2010.

[Gin75] E. Giné. Invariant tests for uniformity on compact Riemannian manifolds based on Sobolev norms. *The Annals of Statistics*, 3(6):1243–1266, 1975.

[GKKW02] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer-Verlag New York, 2002.

[GL11] A. Goldenshluger and O. V. Lepski. Bandwith selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608–1632, 2011.

[GN15] E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press, 2015.

[GP69] G. L. Gerstein and D. H. Perkel. Simultaneous recorded trains of action potentials: analysis and functional interpretation. *Science*, 164:828–830, 1969.

[GP01] G. Gayraud and C. Pouet. Minimax testing composite null hypotheses in the discrete regression scheme. *Mathematical Methods of Statistics*, 10(4):375–394, 2001.

[GP05] G. Gayraud and C. Pouet. Adaptive minimax testing in the discrete regression scheme. *Probability Theory and Related Fields*, 133(4):531–558, 2005.

[GR03] F. Grammont and A. Riehle. Spike synchronisation and firing rate in a population of motor cortical neurons in relation to movement direction and reaction time. *Biological Cybernetics*, 88:360–373, 2003.

[Grü96] S. Grün. *Unitary joint-events in multiple-neuron spiking activity: Detection, significance and interpretation.* PhD thesis, Thun: Verlag Harri Deutsch, 1996.

[GS05] G. Gusto and S. Schbath. FADO: a statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes' model. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

[GS10] J. J. Goeman and A. Solari. The sequential rejection principle of familywise error control. *The Annals of Statistics*, 38(6):3782–3810, 2010.

[GW04] C. Genovese and L. Wasserman. A stochastic process approach to false discovery control. *The Annals of Statistics*, 32(3):1035–1061, 2004.

[GZ84] E. Giné and J. Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, 12(4):929–998, 1984.

[GZ89] E. Giné and J. Zinn. Necessary conditions for the bootstrap of the mean. *The Annals of Statistics*, 17(2):684–691, 1989.

[GZ90] E. Giné and J. Zinn. Bootstrapping general empirical measures. *The Annals of Probability*, 18(2):851–869, 1990.

[Háj61] J. Hájek. Some extensions of the Wald-Wolfowitz-Noether theorem. *The Annals of Mathematical Statistics*, 32(2):506–523, 1961.

[Hau95] D. Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory*, A 69(2):217–232, 1995.

[HC78] S. T. Ho and L. H. Y. Chen. An L_p bound for the remainder in a combinatorial central limit theorem. *The Annals of Probability*, 6(2):231–249, 1978.

[Hen88] N. Henze. A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, 16(2):772–783, 1988.

[HH88] G. Hommel and T. Hoffmann. Controlled uncertainty. In *Multiple Hypothesenprüfung/Multiple Hypotheses Testing*, pages 154–161. Springer-Verlag Berlin Heidelberg, 1988.

[HH90] P. Hall and J. D. Hart. Bootstrap test for difference between means in nonparametric regression. *Journal of the American Statistical Association*, 85(412):1039–1049, 1990.

[HJ93] M. Hušková and P. Janssen. Consistency of the generalized bootstrap for degenerate U-statistics. *The Annals of Statistics*, 21(4):1811–1823, 1993.

[HK99] W. Härdle and A. Kneip. Testing a regression model when we have smooth alternatives in mind. *Scandinavian Journal of Statistics*, 26(2):221–238, 1999.

[HLW94] D. Haussler, N. Littlestone, and M. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

[HMLW02] N. Hengartner, E. Matzner-Lober, and M. Wegkamp. Bandwidth selection for local linear regression. *Journal of the Royal Statistical Society, Series B*, 64:1–14, 2002.

[Hoe48a] W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

[Hoe48b] W. Hoeffding. A non-parametric test of independence. *The Annals of Mathematical Statistics*, 19(4):546–557, 1948.

[Hoe52] W. Hoeffding. The large-sample power of tests based on permutation of the observations. *The Annals of Mathematical Statistics*, 23(2):169–192, 1952.

[Hoe61] W. Hoeffding. The strong law of large numbers for U-statistics. Institute of Statistics, Mimeograph Series No. 302, 1961.

[Hol79] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[HP36] H. Hotelling and M. R. Pabst. Rank correlation and tests of significance involving no assumption of normality. *The Annals of Mathematical Statistics*, 7(1):29–43, 1936.

[HPS08] P. Hall, B. U. Park, and R. J. Samworth. Choice of neighbor order in nearest-neighbor classification. *The Annals of Statistics*, pages 2135–2152, 2008.

[HRB03] C. Houdré and P. Reynaud-Bouret. Exponential inequalities, with constants, for U-statistics of order two. In *Stochastic inequalities and applications*, pages 55–69. Springer Birkhäuser Basel, 2003.

[HS01] J. L. Horowitz and V. G. Spokoiny. An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica*, 69(3):599–631, 2001.

[IK07] Yu. I. Ingster and Yu. A. Kutoyants. Nonparametric hypothesis testing for intensity of the Poisson process. *Mathematical Methods of Statistics*, 16(3):217–245, 2007.

[IKL97] T. Inglot, W. C. Kallenberg, and T. Ledwina. Data driven smooth tests for composite hypotheses. *The Annals of Statistics*, 25(3):1222–1250, 1997.

[Ing82] Yu. I. Ingster. Minimax nonparametric detection of signals in white Gaussian noise. *Problems of Information Transmission*, 18(2):130–140, 1982.

[Ing84] Yu. I. Ingster. Asymptotic minimax testing of nonparametric hypothesis on the distribution density of an independent sample. *Zapiski Nauchn. Seminar. Leningrad Otdel. Mat. Inst. Steklov*, 136:74–96, 1984.

[Ing93] Yu. I. Ingster. Asymptotically minimax testing for nonparametric alternatives I-II-III. *Mathematical Methods of Statististics*, 2:85–114, 171–189, 249–268, 1993.

[Ing00] Yu. I. Ingster. Adaptive chi-square tests. *Journal of the American Statistical Association*, 89(427):1000–1005, 2000.

[IS11] Yu. I. Ingster and N. Stepanova. Estimation and detection of functions from anisotropic Sobolev classes. *Electronic Journal of Statistics*, 5:484–506, 2011.

[ITV10] Yu. I. Ingster, A. B. Tsybakov, and N. Verzelen. Detection boundary in sparse regression. *Electronic Journal of Statistics*, 4:1476–1526, 2010.

[Jan94] A. Janssen. Two-sample goodness-of-fit tests when ties are present. *Journal of Statistical Planning and Inference*, 39(3):399–424, 1994.

[Ken38] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1–2):81–93, 1938.

[KHW91] E. King, J. D. Hart, and T. E. Wehrly. Testing the equality of two regression curves using linear smoothers. *Statistics & Probability Letters*, 12(3):239–247, 1991.

[KK06] M. Kohler and A. Krzyżak. Rate of convergence of local averaging plug-in classification rules under margin condition. In *2006 IEEE International Symposium on Information Theory*, pages 2176–2179. IEEE, 2006.

[KL95] W. C. Kallenberg and T. Ledwina. Consistency and Monte Carlo simulation of a data driven version of smooth goodness-of-fit tests. *The Annals of Statistics*, 23(5):1594–1608, 1995.

[Kol81] V. Koltchinskii. On the central limit theorem for empirical measures. *Theory of Probability and Mathematical Statistics*, 24:71–82, 1981.

[Kol01] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

[Kol06] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

[KP99] V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–459. Birkhäuser Boston, 1999.

[KP00] G. Kerkyacharian and D. Picard. Thresholding algorithms, maxisets and well-concentrated bases. *Test*, 9(2):283–344, 2000.

[KRPA+09] B. E. Kilavik, S. Roux, A. Ponce-Alvarez, J. Confais, S. Grün, and A. Riehle. The control of the false discovery rate in multiple testing under dependency. *Journal of Neuroscience*, 29(40):12653–12663, 2009.

[KTMS04] E. L. Korn, J. F. Troendle, L. M. McShane, and R. Simon. Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, 124(2):379–398, 2004.

[KW92] C. Klaassen and J. A. Wellner. KAC empirical processes and the bootstrap. In *Probability in Banach spaces, 8*, pages 411–429. Birkhäuser Boston, 1992.

[Lat99] R. Latała. Tail and moment estimates for some types of chaos. *Studia Mathematica*, 135(1):39–53, 1999.

[Lau05] B. Laurent. Adaptive estimation of a quadratic functional of a density by model selection. *ESAIM: Probability and Statistics*, 9:1–18, 2005.

[Lec07a] G. Lecué. Simultaneous adaptation to the margin and to complexity in classification. *The Annals of Statistics*, 35(4):1698–1721, 2007.

[Lec07b] G. Lecué. Suboptimality of penalized empirical risk minimization in classification. In *Learning Theory: 20th Annual Conference On Learning Theory, COLT 2007*, pages 142–156. Springer, 2007.

[Ler12] M. Lerasle. Optimal model selection in density estimation. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(3):884–908, 2012.

[LLM] B. Laurent, J.-M. Loubes, and C. Marteau. Non asymptotic minimax rates of testing in signal detection with heterogeneous variances. `http://arxiv.org/abs/0912.2423v1`.

[LLM12] B. Laurent, J.-M. Loubes, and C. Marteau. Non asymptotic minimax rates of testing in signal detection with heterogeneous variances. *Electronic Journal of Statistics*, 6:91–122, 2012.

[LM00] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

[LN09] A. Leucht and M. H. Neumann. Consistency of general bootstrap methods for degenerate U-type and V-type statistics. *Journal of Multivariate Analysis*, 100(8):1622–1633, 2009.

[LN11] K. Lounici and R. Nickl. Global uniform risk bounds for wavelet deconvolution estimators. *The Annals of Statistics*, 39(1):201–231, 2011.

[Lo87] A. Y. Lo. A large sample study of the Bayesian bootstrap. *The Annals of Statistics*, 15(1):360–375, 1987.

[Loz00] F. Lozano. Model selection using Rademacher penalization. In *Proceedings of the 2nd ICSC Symp. on Neural Computation (NC2000). Berlin, Germany.* ICSC Academic Press, 2000.

[LR05] E. L. Lehmann and J. P. Romano. Generalizations of the familywise error rate. *The Annals of Statistics*, 33(3):1138–1154, 2005.

[LRS05] E. L. Lehmann, J. P. Romano, and J. P. Shaffer. On optimality of stepdown and stepup multiple test procedures. *The Annals of Statistics*, 33(3):1084–1108, 2005.

[LS99] O. V. Lepski and V. G. Spokoiny. Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative. *Bernoulli*, 5(2):333–358, 1999.

[LT00] O. V. Lepski and A. B. Tsybakov. Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point. *Probability Theory and Related Fields*, 117(1):17–48, 2000.

[Lug02] G. Lugosi. Pattern classification and learning theory. In *Principles of Nonparametric Learning*, pages 1–56. Springer Wien, 2002.

[LW04] G. Lugosi and M. Wegkamp. Complexity regularization via localized random penalties. *The Annals of Statistics*, 32(4):1679–1697, 2004.

[LZ96] G. Lugosi and K. Zeger. Concept learning using complexity regularization. *IEEE Transactions on Information Theory*, 42(1):48–54, 1996.

[Mam92] E. Mammen. Bootstrap, wild bootstrap, and asymptotic normality. *Probability Theory and Related Fields*, 93(4):439–455, 1992.

[Mas07] P. Massart. *Concentration inequalities and model selection.* Springer-Verlag Berlin Heidelberg, 2007.

[McD89] C. McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[MN06] P. Massart and É. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

[Mot56] M. Motoo. On the Hoeffding's combinatorial central limit theorem. *Annals of the Institute of Statistical Mathematics*, 8(1):145–154, 1956.

[MPG76] R. Marcus, E. Peritz, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

[MT99] E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

[Ney37] J. Neyman. « Smooth test » for goodness of fit. *Scandinavian Actuarial Journal*, 1937(3-4):149–199, 1937.

[PC09] C. Pouzat and A. Chaffiol. Automatic spike train analysis and report generation. An implementation with R, R2HTML and STAR. *Journal of Neuroscience Methods*, 2009.

[PDG03] G. Pipa, M. Diesmann, and S. Grün. Significance of joint-spike events based on trial-shuffling by efficient combinatorial methods. *Complexity*, 8(4):1–8, 2003.

[Pea00] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[Pea11] K. Pearson. On the probability that two independent distributions of frequency are really samples from the same population. *Biometrika*, 8:250–254, 1911.

[PG03] G. Pipa and S. Grün. Non-parametric significance estimation of joint-spike events by shuffling and resampling. *Neurocomputing*, 52–54:31–37, 2003.

[Poi37] S. D. Poisson. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des Probabilités*. Bachelier, Paris, 1837.

[Pol82] D. Pollard. A central limit theorem for empirical processes. *Journal of the Australian Mathematical Society*, A 33(2):235–248, 1982.

[Pou02] C. Pouet. Test asymptotiquement minimax pour une hypothèse nulle composite dans le modèle de densité. *Comptes Rendus de l'Académie des Sciences, Mathématiques*, 334(10):913–916, 2002.

[Præ95] J. T. Præstgaard. Permutation and bootstrap Kolmogorov-Smirnov tests for the equality of two distributions. *Scandinavian Journal of Statistics*, pages 305–322, 1995.

[PS10] F. Pesarin and L. Salmaso. *Permutation tests for complex data: Theory, Applications and Software*. Series in Probability and Statistics. Wiley, 2010.

[PW93] J. T. Præstgaard and J. A. Wellner. Exchangeably weighted bootstraps of the general empirical process. *The Annals of Probability*, 21(4):2053–2086, 1993.

[Que49] M. H. Quenouille. Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society, Series B*, 11(1):68–84, 1949.

[RBR10] P. Reynaud-Bouret and V. Rivoirard. Near optimal thresholding estimation of a Poisson intensity on the real line. *Electronic Journal of Statistics*, 4:172–238, 2010.

[RBRGTM14] P. Reynaud-Bouret, V. Rivoirard, F. Grammont, and C. Tuleau-Malot. Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis. *Journal of Mathematical Neuroscience*, 4(3), 2014.

[RGDG00] A. Riehle, F. Grammont, M. Diesmann, and S. Grün. Dynamical changes and temporal precision of synchronised spiking activity in monkey motor cortex during movement preparation. *Journal of Physiology*, 94:569–582, 2000.

[RGM06] A. Riehle, F. Grammont, and A. MacKay. Cancellation of a planned movement in monkey motor cortex. *Neuroreport*, 17(3):281–285, 2006.

[Riv02] V. Rivoirard. *Estimation bayésienne non paramétrique*. PhD thesis, Université Paris VII-Denis Diderot, 2002.

[Riv06] V. Rivoirard. Nonlinear estimation over weak Besov spaces and minimax Bayes method. *Bernoulli*, 12(4):609–632, 2006.

[Rom87] J. P. Romano. Bootstrap and randomization tests of some nonparametric hypotheses. Technical Report 270, Dept. Statistics, Stanford Univ., 1987.

[Rom89] J. P. Romano. Bootstrap and randomization tests of some nonparametric hypotheses. *The Annals of Statistics*, 17(1):141–159, 1989.

[RRS05] S. Robin, F. Rodolphe, and S. Schbath. *DNA, words and models: Statistics of exceptional words*. Cambridge University Press, 2005.

[RS02] J. Ramsay and B. Silverman. *Applied functional data analysis. Methods and case studies*. Springer-Verlag New York, 2002.

[RS05] J. Ramsay and B. Silverman. *Functional data analysis*. Springer-Verlag New York, 2005.

[RS06] J. P. Romano and A. M. Shaikh. Stepup procedures for control of generalizations of the familywise error rate. *The Annals of Statistics*, 34(4):1850–1873, 2006.

[RSW11] J. P. Romano, A. M. Shaikh, and M. Wolf. Consonance and the closure method in multiple testing. *The International Journal of Biostatistics*, 7(1):1–25, 2011.

[Rub81] D. B. Rubin. The Bayesian bootstrap. *The Annals of Statistics*, 9:130–134, 1981.

[RV80] H. Rubin and R. A. Vitale. Asymptotic distribution of symmetric statistics. *The Annals of Statistics*, pages 165–170, 1980.

[RV05] F. Rossi and N. Villa. Classification in Hilbert spaces with support vector machines. In *Proceedings of ASMDA 2005*, pages 635–642, 2005.

[RW05] J. P. Romano and M. Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108, 2005.

[RW07] J. P. Romano and M. Wolf. Control of generalized error rates in multiple testing. *The Annals of Statistics*, 35(4):1378–1408, 2007.

[RW10] J. P. Romano and M. Wolf. Balanced control of generalized error rates. *The Annals of Statistics*, 38(1):598–633, 2010.

[Sau] A. Saumard. Nonasymptotic quasi-optimality of AIC and the slope heuristics in maximum likelihood estimation of density using histogram models. hal-01107321.

[Sch86] M. F. Schilling. Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81:799–806, 1986.

[SFG+10] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. Lanckriet. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in Neural Information Processing Systems 22 (NIPS 2009)*, pages 1750–1758, 2010.

[SFL11] B. K. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *The Journal of Machine Learning Research*, 12:2389–2410, 2011.

[SGF+10] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. Lanckriet. Hilbert spaces embeddings and metrics on probability distributions. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

[Sim86] R. J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

[Sin81] K. Singh. On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics*, 9(6):1187–1195, 1981.

[Sin93] W. Singer. Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology*, 55:349–374, 1993.

[Spj72] E. Spjøtvoll. On the optimality of some multiple comparison procedures. *The Annals of Mathematical Statistics*, 43(2):398–411, 1972.

[Spo96] V. G. Spokoiny. Adaptive hypothesis testing using wavelets. *The Annals of Statistics*, 24(6):2477–2498, 1996.

[Spo98] V. G. Spokoiny. Adaptive and spatially adaptive testing of a nonparametric hypothesis. *Mathematical Methods of Statistics*, 7(3):245–273, 1998.

[SS02] B. Schölkopf and A. J. Smola. *Learning with kernels*. MIT Press, 2002.

[SSGF13] D. Sejdinovic, B. K. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

[ST95] J. Shao and D. Tu. *The Jackknife and Bootstrap*. Springer-Verlag New York, 1995.

[STM15] L. Sansonnet and C. Tuleau-Malot. A model of Poissonian interactions and detection of dependence. *Statistics and Computing*, 25(2):449–470, 2015.

[Sto77] C. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5:595–645, 1977.

[TMRGRB14] C. Tuleau-Malot, A. Rouis, F. Grammont, and P. Reynaud-Bouret. Multiple tests based on a Gaussian approximation of the Unitary Events method. *Neural Computation*, 26(7), 2014.

[Tri06] H. Triebel. *Theory of function spaces*. Birkhäuser Verlag, Basel, 2006.

[Tsy04] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[Tsy09] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer-Verlag New York, 2009.

[Tuk58] J. W. Tukey. Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics*, 29:614, 1958.

[Tul05] C. Tuleau. *Sélection de variables pour la discrimination en grande dimension et classification de données fonctionnelles*. PhD thesis, Université Paris Sud-Paris XI, 2005.

[Vap82] V. Vapnik. *Estimation of dependences based on empirical data*. Springer-Verlag New York, 1982.

[Var58] V. S. Varadarajan. On the convergence of sample probability distributions. *Sankhyà*, 19(1–2):23–26, 1958.

[VC71] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

[VC74] V. Vapnik and A. Chervonenkis. *Teoriya raspoznavaniya obrazov. Statisticheskie problemy obucheniya (Theory of pattern recognition. Statistical problems of learning)*. Nauka, Moscow, 1974.

[VdV00] A. W. Van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000.

[VdVW96] A. W. Van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag New York, 1996.

[VV10] N. Verzelen and F. Villers. Goodness-of-fit tests for high-dimensional gaussian linear models. *The Annals of Statistics*, 38(2):704–752, 2010.

[Wat78] G. S. Watson. Estimating the intensity of a Poisson process. *Applied time series analysis*, pages 325–345, 1978.

[Wel79] J. A. Wellner. Permutation tests for directional data. *The Annals of Statistics*, 7(5):929–943, 1979.

[Wen89] C.-S. Weng. On a second-order asymptotic property of the Bayesian bootstrap mean. *The Annals of Statistics*, 17(2):705–710, 1989.

[Wol42] J. Wolfowitz. Additive partition functions and a class of statistical hypotheses. *The Annals of Mathematical Statistics*, 13(3):247–279, 2042.

[WW40] A. Wald and J. Wolfowitz. On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11(2):147–162, 1940.