ESAIM: PS 16 (2012) 61–85
DOI: 10.1051/ps/2010012

ESAIM: Probability and Statistics
www.esaim-ps.org

ADAPTIVE NON-ASYMPTOTIC CONFIDENCE BALLS IN DENSITY ESTIMATION

Matthieu Lerasle[1]

Abstract. We build confidence balls for the common density s of a real valued sample $X_1,\dots,X_n$. We use resampling methods to estimate the projection of s onto finite dimensional linear spaces and a model selection procedure to choose an optimal approximation space. The covering property is ensured for all $n\geq 2$ and the balls are adaptive over a collection of linear spaces.

Mathematics Subject Classification. 62G07, 62G09, 62G10, 62G15.

Received November 14, 2008. Revised January 11, 2010.

1. Introduction

In this paper, we discuss the problem of adaptive confidence balls, from a non-asymptotic point of view, in the particular context of density estimation. Let S be a set of densities with respect to the Lebesgue measure μ on $\mathbb{R}$. Given an i.i.d sample $X_{1:n}=(X_1,\dots,X_n)$ and a confidence level $\beta\in(0,1)$, a confidence set (hereafter CS) $\hat{B}_\beta(X_{1:n})$ on S is a subset of S satisfying the following covering property:

$$\forall s\in S,\ \mathbb{P}_s\left(s\in\hat{B}_\beta(X_{1:n})\right)\geq 1-\beta \tag{1.1}$$

where, for all s in S, $\mathbb{P}_s$ denotes the distribution of $X_{1:n}$ when the marginals have common density s. All the CS considered in this paper are L^2-balls, centered on estimators $\hat{s}$ of s, and with random radius $\hat{\rho}_\beta$. The quality of a CS is measured with the quantiles of $\hat{\rho}_\beta$. We are looking for adaptive CS, which means that, given a collection $(S_m)_{m\in\mathcal{M}_n}$ of subsets of S, $\hat{\rho}_\beta$ should be as small as possible over all the sets $(S_m)_{m\in\mathcal{M}_n}$.

This problem was mostly considered in regression frameworks, see among others Li [25], Lepski [23], Juditski and Lepski [20], Hoffmann and Lepski [14], Juditski and Lambert-Lacroix [19], Baraud [4], Beran [5], Beran and Dümbgen [6], Cai and Low [9], Genovese and Wassermann [12,13]. Robins and van der Vaart [28] considered a more general Hilbertian framework that includes in particular density estimation and some regression frameworks.

Our adaptive balls are derived from a model selection procedure, which is essentially the one of Baraud [4]. We start with a collection of linear spaces $(S_m)_{m\in\mathcal{M}_n}$ and associate to each of these, the projection estimator $\hat{s}_m$ of s and some positive number $\hat{\rho}(m)$. The $\hat{\rho}(m)$'s are suitably calibrated to satisfy the property that, with probability close to one the distance between s and its projection estimator $\hat{s}_m$ is not larger than $\hat{\rho}(m)$. We

Keywords and phrases. Confidence balls, density estimation, resampling methods.

[1] Institut de Mathématiques (UMR 5219), INSA de Toulouse, Université de Toulouse, France. lerasle@gmail.com

Article published by EDP Sciences

then select $\hat{m}$ as the minimizer of $\hat{\rho}(m)$ and define the confidence ball as the L^2-ball centered at $\hat{s}_{\hat{m}}$ of radius $\hat{\rho}(\hat{m})$.

We use two different ingredients to compute $\hat{\rho}(m)$. The first one is a resampling estimator of $\|s_m - \hat{s}_m\|^2$, where s_m denotes the projection of s onto S_m. It is naturally derived from Efron's heuristic (see Efron [10]), in the same way as Arlot *et al.* [3]. This allows us in particular to keep all the sample to build $\hat{s}_m$. This is an improvement compared with Robins and van der Vaart [28] or Cai and Low [9], who cut the sample into two parts, the first one being used to build an estimator $\hat{s}$ of s and the other to evaluate the distance $\|\hat{s} - s\|^2$.

The second ingredient is an estimator of $\|s - s_m\|^2$, based on U-statistics, as in Laurent [21,22]. The proofs are handled thanks to a concentration inequality for U-statistics, derived from Houdré and Reynaud–Bouret [15]. The main advantage of a model selection's approach is that the resulting CS are non asymptotic, *i.e.* (1.1) holds for all n. Moreover, the CS behaves well even if s does not belong to S, which outperforms, in that case, the result of Li [25].

Let S be a linear space with dimension d and let $(S_m)_{m\in\mathcal{M}_n}$ be a collection of linear subspaces of S, with respective dimensions $(d_m)_{m\in\mathcal{M}_n}$. The diameter of our CS on S is upper bounded, for any s in S_m, by $C(\sqrt{d} \vee d_m)/n$, where C is a constant, free from d, d_m, and n. This bound is optimal in the minimax sense. Hence, adaptation is possible over collections of subspaces with dimension $d_m \geq \sqrt{d}$ for L^2-balls. This positive result does not hold in general, in particular, adaptation is impossible for L^∞-balls (Low [26]). However, the adaptation property is strongly limited since it is impossible over spaces with dimension $d_m \leq \sqrt{d}$. This negative result was already proved asymptotically in Li [25], Hoffmann and Lepski [14], Juditski and Lambert-Lacroix [19], Robins and van der Vaart [28]. It was proved non-asymptotically in a regression framework in Baraud [4]. We use the method of Baraud [4] and extend his result to the density estimation framework.

The paper is decomposed as follows. Section 2 introduces the notations and the main assumptions. Section 3 presents the technical tools required for the construction of our CS. Section 4 gives the main results, we build our CS, give upper bounds on their size and prove their optimality in the minimax sense. Section 5 presents a short simulation study, where we illustrate the behavior of our resampling-based estimators. All the proofs are postponed to Section 6. We add in an Appendix the proofs of some technical lemmas.

2. Notations and assumptions

2.1. Notations

Hereafter, $L^2(\mu)$ denotes the space of all measurable functions $t:\mathbb{R}\to\mathbb{R}$ such that $\int_{\mathbb{R}} t^2(x)\mathrm{d}\mu(x) < \infty$. It is endowed by its classical scalar product defined, for all t, t' in $L^2(\mu)$ by $\langle t, t'\rangle = \int_{\mathbb{R}} t(x)t'(x)\mathrm{d}\mu(x)$ and by the associated L^2-norm defined, for t in $L^2(\mu)$ by $\|t\| = \sqrt{\langle t, t\rangle}$.

For any density s, we denote by $\mathbb{P}_s$ the distribution of an iid sample $X_{1:n} = (X_1, \ldots, X_n)$ with common marginal density s and by $\mathbb{E}_s$ the expectation with respect to $\mathbb{P}_s$.

Hereafter, S, with various subscripts, denotes a linear subspace of $L^2(\mu)$ and S^* the set of densities in S. For all sets $\mathcal{F}$ in $L^2(\mu)$, the L^2-diameter of $\mathcal{F}$ is defined by

$$\Delta(\mathcal{F}) = \sup_{(t,t')\in\mathcal{F}^2} \|t - t'\|.$$

For a random set $\hat{B}$ in $L^2(\mu)$, a linear space S of measurable functions and a real number α in $(0,1)$, we define the (S,α)-size of $\hat{B}$ as

$$\Delta_{(S,\alpha)}(\hat{B}) = \inf\left\{\delta > 0,\ \sup_{s\in S^*} \mathbb{P}_s(\Delta(\hat{B}) > \delta) \leq \alpha\right\}. \tag{2.1}$$

For all indexes sets Λ, $(\psi_\lambda)_{\lambda\in\Lambda}$ will always denote an orthonormal system in $L^2(\mu)$.

2.2. Efron's resampling heuristic

Let $X, X_1, \ldots, X_n$ be i.i.d random variables with common density s, let P_s and P_n denote the following processes defined respectively for all functions t in $L^2(\mu)$ and for all measurable functions t by

$$P_s t = \langle s, t \rangle = \int_{\mathbb{R}} t(x)s(x)\mathrm{d}\mu(x) = \mathbb{E}(t(X)),\ P_n t = \frac{1}{n}\sum_{i=1}^{n} t(X_i).$$

Hereafter, a resampling scheme $(W_1, \ldots, W_n)$ is a vector of real valued random variables, independent of $(X_1, \ldots, X_n)$ and exchangeable, which means that, for all permutations τ of $1, \ldots, n$,

$$(W_{\tau(1)}, \ldots, W_{\tau(n)}) \text{ has the same law as } (W_1, \ldots, W_n).$$

Let $(W_1, \ldots, W_n)$ be a resampling scheme, let $\bar{W}_n = \sum_{i=1}^{n} W_i/n$ and let P_n^W denotes the resampling-based empirical process defined, for all measurable functions t, by

$$P_n^W t = \frac{1}{n}\sum_{i=1}^{n} W_i t(X_i).$$

For all random variables $F(X_1, \ldots, X_n, W_1, \ldots, W_n)$, we denote by

$$\mathbb{E}_W \left(F(X_1, \ldots, X_n, W_1, \ldots, W_n)\right) = \mathbb{E}\left(F(X_1, \ldots, X_n, W_1, \ldots, W_n) | X_1, \ldots, X_n\right).$$

Let F be a known functional and $F_n = F(P_n, P_s)$, we define the resampling estimator of F_n by

$$F_n^W = C_W \mathbb{E}_W \left(F(P_n^W, \bar{W}_n P_n)\right),$$

where C_W is a constant depending only on the functional F and the law of the resampling scheme. Efron's heuristics states that F_n^W provides a sharp estimator of F_n when the constant C_W is well chosen.

2.3. Balls in functional spaces

Our method is strongly based on empirical process methods, in particular on Talagrand's concentration inequality. This inequality involves some L^∞-norms, this is why we introduce the following notations. Let S be a linear space of measurable functions. For any function t in $L^2(\mu) \cap L^\infty(\mu)$, let $\pi_S(t)$ denote its orthogonal projection onto S, let $\|t\|_\infty$ be its L^∞-norm. For all C, C', η in $\bar{\mathbb{R}}_+$, for all t in $L^2(\mu)$, let

$$B_2(t, C, S) = \{t' \in S,\ \|t' - t\| \le C\},\ B(S) = B_2(0, 1, S) = \{t \in S,\ \|t\| \le 1\}. \tag{2.2}$$

$$B_{2,\infty}(C, C', \eta, S) = \left\{t \in L^2(\mu) \cap L^\infty(\mu),\ \|t\| \le C,\ \|t\|_\infty \le C',\ \|t - \pi_S(t)\| \le \eta\right\}. \tag{2.3}$$

2.4. Basic definitions

Definition 2.1 (confidence sets)**.** Let $(X_1, \ldots, X_n)$ be an i.i.d. sample of real valued random variables, let $S \subset L^2(\mu)$ and let β be a real number in $(0, 1)$. The set $CS(S, \beta)$ of $(1 - \beta)$-confidence balls on S is defined as the collection of all subsets $\hat{B}_\beta = B_2(\hat{s}, \hat{\rho}_\beta, S)$ of $L^2(\mu)$, where $\hat{s}$ and $\hat{\rho}_\beta$ are measurable with respect to $\sigma(X_1, \ldots, X_n)$ such that

$$\forall s \in S^*,\ \mathbb{P}_s\left(s \in \hat{B}_\beta\right) \ge 1 - \beta.$$

Definition 2.2 (minimax rate of convergence for confidence sets)**.** Let $(X_1,\dots,X_n)$ be an i.i.d. sample of real valued random variables, let $S'\subset S\subset L^2(\mu)$ and let α, β be real numbers in $(0,1)$. The (α,β)-minimax rate of convergence over S' for CS on S is defined as

$$\phi_n(\alpha,\beta,S,S') = \inf_{\hat{B}_\beta\in CS(S,\beta)} \Delta_{(S',\alpha)}(\hat{B}_\beta).$$

Definition 2.3 (adaptive confidence sets)**.** Let $(X_1,\dots,X_n)$ be an i.i.d. sample of real valued random variables, let $S\subset L^2(\mu)$, let $(S_m)_{m\in\mathcal{M}_n}$ be a collection of subsets of S and let α, β be real numbers in $(0,1)$. A CS $\hat{B}_\beta$ in $CS(S,\beta)$ is said to be optimal, or adaptive over $(S_m)_{m\in\mathcal{M}_n}$, if the following condition holds.

For all fixed α in $(0,1)$, there exists a constant $c(\alpha,\beta)>0$ free from n, S and $(S_m)_{m\in\mathcal{M}_n}$ such that, for all m in $\mathcal{M}_n$,

$$\Delta_{S_m,\alpha}(\hat{B}_\beta) \le c(\alpha,\beta)\phi_n(\alpha,\beta,S,S_m).$$

Definition 2.4 (test)**.** Let $(X_1,\dots,X_n)$ be an i.i.d. sample of real valued random variables. Let S be a family of densities on $\mathbb{R}$. Let S_0, S_1 be two disjoint subsets in S. A test T of the assumption $H_0: s\in S_0$ against the alternative $H_1: s\in S_1$ is a function $T:\mathbb{R}^n\to\{0,1\}$. The test T is said to have a confidence level $1-\alpha\in(0,1)$ when

$$\forall s\in S_0,\ \mathbb{P}_s\left(T(X_1,\dots,X_n)=0\right)\ge 1-\alpha.$$

It is said to have a power $1-\beta\in(0,1)$ when

$$\forall s\in S_1,\ \mathbb{P}_s\left(T(X_1,\dots,X_n)=1\right)\ge 1-\beta.$$

2.5. Main assumptions

Let $(S_m)_{m\in\mathcal{M}_n}$ be a collection of linear subspaces of $L^2(\mu)$, with finite dimensions respectively denoted by $(d_m)_{m\in\mathcal{M}_n}$. We make the following assumptions on this collection.

H1: There exists m_n in $\mathcal{M}_n$ such that $S_{m_n}=\mathrm{Span}\left(\bigcup_{m\in\mathcal{M}_n} S_m\right)$.

H2: There exists a constant C_1 such that, for all m in $\mathcal{M}_n$, for all t in S_m

$$\|t\|_\infty \le C_1\sqrt{d_m}\|t\|.$$

The last assumption is only technical and let us simplify the results. Let β be a real number in $(0,1)$.

H3: $(\mathcal{M},\beta)$: For all $n\ge 2$ $N_n=\mathrm{Card}(\mathcal{M}_n)$ is finite and there exists a constant $C_{\mathcal{M}}$ such that, for all $n\ge 2$,

$$\frac{2\sqrt{d_n}\ln(6N_n/\beta)}{n}\le C_{\mathcal{M}}.$$

Four examples are usually developed as fulfilling this set of assumptions:

[Hist] regular histogram spaces: for all m in $\mathbb{N}^*$, S_m is the space of all the functions constant on the partition $(I_{[k/m,(k+1)/m)})_{k=0,\dots,m-1}$ of $[0,1]$, $d_m=m$.

[T] trigonometric spaces: S_m is the linear span of the functions $\psi_{0,0}(x)=1_{[0,1]}$, $\psi_{j,1}(x)=\sqrt{2}\cos(2\pi jx)1_{[0,1]}(x)$ and $\psi_{j,2}(x)=\sqrt{2}\sin(2\pi jx)1_{[0,1]}(x)$ for all $1\le j\le J_m$. $d_m=2J_m+1$.

[P] regular piecewise polynomial spaces: S_m is the linear span of the functions $(\psi_{j,k})$ for $j=1,\dots,J_m$, $k=0,\dots,r-1$, where, for all $j=1,\dots,J_m$ and $k=0,\dots,r-1$, $\psi_{j,k}$ is a polynomial of degree k on $[(j-1)/J_m, j/J_m]$. $d_m=rJ_m$.

[W] spaces spanned by dyadic wavelets with regularity r.

We have to choose $d_{m_n}\le Cn^2/(\ln n)^2$ and $\beta\ge n^{-r}$ for some $r>0$ in order to fulfill Assumption **H3**$(\mathcal{M},\beta)$. For a description of those spaces and their properties, we refer to Birgé and Massart [7]. Hereafter, in order to simplify the notations, we will often write S_n, d_n, s_n,… instead of S_{m_n}, d_{m_n}, s_{m_n},…

3. Technical tools

This section presents the results required in Section 4 to build our adaptive confidence sets. Let s be a density in $L^2(\mu)$ and let s_m and s_n denote respectively its orthogonal projections onto the linear spaces S_m and S_n, where $S_m \subset S_n$. We recall the definition and some basic properties of the projection estimator $\hat{s}_m$ of s on S_m in Section 3.1. From Pythagoras theorem, it satisfies

$$\|s - \hat{s}_m\|^2 = \|s - s_n\|^2 + \|s_n - s_m\|^2 + \|s_m - \hat{s}_m\|^2. \tag{3.1}$$

Section 3.2 deals with the estimation of $\|s_m - \hat{s}_m\|^2$. We introduce our resampling estimator and state a very important concentration inequality (Thm. 3.3). In Section 3.3, we introduce our estimator of $\|s_n - s_m\|^2$ based on U-statistics.

3.1. Projection estimators

Definition 3.1 (projection estimators)**.** Let $X_1, \dots, X_n$ be i.i.d random variables with common density s in $L^2(\mu)$. Let S_m be a linear subspace of $L^2(\mu)$. The projection estimator of s on S_m is defined by

$$\hat{s}_m = \inf_{t \in S_m} \|t\|^2 - 2P_n t.$$

Classical computations show the following lemma:

Lemma 3.2. *Let $X_1, \dots, X_n$ be i.i.d random variables with common density s in $L^2(\mu)$. Let S_m be a linear subspace of $L^2(\mu)$ and let $(\psi_\lambda)_{\lambda \in \Lambda_m}$ be an orthonormal basis of S_m. Let s_m be the orthogonal projection of s onto S_m and let $\hat{s}_m$ be the projection estimator of s onto S_m. Then,*

$$s_m = \sum_{\lambda \in \Lambda_m} (P_s \psi_\lambda)\psi_\lambda,\ \hat{s}_m = \sum_{\lambda \in \Lambda_m} (P_n \psi_\lambda)\psi_\lambda,\ \|s_m - \hat{s}_m\|^2 = \sum_{\lambda \in \Lambda_m} \left[(P_n - P_s)\psi_\lambda\right]^2.$$

3.2. Estimation of $\|s_m - \hat{s}_m\|^2$ by resampling methods

Let s be a density in $L^2(\mu)$. Let S_m be a finite dimensional linear subspace of $L^2(\mu)$, let $(\psi_\lambda)_{\lambda \in \Lambda_m}$ be an orthonormal basis of S_m. Let s_m denote the orthogonal projection of s onto S_m and let $\hat{s}_m$ denote the projection estimator of s onto S_m. $\|s_m - \hat{s}_m\|^2$ is a functional of P_n and P_s, therefore, it can be estimated by resampling. Indeed, let $(W_1, \dots W_n)$ be a resampling scheme and let $\bar{W}_n = \sum_{i=1}^n W_i/n$. The resampling estimator of $\|s_m - \hat{s}_m\|^2$ given by Efron's heuristic (see Sect. 2.2) is defined for this resampling scheme and a suitably chosen constant C_W by:

$$p_W(S_m) = C_W \sum_{\lambda \in \Lambda_m} \mathbb{E}_W \left([(P_n^W - \bar{W}_n P_n)\psi_\lambda]^2 \right). \tag{3.2}$$

$p_W(S_m)$ is well defined since we can check with Cauchy–Schwarz inequality that

$$p_W(S_m) = C_W \mathbb{E}_W \left(\left[\sup_{t \in S_m, \|t\| \le 1} (P_n^W - \bar{W}_n P_n)t \right]^2 \right).$$

The deviations of $p_W(S_m)$ are given by the following theorem.

Theorem 3.3. *Let S_m be a linear subspace of $L^2(\mu)$ with finite dimension d_m, satisfying* **H2** *and let $C_3 > 0$. Let $X_1, \dots, X_n$ be an i.i.d. sample, let $(W_1, \dots W_n)$ be a resampling scheme and let $p_W(S_m)$ be the associated random variables defined in (3.2) for $C_W = \mathrm{Var}\left(W_1 - \bar{W}_n\right)$. There exists a constant $\kappa_v(C_1, C_3)$ such that, for all $2 \le x \le C_3 n/\sqrt{d_m}$, for all densities s in $L^2(\mu) \cap L^\infty(\mu)$,*

$$\mathbb{P}_s\left(\|s_m - \hat{s}_m\|^2 > p_W(S_m) + \kappa_v(C_1, C_3)(1 + \sqrt{\|s\|_\infty \wedge \|s\|\, d_m^{1/2} \wedge d_m})\frac{\sqrt{d_m}x}{n}\right) \le \mathrm{e}^{-x/2}.$$

Comments:

- This theorem is one of the main contributions of the article. It provides a sharp control of the variance term. It is the main difference with the article of Baraud who worked in a Gaussian framework and handled this term with a concentration inequality for χ^2-statistics of Birgé [7]. Our new construction is more general and can be easily adapted to other frameworks, which is not the case in Baraud [4].
- It is proved thanks to a technical lemma (Lem. 6.1) and a sharp concentration inequality (Lem. 6.2). Lemma 6.1 shows that, with our choice of C_W, $\|s_m - \hat{s}_m\|^2 - p_W(S_m)$ is a totally degenerate U-statistics of order 2. Lemma 6.2 is a concentration inequality for U-statistics of order 2.
- The Proof of Lemma 6.2 is derived from Houdré and Reynaud–Bouret [15], it follows mainly the one of Fromont and Laurent [11]. The main improvement compared with Fromont and Laurent [11] is that we work with general linear spaces S_m.
- The bound involves a term $\sqrt{\|s\|_\infty} \wedge \sqrt{\|s\| d_m^{1/4}} \wedge \sqrt{d_m}$. From a theoretical point of view, the term $\sqrt{\|s\| d_m^{1/4}} \wedge \sqrt{d_m}$ is useless asymptotically when $\|s\|_\infty$ is finite. In practice the L^2-norm of s is often much smaller than its L^∞-norm. Moreover, our control can also be used when $\|s\|_\infty$, $\|s\|$ or both of these quantities are unknown, since $\kappa_v(C_1, C_3)$ is free from $\|s\|$, $\|s\|_\infty$.
- The condition on x is not a problem in practice. We are interested in cases where $1 - \mathrm{e}^{-x/2}$ is large, therefore, $2 \le x$ will always be satisfied. Moreover, we will see in Section 4 that the assumptions $\mathbf{H3}(\mathcal{M}, \beta)$ are designed to ensure that the interesting x satisfy $x \le C_3 n/\sqrt{d_m}$ provided that C_3 is sufficiently large.
- This theorem can be used to build a model selection procedure of density estimation. Actually, an ideal penalty in this problem is given by $2\|s_m - \hat{s}_m\|^2$ and the aim of model selection is to evaluate this ideal penalty as precisely as possible. Theorem 3.3 provides such a control. This important application is discussed in detail in [24]. For an introduction to model selection, we refer to Massart [27]. The concept of ideal penalty is defined in Arlot [1].
- In order to keep the result as readable as possible, we only give the explicit form of the constant $\kappa_v(C_1, C_3)$ in the Proof of Theorem 3.3.

Corollary 3.4. *Let $X_1, \dots, X_n$ be i.i.d. real valued random variables. Let $(S_m)_{m \in \mathcal{M}_n}$ be a collection of finite dimensional linear spaces satisfying* **H1**, **H2**. *Let β be a real number in $(0,1)$ such that $\mathbf{H3}(\mathcal{M}, \beta)$ holds and let $M_2 > 0$, $M_\infty > 0$. Let $(W_1, \dots, W_n)$ be a resampling scheme and let $p_W(S_m)$ be the associated resampling estimator defined in Theorem 3.3. Let $\kappa_v(C_1, C_{\mathcal{M}})$ be the constant defined in Theorem 3.3 for $C_3 = C_{\mathcal{M}}$, let $x_n = 2\ln(2N_n/\beta) \vee 2$ and let*

$$V(m, \beta, X_1, \dots, X_n) = p_W(S_m) + \kappa_v(C_1, C_{\mathcal{M}})\left(1 + \sqrt{M_\infty \wedge M_2 d_m^{1/2} \wedge d_m}\right)\frac{\sqrt{d_m}x_n}{n}\cdot \qquad (3.3)$$

Then, for all densities s in $L^2(\mu) \cap L^\infty(\mu)$ such that $\|s\| \le M_2$ and $\|s\|_\infty \le M_\infty$,

$$\mathbb{P}_s\left(\exists m \in \mathcal{M}_n,\ \|s_m - \hat{s}_m\|^2 > V(m, \beta, X_1, \dots, X_n)\right) \le \frac{\beta}{2}\cdot$$

Comments:

- This corollary gives a uniform upper bound $V(m, \beta, X_1, \dots X_n)$ on the variance term.
- The size of this uniform bound, in the sense of (2.1), is given by the following theorem.

Theorem 3.5. *Let $X_1, \dots, X_n$ be i.i.d. real valued random variables. Let $(S_m)_{m \in \mathcal{M}_n}$ be a collection of linear spaces satisfying* **H1**, **H2**. *Let α, β be real numbers in $(0, 1)$ such that this collection satisfies also* **H3**$(\mathcal{M}, \alpha)$ *and* **H3**$(\mathcal{M}, \beta)$. *Let $M_2 > 0$, $M_\infty > 0$ and let $V_{m,\beta} = V(m, \beta, X_1, \dots, X_n)$ be the associated random variables defined in (3.3). There exists a constant κ, free from d_m, M_2, M_∞, α, β, such that, for all m in $\mathcal{M}_n$,*

$$\Delta^2_{B_{2,\infty}(M_2, M_\infty, 0, L^2(\mu)), \alpha}(V_{m,\beta}) \leq \kappa \left[\frac{d_m}{n} + \left(1 + \sqrt{M_\infty \wedge M_2 d_m^{1/2} \wedge d_m} \right) \frac{\sqrt{d_m}}{n} \ln \left[\frac{N_n}{\alpha \beta} \right] \right].$$

Comments:

- For fixed confidence level α, β, the asymptotic order of magnitude of $V_{m,\beta}$ is d_m/n for all models with dimension $d_m \geq (\ln N_n)^2$.

3.3. Estimation of $\|s_n - s_m\|^2$

The simple following lemma is important to understand our procedure.

Lemma 3.6. *Let $X_1, \dots, X_n$ be i.i.d. real valued random variables with common density s in $L^2(\mu)$. Let $S_m \subset S_n$ be two linear subspaces of $L^2(\mu)$, with respective finite dimensions d_m and d_n. Let s_m and s_n be the orthogonal projections of s respectively onto S_m and S_n. Let $(\psi_\lambda)_{\lambda \in \Lambda_n}$ be an orthonormal basis of S_n such that $(\psi_\lambda)_{\lambda \in \Lambda_m}$ is an orthonormal basis of S_m, with $\Lambda_m \subset \Lambda_n$. Then*

$$\|s_n - s_m\|^2 = \sum_{\lambda \in \Lambda_n - \Lambda_m} (P_s \psi_\lambda)^2 = \mathbb{E}_s \left(\frac{1}{n(n-1)} \sum_{i \neq j = 1}^{n} \sum_{\lambda \in \Lambda_n - \Lambda_m} \psi_\lambda(X_i) \psi_\lambda(X_j) \right). \tag{3.4}$$

Based on this kind of lemma, Laurent [21,22] introduced the estimators based on U-statistics to estimate quadratic functionals of a density. These estimators were successfully used by Fromont and Laurent [11] for goodness of fit tests in a density estimation model, and by Robins and van der Vaart [28] to build adaptive confidence sets. We follow the same steps here and define, for any observation $X_1, \dots X_n$, for all finite dimensional linear spaces $S_m \subset S_n$, for all orthonormal basis $(\psi_\lambda)_{\lambda \in \Lambda_n}$ of S_n such that $(\psi_\lambda)_{\lambda \in \Lambda_m}$ is an orthonormal basis of S_m, with $\Lambda_m \subset \Lambda_n$,

$$p_b(S_m, S_n) = \frac{1}{n(n-1)} \sum_{i \neq j = 1}^{n} \sum_{\lambda \in \Lambda_n - \Lambda_m} \psi_\lambda(X_i) \psi_\lambda(X_j). \tag{3.5}$$

$p_b(S_m, S_n)$ is well defined since we can prove with Cauchy–Schwarz inequality that, if $S_n^{\perp m}$ denotes the orthogonal of S_m in S_n,

$$p_b(S_m, S_n) = \frac{1}{n-1} \left(n \sup_{t \in B_2(S_n^{\perp m})} (P_n t)^2 - P_n \left(\sup_{t \in B_2(S_n^{\perp m})} t^2 \right) \right).$$

The deviations of $p_b(S_m, S_n)$ are given by the following result:

Lemma 3.7. *Let $X_1, \dots, X_n$ be i.i.d. real valued random variables. Let $S_m \subset S_n$ be two linear subspaces of $L^2(\mu)$, with respective finite dimensions d_m and d_n and let $p_b(S_m, S_n)$ be the estimator defined in (3.5). For any density s in $L^2(\mu)$, let s_n and s_m denote its orthogonal projections respectively onto S_n and S_m. For all $C_3 > 0$ and all ϵ in $(0, 1)$, there exists a real constant $\kappa_b(\epsilon, C_3)$ such that, for all $2 \leq x \leq C_3 n / \sqrt{d_n}$, for all densities s in $L^2(\mu) \cap L^\infty(\mu)$, with $\mathbb{P}_s$-probability larger than $1 - 3\mathrm{e}^{-x/2}$,*

$$\left| p_b(S_m, S_n) - \|s_n - s_m\|^2 \right| \leq \epsilon \|s_n - s_m\|^2 + \kappa_b(\epsilon, C_3) \left(1 + \sqrt{\|s\|_\infty \wedge \|s\|_2 \, d_n^{1/2}} \right) \frac{\sqrt{d_n} x}{n}.$$

Thanks to this lemma, we can derive the following corollary that gives our estimation of $\|s_n - s_m\|$.

Corollary 3.8. *Let $X_1,\dots,X_n$ be i.i.d. real valued random variables. Let $(S_m)_{m\in\mathcal{M}_n}$ be a collection of linear spaces satisfying assumptions* **H1**, **H2**. *Let β be a real number in $(0,1)$ such that this collection satisfies also* **H3**$(\mathcal{M},\beta)$. *Let $M_2>0$, $M_\infty>0$, $x_n=2\ln(6N_n/\beta)\vee 2$. Let p_b be defined in (3.5) and, for all ϵ in $(0,1)$, let $\kappa_b(\epsilon,C_{\mathcal{M}})$ be the constant defined in Lemma 3.7 for $C_3=C_{\mathcal{M}}$. For all $m\in\mathcal{M}_n$, let*

$$K(m,\beta,X_1,\dots,X_n)=\inf_{\epsilon\in(0,1)}\frac{p_b(S_m,S_n)}{1-\epsilon}+\frac{\kappa_b(\epsilon,C_{\mathcal{M}})}{1-\epsilon}\left(1+\sqrt{M_\infty\wedge M_2 d_n^{1/2}}\right)\frac{\sqrt{d_n}x_n}{n}\cdot \tag{3.6}$$

Then, for all densities s in $B_{2,\infty}(M_2,M_\infty,0,L^2(\mu))$,

$$\mathbb{P}_s\left(\exists m\in\mathcal{M}_n,\ \|s_n-s_m\|^2>K(m,\beta,X_1,\dots,X_n)\right)\le\frac{\beta}{2}\cdot$$

Comments:

- This corollary gives a sharp estimation of the bias term. In particular, we will see in the following section that the term $\sqrt{d_n}x_n/n$ is essentially necessary.
- We obtain a bound valid for all the models in the collection $\mathcal{M}_n$. Combined with Corollary 3.4, it gives all the tools required to apply our method of selection.

4. Main results

4.1. Adaptive confidence balls

We can now easily present our model selection procedure to obtain CS.

Construction of the adaptive CS

Let β be a real number in $(0,1)$, let $M_2>0$, $M_\infty>0$, let $(S_m)_{m\in\mathcal{M}_n}$ be a collection of finite dimensional linear spaces and let $S_n=\mathrm{Span}\left(\bigcup_{m\in\mathcal{M}_n}S_m\right)$. Let $(V(m,\beta,X_1,\dots,X_n))_{m\in\mathcal{M}_n}$ be the collection defined in (3.3), let $(K(m,\beta,X_1,\dots,X_n))_{m\in\mathcal{M}_n}$ be the collection defined in (3.6) and let η be a positive real number. For all m in $\mathcal{M}_n$, let

$$\hat\rho(m,\eta,\beta)=\sqrt{\eta^2+K(m,\beta,X_1,\dots,X_n)+V(m,\beta,X_1,\dots,X_n)}.$$

Recall the definition of the L^2-ball centered in an element t of $L^2(\mu)$ with radius C in $\mathbb{R}$ given in (2.2). Our final CS is defined by

$$\hat B_{\beta,\eta}=B_2(\hat s_{\hat m},\hat\rho(\hat m,\eta,\beta),L^2(\mu)),\ \text{where}\quad \hat m=\arg\min_{m\in\mathcal{M}_n}\{\hat\rho(m,\eta,\beta)\}. \tag{4.1}$$

Performances of our CS

Theorem 4.1. *Let $X_1,\dots,X_n$ be i.i.d real valued random variables. Let $(S_m)_{m\in\mathcal{M}_n}$ be a collection of models satisfying assumptions* **H1**, **H2**. *Let β be a real number in $(0,1)$ such that this collection satisfies also* **H3**$(\mathcal{M},\beta)$. *Let $M_2>0$, $M_\infty>0$, $\eta>0$ and let $B_{2,\infty}(M_2,M_\infty,\eta,S_n)$ be the ball defined in (2.3).*
Then $\hat B_{\beta,\eta}$, defined in (4.1), belongs to $CS(B_{2,\infty}(M_2,M_\infty,\eta,S_n),\beta)$.

Moreover, there exists a constant κ such that for all m in $\mathcal{M}_n$, for all $\eta_m>0$ and all α such that $(S_m)_{m\in\mathcal{M}_n}$ satisfies also **H3**$(\mathcal{M},\alpha)$

$$\Delta_{B_{2,\infty}(M_2,M_\infty,\eta_m,S_m),\alpha}(\hat B_{\beta,\eta})\le\kappa\left(\left(\eta_m^2+\frac{d_m}{n}\right)\vee\left(\eta^2+\frac{\sqrt{d_n}\ln(N_n/(\alpha\beta))}{n}\right)\right)\cdot \tag{4.2}$$

Comments:

- Theorem 4.1 gives CS over $B_{2,\infty}(M_2,M_\infty,\eta,S_n)$, with prescribed confidence level β, valid for all $n\ge 2$.

- The size of these CS is upper bounded by the maximum of two terms. $\eta^2 + \sqrt{d_n}/n$ is the minimax separation rate for the tests $H_0: \ s = s_0$ against the alternative $H_1: \ s \in B_{2,\infty}(M_2, M_\infty, \eta, S_n) - \{s_0\}$, where s_0 is some element in S_m^*. $\eta_m^2 + d_m/n$ is the minimax estimation rate over $B_{2,\infty}(M_2, M_\infty, \eta_m, S_m)$.
- Robins and van der Vaart [28] proved that these rates are optimal asymptotically. We will show in Theorem 4.2 below that this property holds also non asymptotically.
- $\hat{\rho}(m, \eta, \beta)$ has basically the following form

$$\hat{\rho}^2(m, \eta, \beta) = \eta^2 + p_b(S_m, S_n) + p_W(S_m) + \kappa(M_2, M_\infty) \frac{\sqrt{d_n} \ln(N_n/(\alpha\beta))}{n}.$$

  It depends in practice on two unknown constants, η and $\kappa(M_2, M_\infty)$. We believe that some "slope heuristic" (see Birgé and Massart [8], Arlot and Massart [2] or [24]) method can be developed for CS in order to obtain a data driven estimate of $\kappa(M_2, M_\infty)$. This estimate would probably be more reasonable than the upper bound given in our proof. On the other hand, we believe that the constant η can only be handled with suitably chosen assumptions. For example, some regularity assumption as in Section 4.3 bellow.
- Baraud [4] used a procedure almost similar in a regression framework. He defined, for all m in $\mathcal{M}_n$, a test T_m to test the null hypothesis $s_n \in S_m$ against the alternative $s_n \in S_n - S_m$ and some positive number $\hat{\rho}(m)$. His $\hat{\rho}(m)$'s are calibrated to satisfy the property that, if T_m accepts the null, then, with probability close to one, the distance between s and its projection estimator $\hat{s}_m$ is not larger than $\hat{\rho}(m)$. He selected $\hat{m}$ as the minimizer of $\hat{\rho}(m)$ among those m for which T_m accepts the null and defined the confidence ball as the L^2-ball centered at $\hat{s}_{\hat{m}}$ of radius $\hat{\rho}(\hat{m})$. The main difference with this general scheme is that our procedure does not require a series of tests to work as the bound given in Corollary 3.8 holds for all m.

4.2. **Optimality of our balls**

In this section we prove that the rate given in (4.2) can not be improved in general, from a minimax point of view. The result is stated in the following theorem:

Theorem 4.2. *Let S_n be the set of histograms on $\{[k/d_n, (k+1)/d_n), \ k = 0, \dots, d_n - 1\}$ and let S_m be the linear subspace of S_n of histograms on $\{[k/d_m, (k+1)/d_m), \ k = 0, \dots, d_m - 1\}$. Let α, β be real numbers in $(0,1)$ such that $2\alpha + \beta < 1$. There exists a constant $C(\alpha, \beta)$, such that*

$$\phi_n^2(\alpha, \beta, S_n, S_m) \geq C(\alpha, \beta) \left(\frac{\sqrt{d_n}}{n} \vee \frac{d_m}{n} \right).$$

Comments:

- Theorem 4.2 gives the optimality of the rate given in (4.2), since the terms η and η_m can obviously not be avoided also.
- The key point of the proof (Lem. 6.8) is that we can not build a test of null hypothesis $H_0 : s \in S_m$ against the alternative $H_1 : s \in S_n, \ s \notin S_m$ with separation rate smaller than $C_{\alpha,\beta}\sqrt{d_n}/n$. This extends the result of Ingster [16–18] to a non asymptotical framework and the result of Baraud [4] to density estimation. For a definition of the separation rate, we refer to Ingster [16–18].
- The proof follows the methodology described in Baraud [4].

4.3. **Application to regular density**

This section presents the application of Theorem 4.1 to regular densities. In particular, we extend the result of Robins and van der Vaart [28] since (1.1) is obtained for all n.

Fourier spaces:
For all k in $\mathbb{N}^*$, for all x in $\mathbb{R}$, let

$$\psi_{1,k}(x) = \sqrt{2}\cos(2\pi kx)I_{[0,1]}(x),\ \psi_{2,k}(x) = \sqrt{2}\sin(2\pi kx)I_{[0,1]}(x).$$

For all d in $\mathbb{N}$, let F_d be the linear span of $I_{[0,1]},\ \psi_{1,k},\ \psi_{2,k}$, for all k in $\{1,\dots,d\}$. F_d is a subspace of $L^2(\mu)$. It is a classical result (see for example Birgé and Massart [7]) that any sub-collection of $(F_{d_m})_{0\leq d_m\leq n^2(\ln n)^{-2}}$ satisfies **H1, H2** with $C_1 = 1$. We can also easily check that, for all $\beta \geq n^{-2}$, it satisfies also **H3**$(\mathcal{M},\beta)$ with $C_{\mathcal{M}} = 4$.

Sobolev Spaces:
For all functions t in $L^2(\mu)$, let

$$t_0 = \int_{\mathbb{R}} t(x)I_{[0,1]}(x)\mathrm{d}\mu(x) = \int_0^1 t(x)\mathrm{d}\mu(x)$$

and for all $k \in \mathbb{N}^*$, let

$$t_{1,k} = \int_{\mathbb{R}} t(x)\psi_{1,k}(x)\mathrm{d}\mu(x),\ t_{2,k} = \int_{\mathbb{R}} t(x)\psi_{2,k}(x)\mathrm{d}\mu(x).$$

For all $\gamma \in \mathbb{R}_+^*$, for all M in $\mathbb{R}_+$, we denote by $S(\gamma, M)$, the set of functions t in $L^2(\mu)$ such that

$$t_0^2 + \sum_{i\in\mathbb{N}^*}\left(t_{1,i}^2 + t_{2,i}^2\right) i^{2\gamma} \leq M^2.$$

It is clear that for all t in $S(\gamma, M)$, $\|t\| \leq M$ and for all d in $\mathbb{N}$, if $\pi_{F_d}(t)$ denotes the orthogonal projection of t onto F_d,

$$\|t - \pi_{F_d}(t)\|^2 = \sum_{i>d}\left(t_{1,i}^2 + t_{2,i}^2\right) \leq \frac{1}{(d+1)^{2\gamma}}\sum_{i>d}\left(t_{1,i}^2 + t_{2,i}^2\right) i^{2\gamma} \leq \frac{M^2}{(d+1)^{2\gamma}}.$$

We can also use Cauchy–Schwarz inequality to prove that, when $\gamma > 1/2$, for all x in $[0,1]$,

$$|t(x)| \leq |t_0| + \sqrt{2\left(\sum_{i\in\mathbb{N}}(t_{1,i}^2 + t_{2,i}^2)^2(i+1)^{2\gamma}\right)\left(\sum_{i\in\mathbb{N}}\frac{\cos^2(2\pi ix) + \sin^2(2\pi ix)}{(i+1)^{2\gamma}}\right)}.$$

Hence, when $\gamma > 1/2$, for all t in $S(\gamma, M)$, $\|t\|_\infty \leq 2M\sqrt{\sum_{i\in\mathbb{N}}(i+1)^{-2\gamma}}$. When $\gamma > 1/2$, let $M_\infty = 2M\sqrt{\sum_{i\in\mathbb{N}}(i+1)^{-2\gamma}}$ and when $\gamma \leq 1/2$, let M_∞ denote a positive real number. We have obtained that

$$S(\gamma, M, M_\infty) := \{t \in S(\gamma, M),\ \|t\|_\infty \leq M_\infty\} \subset B_{2,\infty}\left(M, M_\infty, M(d+1)^{-\gamma}, F_d\right). \tag{4.3}$$

Hence, the following proposition holds.

Proposition 4.3. *We keep the previous notations. Let γ, M, M_∞ be strictly positive real numbers, let d_n denotes the integer part of $n^{(2\gamma+1/2)^{-1}} \wedge n^2(\ln n)^{-2}$ and let $\mathcal{M}_n = \{1,\dots,d_n\}$.*

Let $\hat{B}_{\beta,M(d_n+1)^{-\gamma}}$ be the set defined in Theorem 4.1 for the collection $(F_{d_m})_{d_m\in\mathcal{M}_n}$. Then, $\hat{B}_{\beta,M(d_n+1)^{-\gamma}}$ belongs to $CS(S(\gamma, M, M_\infty), \beta)$.

There exists a constant κ free from n such that, for all $\gamma' \geq \gamma$,

$$\Delta_{S(\gamma',M,M_\infty),\alpha}\left(\hat{B}_{\beta,M(d_n+1)^{-\gamma}}\right) \leq \kappa\left(n^{-\gamma'/(2\gamma'+1)} \vee (\ln n)n^{-2\gamma/(4\gamma+1)}\right).$$

Comments:

- This result can be compared with the one of Robins and van der Vaart [28]. Our balls satisfy the covering property (1.1) for all n and not asymptotically as in their paper. They proved that the rate $n^{-\gamma'/(2\gamma'+1)} \vee n^{-2\gamma/(4\gamma+1)}$ is asymptotically optimal.
- It is a straightforward consequence of Theorem 4.1, applied with $\eta_m = M(d_m+1)^{-\gamma'}$, $\eta = M(d_n+1)^{-\gamma}$ and the previous computations, therefore, the proof is omitted.

5. Simulation study

In this section, our first goal is to illustrate Theorem 3.3. We proved that the difference $\|s_m - \hat{s}_m\|_2^2 - p_W(S_m)$ is upper bounded by $\sqrt{d_m}/n$, we will show that this bound is sharp on some simulations. Then, we will consider a more general version of Efron's heuristics, which states that, for a good choice of the constant C_W, the distribution of $\|s_m - \hat{s}_m\|_2^2$ is close to the conditional distribution $\mathcal{D}^W\left(C_W \sum_{\lambda\in\Lambda_m}[(P_n^W - \bar{W}_n)\psi_\lambda]^2\right)$. The quantiles of $\|s_m - \hat{s}_m\|_2^2$ must then be close to their resampled counterpart. In a second simulation, we test this method and remark that it gives very good practical results.

5.1. **Illustration of Theorem 3.3**

In this simulation, s is the uniform density on $[0,1]$, S_m is the set of histograms on the partition $([(k-1)/d_m, k/d_m))_{k=1,\ldots,d_m}$. $(W_1,\ldots,W_n)$ are Efron's weights, *i.e.* the distribution $\mathcal{D}(W_1,\ldots,W_n)$ is the multinomial distribution $\mathcal{M}(n,1/n,\ldots,1/n)$. In order to compute $p_W(S_m)$, we estimate the conditional expectation $\mathbb{E}^W(\sum_{\lambda\in\Lambda}[(P_n^W - \bar{W}_n)\psi_\lambda]^2)$ by a Monte Carlo method with n_b repetitions. Finally, we repeat $p = 1000$ times the experiment. We plot the histograms of the p values of the normalized difference $n(\|s_m - \hat{s}_m\|_2^2 - p_W(S_m))/\sqrt{d_m}$. The first histogram is obtained with $n = 50, d_m = 10, n_b = 100$ and the second for $n = 200, d_m = 50, n_b = 500$.

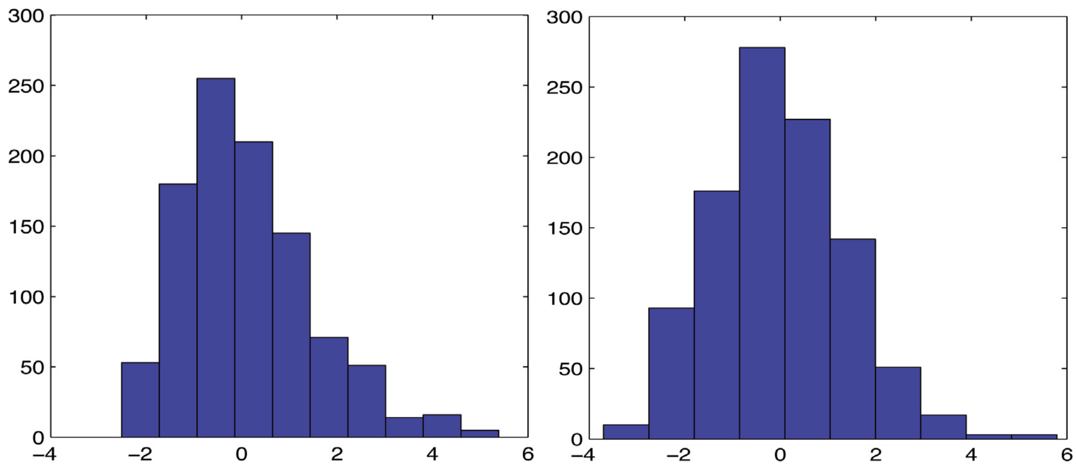


FIGURE 1. $\frac{n}{\sqrt{d_m}}(\|s_m - \hat{s}_m\|_2^2 - p_W(S_m))$.

Comments:

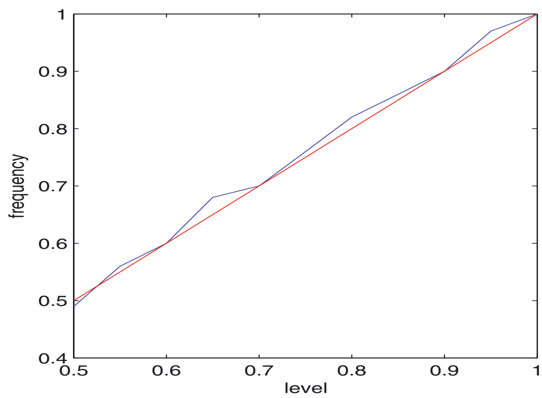
- The distribution of $n(\|s_m - \hat{s}_m\|_2^2 - p_W(S_m))/\sqrt{d_m}$ does not change with n or d_m. This shows that the result of Theorem 3.3 is sharp in this example, at least, up to the constant in front of the remainder term.

5.2. Illustration of the second Efron's heuristic

In this simulation, we keep the same s and the same resampling scheme. S_m is the set of functions constant on the partition $([(k-1)/d_m, k/d_m))_{k=1,\dots,d_m}$, with $d_m = 50$. $n = 100$, $N = 100$ and $((X_i^J)_{i=1,\dots,n})_{J=1,\dots,N}$ are N independent samples with common law $\mathbb{P}_s$. For all $J = 1,\dots,N$, we compute the projection estimator $\hat{s}_m^J$ on S_m with the sample $(X_i^J)_{i=1,\dots,n}$. Then, we take $n_b = 10\,000$ resampling schemes $(W_1,\dots,W_n)$. For all resampling schemes, we compute the quantity

$$p_W^J(S_m) = \frac{1}{v_W^2}\left(\sum_{\lambda\in\Lambda}[(P_n^{J,W} - \bar{W}_n P_n^J)\psi_\lambda]^2\right)$$

and we obtain an approximation of the $(1-\alpha)$-quantiles $\hat{q}_\alpha^J$ of its conditional distribution $\mathcal{D}^W(p_W^J\ (S_m))$. We plot the frequency of J such that $\left\|s_m - \hat{s}_m^J\right\|^2 \le \hat{q}_\alpha^J$ and the function $f(\alpha) = \alpha$ when α varies in $(0.5, 1)$ in the following curves.



Comments

- The covering property of this empirical ball is very close to the one we would like to obtain. Hence, this method seems to give sharp confidence balls for s_m. The computation time is the same as in the first method.
- We do not prove any theoretical evidence of this covering property. In particular, we cannot guarantee that $\mathbb{P}_s(\|s_m - \hat{s}_m\|_2^2 \le \hat{q}_\alpha) \ge 1-\alpha$ occurs for any n.

6. Proofs

6.1. Proof of Theorem 3.3

The theorem can easily be deduced from the following Lemmas, whose proofs are postponed to the appendix.

Lemma 6.1. *Let $X_1,\dots,X_n$ be an i.i.d sample with common density s in $L^2(\mu)$ and let $(\psi_\lambda)_{\lambda\in\Lambda}$ be an orthonormal system in $L^2(\mu)$. Let $W_1,\dots W_n$ be a resampling scheme, let $\bar{W}_n = n^{-1}\sum_{i=1}^n W_i$ and let $C_W = $* $\mathrm{Var}(W_1 - \bar{W}_n)^{-1}$.

Let $T_s(\Lambda) = \sum_{\lambda\in\Lambda}(\psi_\lambda - P_s\psi_\lambda)^2$,

$$p_s(\Lambda) = \sum_{\lambda\in\Lambda}\left[(P_n - P_s)\psi_\lambda\right]^2, \; p_W(\Lambda) = C_W \mathbb{E}_W\left(\sum_{\lambda\in\Lambda}\left[(P_n^W - \bar{W}_n P_n)\psi_\lambda\right]^2\right),$$

$$U_s(\Lambda) = \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n}\sum_{\lambda\in\Lambda}(\psi_\lambda(X_i) - P_s\psi_\lambda)(\psi_\lambda(X_j) - P_s\psi_\lambda).$$

Then

$$p_s(\Lambda) = \frac{1}{n}P_nT_s(\Lambda) + \frac{n-1}{n}U_s(\Lambda),\; p_W(\Lambda) = \frac{1}{n}P_nT_s(\Lambda) - \frac{1}{n}U_s(\Lambda),\; p_s(\Lambda) - p_W(\Lambda) = U_s(\Lambda).$$

Lemma 6.2. *Let $X_1,\dots,X_n$ be an i.i.d sample with common density s in $L^2(\mu)$ and let $(\psi_\lambda)_{\lambda\in\Lambda}$ be an orthonormal system in $L^2(\mu)$. Let $D_{s,\Lambda} = \sum_{\lambda\in\Lambda}P_s\left((\psi_\lambda - P_s\psi_\lambda)^2\right)$,*

$$U_s(\Lambda) = \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n}\sum_{\lambda\in\Lambda}(\psi_\lambda(X_i) - P_s\psi_\lambda)(\psi_\lambda(X_j) - P_s\psi_\lambda),$$

$$B(\Lambda) = \left\{\sum_{\lambda\in\Lambda}a_\lambda\psi_\lambda;\; \sum_{\lambda\in\Lambda}a_\lambda^2 \leq 1\right\},\; v_{s,\Lambda}^2 = \sup_{t\in B(\Lambda)}P_s\left((t-Pt)^2\right),\; b_\Lambda = \sup_{t\in B(\Lambda)}\|t\|_\infty.$$

For all ξ in $\{-1,1\}$, for all $x>0$, we have

$$\mathbb{P}_s\left(\xi U_s(\Lambda) > 5.7 v_{s,\Lambda}\frac{\sqrt{D_{s,\Lambda}x}}{n} + 8v_{s,\Lambda}^2\frac{x}{n} + 384\sqrt{2}v_{s,\Lambda}b_\Lambda\left(\frac{x}{n}\right)^{3/2} + 2040b_\Lambda^2\left(\frac{x}{n}\right)^2\right) \leq \mathrm{e}\mathrm{e}^{-x}.$$

Lemma 6.3. *Let S be a linear space with finite dimension d satisfying assumption* **H2**. *Let s be a density in $L^2(\mu)\cap L^\infty(\mu)$, let $(\psi_\lambda)_{\lambda\in\Lambda}$ be an orthonormal basis of S. Let*

$$B(\Lambda) = \left\{\sum_{\lambda\in\Lambda}a_\lambda\psi_\lambda;\; \sum_{\lambda\in\Lambda}a_\lambda^2 \leq 1\right\}, v_{s,\Lambda}^2 = \sup_{t\in B(\Lambda)}P_s\left((t-Pt)^2\right),\; b_\Lambda = \sup_{t\in B(\Lambda)}\|t\|_\infty,$$

$$D_{s,\Lambda} = \sum_{\lambda\in\Lambda}P_s\left((\psi_\lambda - P_s\psi_\lambda)^2\right) = P_s\left(\sup_{t\in B(\Lambda)}(t-P_st)^2\right).$$

We have

$$v_{s,\Lambda}^2 \leq \|s\|_\infty \wedge C_1\|s\|\sqrt{d},\; v_{s,\Lambda}^2 \leq D_{s,\Lambda} \leq b_\Lambda^2 \leq C_1^2 d.$$

Let us now explain briefly the Proof of Theorem 3.3. Let $X_1,\dots,X_n$ be an i.i.d sample with common density s in $L^2(\mu)\cap L^\infty(\mu)$. Let $(\psi_\lambda)_{\lambda\in\Lambda_m}$ be an orthonormal basis in S_m. It comes from Lemmas 6.1 and 6.2 that, using the notations of these lemmas, for all $x>0$, there exists an absolute constant $\kappa = 2040$ such that, with probability larger than $1-\mathrm{e}^{-x+1}$

$$\|s_m - \hat{s}_m\|^2 \leq p_W(S_m) + \kappa\left(v_{s,\Lambda_m}\frac{\sqrt{D_{s,\Lambda_m}x}}{4n} + v_{s,\Lambda_m}^2\frac{x}{4n} + v_{s,\Lambda_m}b_{\Lambda_m}\left(\frac{x}{n}\right)^{3/2} + b_{\Lambda_m}^2\left(\frac{x}{n}\right)^2\right). \tag{6.1}$$

Since $x\geq 2$, $\sqrt{x}\leq x$ and $x-1\geq x/2$. We have

$$2v_{s,\Lambda_m}b_{\Lambda_m}\left(\frac{x}{n}\right)^{3/2} \leq v_{s,\Lambda_m}^2\frac{x}{n} + b_{\Lambda_m}^2\left(\frac{x}{n}\right)^2,\; v_{s,\Lambda_m}^2 \leq D_{s,\Lambda_m}.$$

Hence, from (6.1), with probability larger than $1-\mathrm{e}^{-x/2}$,

$$\|s_m-\hat{s}_m\|^2 \le p_W(S_m)+\kappa\left(v_{s,\Lambda_m}\frac{\sqrt{D_{s,\Lambda_m}}x}{n}+\frac{3}{2}b_{\Lambda_m}^2\left(\frac{x}{n}\right)^2\right).$$

Since $\sqrt{d_m}x/n\le C_3$, $d_m x^2/n^2\le C_3\sqrt{d_m}x/n$, from Lemma 6.3,

$$v_{s,\Lambda_m}\frac{\sqrt{D_{s,\Lambda_m}}x}{n}+\frac{3}{2}b_{\Lambda_m}^2\left(\frac{x}{n}\right)^2\le C_1\left(\sqrt{\|s\|_\infty\wedge C_1\|s\|\sqrt{d}\wedge C_1^2 d}+\frac{3}{2}C_1C_3\right)\frac{\sqrt{d_m}x}{n}. \tag{6.2}$$

This concludes the Proof of Theorem 3.3, with $\kappa_v=2040C_1(1\vee C_1\vee 3C_1C_3/2)$.

6.2. **Proof of Corollary 3.4**

We use a union bound to obtain that

$$\begin{aligned}&\mathbb{P}_s\left(\exists m\in\mathcal{M}_n,\ \|s_m-\hat{s}_m\|^2>V(m,\beta,X_1,\dots,X_n)\right)\\ &\le N_n\max_{m\in\mathcal{M}_n}\mathbb{P}_s\left(\|s_m-\hat{s}_m\|^2>V(m,\beta,X_1,\dots,X_n)\right).\end{aligned}$$

All the models satisfy **H2**. From assumption **H3**$(\mathcal{M},\beta)$, x_n satisfies $2\le x_n\le C_3 n/\sqrt{d_m}$ with $C_3=C_{\mathcal{M}}$, thus, from Theorem 3.3, for all m in $\mathcal{M}_n$,

$$\mathbb{P}_s\left(\|s_m-\hat{s}_m\|^2>V(m,\beta,X_1,\dots,X_n)\right)\le \mathrm{e}^{-x_n/2}.$$

Finally, $\mathrm{Card}(\mathcal{M}_n)\mathrm{e}^{-x_n/2}\le\frac{\beta}{2}$, which concludes the Proof of Corollary 3.4.

6.3. **Proof of Theorem 3.5**

Let s be a density in $L^2(\mu)\cap L^\infty(\mu)$, we only have to prove that there exists a constant κ such that, with $\mathbb{P}_s$-probability larger than $1-\alpha$,

$$\forall m\in\mathcal{M}_n,\ p_W(S_m)\le\kappa\left(\frac{d_m}{n}+\left(1+\sqrt{\|s\|_\infty\wedge\|s\|d_m^{1/2}\wedge d_m}\right)\frac{\sqrt{d_m}}{n}\ln\left[\frac{N_n}{\alpha}\right]\right).$$

Let $(\psi_\lambda)_{\lambda\in\Lambda_m}$ be an orthonormal basis of S_m, from Lemma 6.1 and using the notations of this lemma,

$$p_W(\Lambda)=\frac{1}{n}P_nT_s(\Lambda_m)-\frac{1}{n}U_s(\Lambda_m).$$

We follow the Proof of Theorem 3.3. From Lemmas 6.2 and 6.3 and assumptions **H1, H2, H3**$(\mathcal{M},\alpha)$, there exists a constant κ such that

$$\mathbb{P}_s\left(\exists m\in\mathcal{M}_n,\ U_s(\Lambda_m)>\kappa\sqrt{\|s\|_\infty\wedge\|s\|d_m^{1/2}\wedge d_m}\frac{\sqrt{d_m}\ln[N_n/\alpha]}{n}\right)\le\alpha.$$

Moreover, it is easy to check, with Cauchy–Schwarz inequality, that, using the notations of Lemma 6.3

$$T_s(\Lambda_m)=\sup_{t\in B(\Lambda_m)}(t-P_st)^2.$$

Hence, using assumptions **H2**, we obtain

$$P_n T_s(\Lambda_m) \leq \|T_s(\Lambda_m)\|_\infty \leq 2C_1^2 d_m.$$

This conclude the Proof of Theorem 3.5.

6.4. Proof of Lemma 3.7

Let $X_1, \ldots, X_n$ be an i.i.d sample with common density s in $L^2(\mu) \cap L^\infty(\mu)$. Let $(\psi_\lambda)_{\lambda\in\Lambda_n}$ be an orthonormal basis of S_n such that $(\psi_\lambda)_{\lambda\in\Lambda_m}$ is an orthonormal basis of S_m, with $\Lambda_m \subset \Lambda_n$. The Hoeffding's decomposition of the U-statistic $p_b(S_m, S_n)$ can be written

$$\begin{aligned} p_b(S_m, S_n) &= U_s(\Lambda_n - \Lambda_m) + 2P_n\left(\sum_{\lambda\in\Lambda_n-\Lambda_m}(P_s\psi_\lambda)(\psi_\lambda - P_s\psi_\lambda)\right) + \sum_{\lambda\in\Lambda_n-\Lambda_m}(P_s\psi_\lambda)^2 \\ &= U_s(\Lambda_n - \Lambda_m) + 2(P_n - P_s)(s_n - s_m) + \|s_n - s_m\|^2, \end{aligned}$$

where, as usually, for all indexes sets Λ,

$$U_s(\Lambda) = \frac{1}{n(n-1)} \sum_{i\neq j=1}^{n} \sum_{\lambda\in\Lambda} (\psi_\lambda(X_i) - P_s\psi_\lambda)(\psi_\lambda(X_j) - P_s\psi_\lambda).$$

It comes from Lemmas 6.2 and 6.3 that, for all $2 \leq x \leq C_3 n/\sqrt{d_n}$,

$$\mathbb{P}_s\left(|U_s(\Lambda_n - \Lambda_m)| > \kappa_v(C_1, C_3)\left(1 + \sqrt{\|s\|_\infty \wedge \|s\|\, d_n^{1/2}}\right)\frac{\sqrt{d_n}x}{n}\right) \leq 2\mathrm{e}^{-x/2}.$$

If $s_n = s_m$, this concludes the proof. Else, let ϵ in $(0,1)$, the inequality $2ab \leq \epsilon a^2 + \epsilon^{-1}b^2$ gives

$$2|(P_n - P_s)(s_n - s_m)| \leq \epsilon\|s_n - s_m\|^2 + \epsilon^{-1}\left((P_n - P_s)\left(\frac{s_n - s_m}{\|s_n - s_m\|}\right)\right)^2.$$

The function $s_{m,n} = (s_n - s_m)/\|s_n - s_m\|$ satisfies $\|s_{m,n}\| \leq 1$ and, from Bernstein's inequality, for all $x > 0$,

$$\mathbb{P}_s\left(|(P_n - P_s)(s_{n,m})| > \sqrt{2P_s\left[(s_{m,n} - P_s s_{m,n})^2\right]\frac{x}{n}} + \|s_{n,m}\|_\infty \frac{x}{3n}\right) \leq 2\mathrm{e}^{-x}.$$

Since $s_{m,n}$ belongs to S_n, which satisfies **H2**, it comes from Lemma 6.3 that

$$P_s\left[(s_{m,n} - P_s s_{m,n})^2\right] \leq \left(\|s\|_\infty \wedge C_1\|s\|\, d_n^{1/2}\right),\ \|s_{n,m}\|_\infty \leq C_1\sqrt{d_n}.$$

We conclude the Proof of Lemma 3.7 saying that $x \geq 2$ implies $2\mathrm{e}^{-x} \leq \mathrm{e}^{-x/2}$. In this Lemma, we proved that we can choose $\kappa_b(\epsilon, C_3) = \kappa_v(C_1, C_3) + 2\epsilon^{-1}(2 \vee 2C_1 \vee C_3 C_1^2/9)$.

6.5. Proof of Corollary 3.8

Let $X_1, \ldots, X_n$ be an iid sample with common density s in $B_{2,\infty}(M_2, M_\infty, 0, L^2(\mu))$. Let ϵ in $(0,1)$ and let $\Omega_n(\epsilon)$ denote the event

$$\left\{\forall m \in \mathcal{M}_n,\ \left|p_b(S_m, S_n) - \|s_n - s_m\|^2\right| \leq \epsilon\|s_n - s_m\|^2 + \kappa_b(\epsilon, C_\mathcal{M})\sqrt{\|s\|_\infty \wedge \|s\|\, d_n^{1/2}}\frac{\sqrt{d_n}x_n}{n}\right\}.$$

A union bound gives that $\mathbb{P}_s(\Omega_n(\epsilon)^c)$ is upper bounded by the sum over $\mathcal{M}_n$ of

$$\mathbb{P}_s\left(\left|p_b(S_m,S_n)-\|s_n-s_m\|^2\right|>\epsilon\|s_n-s_m\|^2+\kappa_b(\epsilon,C_{\mathcal{M}})\sqrt{\|s\|_\infty\wedge\|s\|\,d_n^{1/2}}\frac{\sqrt{d_n}x_n}{n}\right)\cdot$$

Assumption $\mathbf{H3}(\mathcal{M},\beta)$ ensures that x_n satisfies $2\le x_n\le C_3n/\sqrt{d_m}$ with $C_3=C_{\mathcal{M}}$, thus, Lemma 3.7 gives that this last probability is upper bounded by $3\mathrm{e}^{-x_n/2}$. Our choice of x_n ensures that $3N_n\mathrm{e}^{-x_n/2}\le\beta/2$ and thus that $\mathbb{P}_s(\Omega_n(\epsilon)^c)\le\frac{\beta}{2}\cdot$ The Proof of Corollary 3.8 is concluded because, on $\Omega_n(\epsilon)$,

$$(1-\epsilon)\|s_n-s_m\|^2\le p_b(S_m,S_n)+\kappa_b(\epsilon,C_{\mathcal{M}})\sqrt{\|s\|_\infty\wedge\|s\|\,d_n^{1/2}}\frac{\sqrt{d_n}x_n}{n}\cdot$$

6.6. Proof of Theorem 4.1

The theorem is a straightforward consequence of Corollaries 3.4 and 3.8.

6.7. Proof of Theorem 4.2

We begin the proof with the following proposition, which shows that $\phi_n(\alpha,\beta,S_m,S_m)\ge d_m/(12n)$. Since $\phi_n(\alpha,\beta,S_n,S_m)\ge\phi_n(\alpha,\beta,S_m,S_m)$, the same bound holds also for $\phi_n(\alpha,\beta,S_n,S_m)$.

Proposition 6.4. *Let S be the set of histograms on the partition,*

$$\left\{\left[\frac{k}{d},\frac{k+1}{d}\right),\ k=0,\dots,d-1\right\}.$$

Let $X_1,\dots,X_n$ be an i.i.d sample. Let α,β be real numbers in $(0,1)$ such that $\alpha+\beta<1$. Assume that $d\ge3+18\log(\sqrt{2}/(1-\alpha-\beta))$, then

$$\phi_n(\alpha,\beta,S,S)\ge\frac{d}{12n}\cdot$$

The proof is decomposed in two lemmas.

Lemma 6.5. *Let $\hat{B}_\beta=B_2(\hat{s},\hat{\rho}_\beta,S)$ in $CS(S,\beta)$ and let $\rho_{\alpha,\beta}$ be a real number such that*

$$\forall s\in S,\ \mathbb{P}_s\left(\hat{\rho}_\beta\le\rho_{\alpha,\beta}\right)\ge1-\alpha\cdot$$

Then,

$$\forall s\in S,\ \mathbb{P}_s\left(\|s-\hat{s}\|>\rho_{\alpha,\beta}\right)\le\alpha+\beta. \tag{6.3}$$

Proof of Lemma 6.5.

$$\begin{aligned}\mathbb{P}_s\left[\|s-\hat{s}\|>\rho_{\alpha,\beta}\right]&=\mathbb{P}_s\left[\|s-\hat{s}\|>\rho_{\alpha,\beta}\cap\rho_{\alpha,\beta}\ge\hat{\rho}_\beta\right]+\mathbb{P}_s\left[\|s-\hat{s}\|>\rho_{\alpha,\beta}\cap\rho_{\alpha,\beta}<\hat{\rho}_\beta\right]\\&\le\mathbb{P}_s\left[\|s-\hat{s}\|>\hat{\rho}_\beta\right]+\mathbb{P}_s\left[\rho_{\alpha,\beta}<\hat{\rho}_\beta\right]\le\alpha+\beta.\end{aligned}$$

□

Lemma 6.6. *Let $\delta=\alpha+\beta$ and let ρ_δ be any real number satisfying (6.3). Then we have*

$$\rho_\delta^2\ge\frac{d-1}{2n}-\frac{1}{n}\sqrt{2(d+1)\ln\left[\frac{\sqrt{1+(d+1)n^{-1}}}{1-\delta}\right]}\cdot$$

Remark: When $d \geq 3 + 18\log(\sqrt{2}/(1-\delta))$ and $n \geq d+1$, we have

$$\sqrt{2(d+1)\ln\left[\frac{\sqrt{1+(d+1)n^{-1}}}{1-\delta}\right]} \leq \frac{d-1}{3},$$

thus $\rho_\delta^2 \geq (d-1)/(6n) \geq d/(12n)$.

Proof. We prove that if

$$\rho_\delta^2 = \frac{d-1}{2n} - \frac{1}{n}\sqrt{2(d+1)\ln\left[\frac{\sqrt{1+(d+1)n^{-1}}}{1-\delta}\right]}$$

then

$$\inf_{s\in S} \mathbb{P}_s\left[\|s-\hat{s}\| \leq \rho_\delta\right] \leq 1-\delta.$$

Let $s_0 = 1_{[0,1)}$, $\Lambda = \{1,\dots,[d/2]\}$ and for all λ in Λ, let

$$\psi_\lambda = \sqrt{\frac{d}{2}}\left(1_{[2(\lambda-1)/d,(2\lambda-1)/d)} - 1_{[(2\lambda-1)/d,2\lambda/d)}\right).$$

It is easy to check that $(\psi_\lambda)_{\lambda\in\Lambda}$ is an orthonormal system in S, orthogonal to s_0 such that, for all λ in Λ, $\|\psi_\lambda\|_\infty \leq \sqrt{d/2}$. Let $\hat{s}_0 = \int \hat{s}s_0 \mathrm{d}\mu$ and for all λ in Λ, let

$$\hat{s}_\lambda = \int \hat{s}\psi_\lambda \mathrm{d}\mu.$$

Let $(\xi_\lambda)_{\lambda\in\Lambda}$ be independent Rademacher random variables, independent of $X_1,\dots,X_n$, let ρ be some real number to be chosen later and let $s_\xi = s_0 + \rho\sum_{\lambda\in\Lambda}\xi_\lambda\psi_\lambda$. The ψ_λ have distinct support, thus $\left\|\sum_{\lambda\in\Lambda}|\psi_\lambda|\right\|_\infty \leq \sqrt{d/2}$ and s_ξ is a density if

$$-\sqrt{\frac{2}{d}} \leq \rho \leq \sqrt{\frac{2}{d}}. \tag{6.4}$$

Assume that (6.4) holds, then

$$\inf_{s\in S} \mathbb{P}_s\left[\|s-\hat{s}\| \leq \rho_\delta\right] \quad\leq\quad \mathbb{P}_{s_\xi}\left[\|s_\xi-\hat{s}\| \leq \rho_\delta\right]. \tag{6.5}$$

We have

$$\|s_\xi - \hat{s}\|^2 = (1+s_0)^2 + \sum_{\lambda\in\Lambda}(\rho\xi_\lambda - \hat{s}_\lambda)^2 = \sum_{\lambda\in\Lambda,\ \rho\xi_\lambda\hat{s}_\lambda\leq 0} \rho^2 - 2\rho\xi_\lambda\hat{s}_\lambda + \hat{s}_\lambda^2 \geq \rho^2 N(\xi,\hat{s}), \tag{6.6}$$

where $N(\xi,\hat{s}) = \mathrm{Card}(\{\lambda\in\Lambda,\ \rho\xi_\lambda\hat{s}_\lambda \leq 0\}) = \sum_{\lambda\in\Lambda} 1_{\{\rho\xi_\lambda\hat{s}_\lambda\leq 0\}}$. If we plug (6.6) in (6.5), we obtain

$$\inf_{s\in S} \mathbb{P}_s\left[\|s-\hat{s}\|_2 \leq \rho_\delta\right] \leq \int_0^1 \mathbf{1}_{\rho^2 N(\xi,\hat{s})\leq\rho_\delta} s_\xi \mathrm{d}\mu.$$

We integrate with respect to ξ and we apply Fubini's theorem to obtain

$$\inf_{s\in S} \mathbb{P}_s\left[\|s-\hat{s}\|_2 \leq \rho_\delta^2\right] \leq \mathbb{P}_{s_\xi}\left[\rho^2 N(\xi,\hat{s}) \leq \rho_\delta^2\right] =\leq \int_0^1 \mathbb{E}_\xi\left(\mathbf{1}_{\rho^2 N(\xi,\hat{s})\leq\rho_\delta^2} s_\xi\right)\mathrm{d}\mu. \tag{6.7}$$

From Cauchy–Schwarz inequality,

$$\mathbb{E}_\xi^2\left(\mathbf{1}_{\rho^2 N(\xi,\hat{s})\leq\rho_\delta^2} s_\xi\right) \leq \mathbb{P}_\xi\left(\rho^2 N(\xi,\hat{s}) \leq \rho_\delta^2\right)\mathbb{E}_\xi\left(s_\xi^2\right), \tag{6.8}$$

and $\mathbb{E}_\xi s_\xi^2 = s_0^2 + \rho^2 \sum_{\lambda\in\Lambda} \psi_\lambda^2$. For all λ in Λ, $\int_0^1 \psi_\lambda^2 = 1$, thus

$$\int_0^1 \mathbb{E}_\xi s_\xi^2 \mathrm{d}\mu = 1 + \rho^2 \left[\frac{d}{2}\right]\cdot \tag{6.9}$$

Moreover, conditionally to $\hat{s}$, $N(\xi,\hat{s})$ is a sum of $[d/2]$ independent random variables valued in $\{0,1\}$. Thus, from Hoeffding's inequality,

$$\forall t > 0,\ \mathbb{P}_\xi \left(N(\xi,\hat{s}) \le \mathbb{E}_\xi \left(N(\xi,\hat{s}) \right) - \sqrt{\left[\frac{d}{2}\right] t} \right) \le \mathrm{e}^{-2t}. \tag{6.10}$$

In (6.10), we have $E_\xi \left(N(\xi,\hat{s}) \right) = \sum_{\lambda\in\Lambda} \mathbb{E}_\xi \left(\mathbf{1}_{\xi_\lambda \hat{s}_\lambda \le 0} \right) \ge [d/2]/2$ and we choose

$$t = \ln \left[\frac{\sqrt{1+\rho^2[d/2]}}{1-\delta} \right],\ \rho = \sqrt{\frac{2}{n}} \le \sqrt{\frac{2}{d}}\cdot$$

Since $(d-1)/2 \le [d/2] \le (d+1)/2$,

$$t \le \ln \left[\frac{\sqrt{1+(d+1)/n}}{1-\delta} \right],\ E_\xi \left(N(\xi,\hat{s}) \right) \ge \frac{d-1}{4}\cdot$$

Thus

$$\{\rho^2 N(\xi,\hat{s}) \le \rho_\delta^2\} \subset \left\{ N(\xi,\hat{s}) \le \mathbb{E}_\xi \left(N(\xi,\hat{s}) \right) - \sqrt{[d/2]t} \right\}.$$

Hence, from (6.10),

$$\mathbb{P}_\xi \left(\rho^2 N(\xi,\hat{s}) \le \rho_\delta^2 \right) \le \frac{(1-\delta)^2}{1+\rho^2[d/2]}\cdot \tag{6.11}$$

We plug inequalities (6.9) and (6.11) in (6.8) to obtain

$$\int_0^1 \mathbb{E}_\xi^2 \left(\mathbf{1}_{d\rho^2 N(\xi,\hat{s}) \le \rho_\delta^2} s_\xi \right) \le (1-\delta)^2.$$

Thus, from (6.7) and Jensen inequality,

$$\inf_{s\in S} \mathbb{P}_s \left[\|s-\hat{s}\|_2 \le \rho_\delta \right] \le 1-\delta.$$

We already know thanks to Proposition 6.4 that $\phi_n(\alpha,\beta,S_n,S_m) \ge d_m/(12n)$, therefore, it remains to prove that $\phi_n(\alpha,\beta,S_n,S_m) \ge \sqrt{d_n}/n$. Let $s_0 = I_{[0,1]}$, let $\hat{B}_\beta = B_2(\hat{s},\hat{\rho}_\beta,S_n)$ be a confidence ball in $CS(S_n,\beta)$ and let $\rho_{\alpha,\beta} > 0$ such that for all densities s in S_m,

$$\mathbb{P}_s \left(\hat{\rho}_\beta \le \rho_{\alpha,\beta} \right) \ge 1-\alpha.$$

We will prove that $\rho_{\alpha,\beta} \ge c\sqrt{d_n}/n$, which is sufficient to prove Theorem 4.2. We decompose the proof into two lemmas. □

Lemma 6.7. *Let $S_n(\rho_{\alpha,\beta}) = \{t \in S_n\ ;\ \|t-s_0\|_2 \ge 2\rho_{\alpha,\beta}\}$. There exists a test T of null hypothesis $H_0 : s = s_0$ against the alternative $H_1 : s \in S_n(\rho_{\alpha,\beta})$ with confidence level more than $1-\beta$ and power more than $1-\alpha-\beta$, i.e. such that*

$$\mathbb{P}_{s_0}(T=0) \ge 1-\beta,\ \inf_{s\in S_n(\rho_{\alpha,\beta})} \mathbb{P}_s(T=1) \ge 1-(\alpha+\beta).$$

Proof of Lemma 6.7. Let $T = 1_{s_0 \in \hat{B}_\beta}$. Since s_0 belongs to S_n and $\hat{B}_\beta$ belongs to $CS(S_n, \beta)$, $\mathbb{P}_{s_0}(T = 0) \geq 1 - \beta$. Moreover, for all s in $S_n(\rho_{\alpha,\beta})$,

$$\begin{aligned} \mathbb{P}_s(T = 0) &= \mathbb{P}_s(s_0 \in \hat{B}_\beta) = \mathbb{P}_s(\|s_0 - \hat{s}\| \leq \hat{\rho}_\beta) \\ &\leq \mathbb{P}_s(\|s_0 - s\| - \|s - \hat{s}\| \leq \hat{\rho}_\beta) \leq \mathbb{P}_s(\|s - \hat{s}\| \geq 2\rho_{\alpha,\beta} - \hat{\rho}_\beta). \end{aligned}$$

This last probability is equal to

$$\begin{aligned} &\mathbb{P}_s(\|s - \hat{s}\| \geq 2\rho_{\alpha,\beta} - \hat{\rho}_\beta \cap \hat{\rho}_\beta > \rho_{\alpha,\beta}) + \mathbb{P}_s(\|s - \hat{s}\| \geq 2\rho_{\alpha,\beta} - \hat{\rho}_\beta \cap \hat{\rho}_\beta \leq \rho_{\alpha,\beta}) \\ &\quad \leq \mathbb{P}_s(\hat{\rho}_\beta > \rho_{\alpha,\beta}) + \mathbb{P}_s(\|s - \hat{s}\| \geq \hat{\rho}_\beta) \leq \beta + \alpha. \end{aligned}$$

□

The second lemma gives the separation rate for the test of null hypothesis $H_0 : s = s_0$

Lemma 6.8. *Let $\eta = 2(1 - 2\alpha - \beta)$, let $\rho > 0$. Let Θ_α be the set of tests T_α with confidence level α, of null hypothesis $H_0 : s = s_0$ against the alternative $H_1 : s \in S_n(\rho)$, where $S_n(\rho)$ is the set of all densities s in S_n such that $\|s - s_0\| \geq \rho$.*

Let $\beta(S_n(\rho)) = \inf_{T_\alpha \in \Theta_\alpha} \sup_{s \in S_n(\rho)} \mathbb{P}_s(T_\alpha = 0)$.

If $d_n \geq 10$ and $\rho^2 < \sqrt{\ln(1 + \eta^2)/3.2}(\sqrt{d_n - 1}/n)$ then $\beta(S(\rho)) > \beta + \alpha$.

Comments: From Lemmas 6.7 and 6.8, we deduce that

$$\rho_{\alpha,\beta}^2 \geq \sqrt{\frac{\ln(1 + \eta^2)}{3.2}} \frac{\sqrt{d_n - 1}}{4n} \geq \frac{\sqrt{\ln(1 + \eta^2)}}{11} \frac{\sqrt{d_n}}{n}.$$

Thus the Proof of Lemma 6.8 concludes the Proof of Theorem 4.2.

Proof of Lemma 6.8.

The function $\beta(S_n(\rho))$ is non-increasing with ρ. Thus we take

$$\rho^2 = \sqrt{\ln(1 + \eta^2)/3.2}\sqrt{d_n - 1}/n$$

and we will to prove that $\beta(S_n(\rho)) \geq \alpha + \beta$. Let μ_ρ be a probability measure on $S_n(\rho)$, let $P_{\mu_\rho} = \int P_s \mathrm{d}\mu_\rho$.

$$\begin{aligned} \beta(S_n(\rho)) &\geq \inf_{T_\alpha \in \Theta_\alpha} \mathbb{P}_{\mu_\rho}(T_\alpha = 0) = \inf_{T_\alpha \in \Theta_\alpha} \left(\mathbb{P}_{\mu_\rho}(T_\alpha = 0) - \mathbb{P}_{s_0}(T_\alpha = 0) + \mathbb{P}_{s_0}(T_\alpha = 0)\right) \\ &\geq 1 - \alpha + \inf_{T_\alpha \in \Theta_\alpha} \left(\mathbb{P}_{\mu_\rho}(T_\alpha = 0) - \mathbb{P}_{s_0}(T_\alpha = 0)\right) \qquad (6.12) \\ &\geq 1 - \alpha - \sup_{A\,;\,\mathbb{P}_{s_0}(A) \leq \alpha} \left|\mathbb{P}_{\mu_\rho}(A) - \mathbb{P}_{s_0}(A)\right| \\ &\geq 1 - \alpha - 1/2 \left\|\mathbb{P}_{\mu_\rho} - \mathbb{P}_{s_0}\right\|_{TV} \qquad (6.13) \end{aligned}$$

where $\|.\|_{TV}$ denote the total variation distance. Assume that $\mathbb{P}_{\mu_\rho}$ is absolutely continuous with respect to $\mathbb{P}_{s_0}$. Let $L_{\mu_\rho} = \mathrm{d}\mathbb{P}_{\mu_\rho}/\mathrm{d}\mathbb{P}_{s_0}$, then

$$\left\|\mathbb{P}_{\mu_\rho} - \mathbb{P}_{s_0}\right\|_{TV} = \mathbb{E}_{s_0}\left|L_{\mu_\rho}(X_1, \ldots, X_n) - 1\right| \leq \left(\mathbb{P}_{s_0}\left(L_{\mu_\rho}^2\right) - 1\right)^{1/2}$$

and then

$$\beta(S_n(\rho)) \geq 1 - \alpha - \frac{\sqrt{\mathbb{E}_{s_0}\left(L_{\mu_\rho}^2\right) - 1}}{2}. \qquad (6.14)$$

From (6.14), $\beta(S_n(\rho))\geq\alpha+\beta$ if $\mathbb{E}_{s_0}\left(L^2_{\mu_\rho}\right)\leq 1+\eta^2$. Let us now give a probability measure on $S_n(\rho)$, absolutely continuous with respect to P_{s_0}, such that $\mathbb{E}_{s_0}\left(L^2_{\mu_\rho}\right)\leq 1+\eta^2$.

Let $(\psi_\lambda)_{\lambda=1,\dots,[d_n/2]}$ be the following orthonormal system. Let $\psi_0=s_0$, $\phi=1_{[0,1/2)}-1_{[1/2,1)}$ and for all $\lambda=1,\dots,[d_n/2]$, $\psi_\lambda=\sqrt{d_n/2}\phi(d_nx/2-(\lambda-1))$. Let $\xi=(\xi_\lambda)_{\lambda=1,\dots,[d_n/2]}$ be independent Rademacher random variables and let μ_ρ be the distribution of $s_\xi=s_0+\rho\sum_{\lambda=1}^{[d_n/2]}\xi_\lambda\psi_\lambda/\sqrt{[d_n/2]}$. Let us check that μ_ρ satisfies the required properties. The functions $(\psi_\lambda)_{\lambda=1,\dots,[d_n/2]}$ have distinct support, thus

$$\left\|\sum_{\lambda=1}^{[d_n/2]}|\psi_\lambda|\right\|_\infty\leq\sqrt{d_n/2}.$$

s_ξ is a real density if $\rho\leq 1$. Since $2\alpha+\beta<1$, $\eta^2\leq 4$ and $\ln(1+\eta^2)\leq\ln(5)$. $\sqrt{d_n}\leq n$, hence

$$\rho^2\leq\sqrt{\frac{\ln(5)}{3.2}}\frac{\sqrt{d_n-1}}{n}\leq 1.$$

Since $(\psi_\lambda)_{\lambda=1,..,[d_n/2]}$ is an orthonormal system, $\|s_\xi-s_0\|=\rho$, thus s_ξ belongs to $S_n(\rho)$ and μ_ρ is a law on $S_n(\rho)$. Moreover

$$\frac{\mathrm{d}\mathbb{P}_{s_\xi}}{\mathrm{d}\mathbb{P}_{s_0}}(x_1,..,x_n)=\prod_{\alpha=1}^n\left(1+\frac{\rho}{\sqrt{[d_n/2]}}\sum_{\lambda=1}^{[d_n/2]}\xi_\lambda\psi_\lambda(x_\alpha)\right).$$

Thus

$$L_{\mu_\rho}(x_1,..,x_n)=\frac{1}{2^{[d_n/2]}}\sum_{\xi\in\{-1,1\}^{[d_n/2]}}\prod_{\alpha=1}^n\left(1+\frac{\rho}{\sqrt{[d_n/2]}}\sum_{\lambda=1}^{[d_n/2]}\xi_\lambda\psi_\lambda(x_\alpha)\right).$$

Hereafter, in order to symplify the notations, we write $\sum_\xi$ instead of $\sum_{\xi\in\{-1,1\}^{[d_n/2]}}$ and $\sum_\lambda$ instead of $\sum_{\lambda=1}^{[d_n/2]}$. Let $\phi(\rho,\xi)=\rho\sum_\lambda\xi_\lambda\psi_\lambda/\sqrt{[d_n/2]}$, we have

$$\begin{aligned}L^2_{\mu_\rho}(x_1,..,x_n)&=\frac{1}{2^{2([d_n/2])}}\sum_{\xi,\xi'}\prod_{\alpha=1}^n\left(1+\phi(\rho,\xi)(x_\alpha)\right)\left(1+\phi(\rho,\xi')(x_\alpha)\right).\\ \mathbb{E}_{s_0}(L^2_{\mu_\rho})&=\frac{1}{2^{2[d_n/2]}}\sum_\xi\sum_{\xi'}\prod_{\alpha=1}^n P_{s_0}\left(1+\phi(\rho,\xi)+\phi(\rho,\xi')+\phi(\rho,\xi)\phi(\rho,\xi')\right).\end{aligned}$$

For all $\lambda\neq\lambda'=1,\dots,[d_n/2]$, $\psi_\lambda\psi_{\lambda'}=0$, thus

$$\phi(\rho,\xi)\phi(\rho,\xi')=\frac{\rho^2}{[d_n/2]}\left(\sum_\lambda\xi_\lambda\psi_\lambda\right)\left(\sum_\lambda\xi'_\lambda\psi_\lambda\right)=\frac{\rho^2}{[d_n/2]}\sum_\lambda\xi_\lambda\xi'_\lambda\psi^2_\lambda.$$

For all $\lambda = 1, \ldots, [d_n/2]$ and all $\alpha = 1, \ldots, n$, $P_{s_0}(\psi_\lambda) = 0$, $P_{s_0}(\psi_\lambda^2) = 1$, thus

$$\begin{aligned}\mathbb{E}_{s_0}(L^2_{\mu_\rho}) &\leq \frac{1}{2^{2[d_n/2]}} \sum_{\xi} \sum_{\xi'} \left(1 + \frac{\rho^2}{[d_n/2]} \sum_{\lambda} \xi_\lambda \xi'_\lambda\right)^n \\ &= \frac{1}{2^{2[d_n/2]}} \sum_{\xi} \sum_{l=0}^{[d_n/2]} \sum_{\xi';\mathrm{Card}(\lambda,\, \xi'_\lambda = \xi_\lambda) = l} \left[1 + \frac{\rho^2}{[d_n/2]}(2l - [d_n/2])\right]^n \\ &= \frac{1}{2^{[d_n/2]}} \sum_{l=0}^{[d_n/2]} C^l_{[d_n/2]} \left[1 + \frac{\rho^2 2l}{[d_n/2]} - \rho^2\right]^n .\end{aligned}$$

For all real numbers $u \geq -1$, we have $0 \leq 1 + u \leq \mathrm{e}^u$, thus $(1+u)^n \leq \mathrm{e}^{nu}$. Since $\rho^2 \leq 1$, we can apply this inequality to all the $u_l = (2l/[d_n/2] - 1)r^2$ and we obtain

$$\mathbb{E}_{s_0}(L^2_{\mu_\rho}) \leq \frac{1}{2^{[d_n/2]}} \sum_{l=0}^{[d_n/2]} C^l_{[d_n/2]} \exp\left(\frac{\rho^2 2nl}{[d_n/2]} - n\rho^2\right) = \frac{\mathrm{e}^{-n\rho^2}}{2^{[d_n/2]}} \left(\exp\left(\frac{\rho^2 2n}{[d_n/2]}\right) + 1\right)^{[d_n/2]} .$$

Thus, $\mathbb{E}_{s_0}\left(L^2_{\mu_\rho}\right) \leq 1 + \eta^2$ if

$$-n\rho^2 + ([d_n/2]) \ln\left(\frac{\exp\left(\frac{\rho^2 2n}{[d_n/2]}\right) + 1}{2}\right) \leq \ln(1+\eta^2).$$

For all positive u, $\ln(1+u) \leq u$, thus, we only have to prove that

$$-n\rho^2 + \frac{[d_n/2]}{2}\left(\exp\left(\frac{\rho^2 2n}{[d_n/2]}\right) - 1\right) \leq \ln(1+\eta^2).$$

$[d_n/2] \geq (d_n - 1)/2$ and $d_n \geq 10$, thus

$$\frac{\rho^2 2n}{[d_n/2]} = 2\sqrt{\frac{\ln(1+\eta^2)}{3.2}}\frac{\sqrt{d_n - 1}}{[d_n/2]} \leq \frac{4 * 0.71}{\sqrt{d_n - 1}} \leq 1.$$

For all real numbers x in $[0,1]$, we have $\mathrm{e}^x \leq 1 + x + 3.2x^2$, thus $\exp\left(\rho^2 2n/([d_n/2])\right) - 1 \leq \rho^2 2n/([d_n/2]) + 3.2\left(\rho^2 n/([d_n/2])\right)^2$. Hence

$$-n\rho^2 + \frac{[d_n/2]}{2}\left(\exp\left(\frac{\rho^2 2n}{[d_n/2]}\right) - 1\right) \leq 1.6\rho^4 n^2/([d_n/2]) \leq \frac{d_n - 1}{2[d_n/2]}\ln(1+\eta^2) \leq \ln(1+\eta^2).$$

Appendix A

A.1 Proof of Lemma 6.1

$\sum_{i=1}^n (W_i - \bar{W}_n) = 0$, thus, for all λ in Λ, $(P_n^W - \bar{W}_n P_n)(P_s\psi_\lambda) = 0$. Moreover, since the weights are exchangeable,

$$\begin{aligned} 0 &= \mathbb{E}\left[\left(\sum_{i=1}^n (W_i - \bar{W}_n)\right)^2\right] \\ &= \sum_{i=1}^n \mathbb{E}\left((W_i - \bar{W}_n)^2\right) + \sum_{i\neq j=1}^n \mathbb{E}(W_i - \bar{W}_n)(W_j - \bar{W}_n) \\ &= n\mathbb{E}\left((W_1 - \bar{W}_n)^2\right) + n(n-1)\mathbb{E}(W_1 - \bar{W}_n)(W_2 - \bar{W}_n). \end{aligned}$$

Thus,

$$v_W^2 = \mathbb{E}\left((W_1 - \bar{W}_n)^2\right) = -(n-1)\mathbb{E}(W_1 - \bar{W}_n)(W_2 - \bar{W}_n).$$

Hence,

$$\begin{aligned} p_W(\Lambda) &= \sum_{\lambda\in\Lambda} \frac{\mathbb{E}_W\left([(P_n^W - \bar{W}_n P_n)(\psi_\lambda)]^2\right)}{v_W^2} = \sum_{\lambda\in\Lambda} \frac{\mathbb{E}_W\left([(P_n^W - \bar{W}_n P_n)(\psi_\lambda - P_s\psi_\lambda)]^2\right)}{v_W^2} \\ &= \sum_{\lambda\in\Lambda} \mathbb{E}_W\left(\frac{1}{n^2}\sum_{i,j=1}^n \frac{(W_i - \bar{W}_n)(W_j - \bar{W}_n)}{v_W^2}(\psi_\lambda(X_i) - P_s\psi_\lambda)(\psi_\lambda(X_j) - P_s\psi_\lambda)\right) \end{aligned}$$

$$\begin{aligned} p_W(\Lambda) &= \frac{1}{n^2}\sum_{\lambda\in\Lambda}\sum_{i=1}^n \frac{\mathbb{E}\left((W_i - \bar{W}_n)^2\right)}{v_W^2}(\psi_\lambda(X_i) - P_s\psi_\lambda)^2 \\ &\quad + \frac{1}{n^2}\sum_{\lambda\in\Lambda}\sum_{i\neq j=1}^n \frac{\mathbb{E}(W_i - \bar{W}_n)(W_j - \bar{W}_n)}{v_W^2}(\psi_\lambda(X_i) - P_s\psi_\lambda)(\psi_\lambda(X_j) - P_s\psi_\lambda) \\ &= \frac{1}{n}\left(P_n T(\Lambda) - U_s(\Lambda)\right). \end{aligned} \tag{A.1}$$

On the other hand, easy algebra leads to

$$\|s_m - \hat{s}_m\|_2^2 = \sum_{\lambda\in\Lambda}\left([(P_n - P_s)(\psi_\lambda)]^2\right) = \frac{1}{n}\left(P_n T(\Lambda) + (n-1)U_s(\Lambda)\right).$$

Thus, we have $\|s_m - \hat{s}_m\|_2^2 - p_W(\Lambda) = U_s(\Lambda)$.

A.2 Proof of Lemma 6.2

We apply Theorem 3.4 in Houdré and Reynaud–Bouret [15]. For all $x > 0$

$$\mathbb{P}_s\left(\xi U(\Lambda) > \frac{1}{n^2}\left(5.7B_1\sqrt{x} + 8B_2 x + 384 B_3 x^{3/2} + 1020 B_4 x^2\right)\right) \leq e e^{-x}, \tag{A.2}$$

where

$$U(x,y) = \textstyle\sum_{\lambda\in\Lambda}(\psi_\lambda(x) - P_s\psi_\lambda)(\psi_\lambda(y) - P_s\psi_\lambda),$$
$$B_1^2 = n^2\mathbb{E}\left[(U(X_1,X_2))^2\right],\ B_3^2 = n\sup_x \mathbb{E}\left[(U(x,X_2))^2\right],\ B_4 = \sup_{x,y} U(x,y),$$

$$B_2 = \sup\left\{\left|\mathbb{E}\sum_{i=1}^{n}\sum_{j=1}^{i-1}U(X_1,X_2)\alpha_i(X_1)\beta_j(X_2)\right|,\ \mathbb{E}\sum_{i=1}^{n}\alpha_i^2(X_1)\le 1,\ \mathbb{E}\sum_{j=1}^{n}\beta_j^2(X_1)\le 1\right\}.$$

From Cauchy–Schwarz inequality, for all real numbers $(b_\lambda)_{\lambda\in\Lambda}$

$$\sum_{\lambda\in\Lambda} b_\lambda^2 = \left(\sup_{\sum a_\lambda^2\le 1}\sum_{\lambda\in\Lambda} a_\lambda b_\lambda\right)^2. \tag{A.3}$$

In particular, since the system $(\psi_\lambda)_{\lambda\in\Lambda}$ is orthonormal, for all x in $\mathbb{R}$, $T(\Lambda) = (\sup_{t\in B(\Lambda)}(t-P_s t))^2$. Thus

$$\|T(\Lambda)\|_\infty \le 2b_\lambda^2. \tag{A.4}$$

Let us now evaluate B_1, B_2, B_3 and B_4.

Evaluation of B_1:

$$\begin{aligned}\frac{B_1^2}{n^2} &= \sum_{\lambda,\lambda'\in\Lambda}\left(P_s\left((\psi_\lambda-P_s\psi_\lambda)(\psi_{\lambda'}-P_s\psi_{\lambda'})\right)\right)^2\\ &= \sum_{\lambda\in\Lambda}\left(\sup_{\sum a_{\lambda'}^2\le 1} P_s\left((\psi_\lambda-P_s\psi_\lambda)\left[\sum_{\lambda'\in\Lambda}a_{\lambda'}\psi_{\lambda'}-P_s\left(\sum_{\lambda'\in\Lambda}a_{\lambda'}\psi_{\lambda'}\right)\right]\right)\right)^2\\ &= \sum_{\lambda\in\Lambda}\left(\sup_{t\in B(\Lambda)}P_s\left((\psi_\lambda-P_s\psi_\lambda)(t-P_s t)\right)\right)^2 \le D_{s,\Lambda}v_{s,\Lambda}^2,\end{aligned}$$

where we use successively the independence of X_1 and X_2, Inequality (A.3), the orthonormality of the system $(\psi_\lambda)_{\lambda\in\Lambda}$ and Cauchy–Schwarz inequality. Thus we obtain

$$B_1 \le n v_{s,\Lambda}\sqrt{D_{s,\Lambda}}. \tag{A.5}$$

Evaluation of B_2: For all real numbers y, z, we have $2yz \le y^2+z^2$, thus, for all i, j in $\{1,\dots,n\}$,

$$2P_s\left((\psi_\lambda-P_s\psi_\lambda)\alpha_i\right)P_s\left((\psi_{\lambda'}-P_s\psi_{\lambda'})\beta_j\right) \le \left(P_s\left((\psi_\lambda-P_s\psi_\lambda)\alpha_i\right)\right)^2+\left(P_s\left((\psi_{\lambda'}-P_s\psi_{\lambda'})\beta_j\right)\right)^2.$$

We apply (A.3) with $b_\lambda = P_s\left((\psi_\lambda-P_s\psi_\lambda)\alpha_i\right)$, since the system $(\psi_\lambda)_{\lambda\in\Lambda}$ is orthonormal, for all i in $\{1,\dots,n\}$,

$$\sum_{\lambda\in\Lambda}\left(P_s\left((\psi_\lambda-P_s\psi_\lambda)\alpha_i\right)\right)^2 = \left(\sup_{t\in B(\Lambda)}P_s(t-P_s t)\alpha_i\right)^2 \le v_{s,\Lambda}^2 P_s\alpha_i^2.$$

Since $\sum_{i=1}^n P_s\alpha_i^2 \le 1$ we deduce that

$$\sum_{i,j=1}^{n}\sum_{\lambda\in\Lambda}\left(P_s\left((\psi_\lambda-P_s\psi_\lambda)\alpha_i\right)\right)^2 \le n v_{s,\Lambda}^2.$$

The same inequality holds for β_j, thus we obtain

$$B_2 \le n v_{s,\Lambda}^2. \tag{A.6}$$

Evaluation of B_3: For all x in $\mathbb{R}$, $\mathbb{E}[(U(x,X_2))^2]$ is the variance of the function $t_x = \sum_{\lambda\in\Lambda}(\psi_\lambda(x) - P_s\psi_\lambda)\psi_\lambda$. t_x is a function in the linear space S spanned by the $(\psi_\lambda)_{\lambda\in\Lambda}$ and, from inequality (A.3),

$$\|t_x\|_2^2 = \sum_{\lambda\in\Lambda}(\psi_\lambda(x) - P_s\psi_\lambda)^2 = \left(\sup_{t\in B(\Lambda)}(t(x) - P_s t)\right)^2 \le 2b_\Lambda^2.$$

Thus $\mathbb{E}[(U(x,X_2))^2] = \mathrm{Var}(t_x(X)) = 2b_\Lambda^2 \mathrm{Var}(t_x(X)/b_\Lambda) \le 2b_\Lambda^2 v_{s,\Lambda}^2$. Thus

$$B_3 \le \sqrt{2n} b_\Lambda v_{s,\Lambda}. \tag{A.7}$$

Evaluation of B_4: We apply Cauchy–Schwarz inequality and we obtain

$$B_4 \le \|T(\Lambda)\|_\infty \le 2b_\Lambda^2. \tag{A.8}$$

Let Ω_x^c be the event defined by inequality (A.2). From (A.5)–(A.8). On Ω_x,

$$\xi U_s(\Lambda) \le \frac{5.7 v_{s,\Lambda}\sqrt{D_{s,\Lambda}x}}{n} + \frac{8v_{s,\Lambda}^2 x}{n} + 384\sqrt{2}v_{s,\Lambda}b_\Lambda\left(\frac{x}{n}\right)^{3/2} + 2040 b_\Lambda\left(\frac{x}{n}\right)^2.$$

A.3 Proof of Lemma 6.3

It comes from Assumption **H2** that

$$b_\Lambda \le C_1\sqrt{d}.$$

It comes from (A.3) that

$$D_{s,\Lambda} \le \sum_{\lambda\in\Lambda} P_s(\psi_\lambda^2) = P_s\left[\left(\sup_{t\in B(\Lambda)} t\right)^2\right] \le \left\|\sup_{t\in B(\Lambda)} t\right\|_\infty^2 \le C_1^2 d.$$

$v_{s,\Lambda}^2 \le \sup_{t\in B(\Lambda)} P_s t^2$, thus

$$v_{s,\Lambda}^2 \le b_\Lambda^2 \le C_1^2 d,\ v_{s,\Lambda}^2 \le \|s\|_\infty \sup_{t\in B(\Lambda)} \|t\|^2 = \|s\|_\infty.$$

Finally, for all t in $B(\Lambda)$,

$$P_s t^2 \le \|t\|_\infty P_s|t| \le \|t\|_\infty \|t\| \|s\| \le C_1\sqrt{d}\|s\|.$$

Thus $v_{s,\Lambda}^2 \le C_1\sqrt{d}\|s\|$.

Acknowledgements. The author would like to thank gratefully Béatrice Laurent and Clémentine Prieur for many fruitful advices. He also would like to thank the reviewers and the associated editors who helped to improve a first version of the article.

References

[1] S. Arlot, Model selection by resampling penalization. *Electron. J. Statist.* **3** (2009) 557–624.
[2] S. Arlot and P. Massart, Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.* **10** (2009) 245–279.
[3] S. Arlot, G. Blanchard and E. Roquain, Resampling-based confidence regions and multiple tests for a correlated random vector, in Learning theory. *Lect. Notes Comput. Sci.* **4539** (2007) 127–141.
[4] Y. Baraud, Confidence balls in Gaussian regression. *Ann. Statist.* **32** (2004) 528–551.
[5] R. Beran, REACT scatterplot smoothers: superefficiency through basis economy. *J. Amer. Statist. Assoc.* **95** (2000) 155–171.

[6] R. Beran and L. Dümbgen, Modulation of estimators and confidence sets. *Ann. Statist.* **26** (1998) 1826–1856.

[7] L. Birgé and P. Massart, From model selection to adaptive estimation, in *Festschrift for Lucien Le Cam*. Springer, New York (1997) 55–87.

[8] L. Birgé and P. Massart, Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields* **138** (2007) 33–73.

[9] T. Cai and M.G. Low, Adaptive confidence balls. *Ann. Statist.* **34** (2006) 202–228.

[10] B. Efron, Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7** (1979) 1–26.

[11] M. Fromont and B. Laurent, Adaptive goodness-of-fit tests in a density model. *Ann. Statist.* **34** (2006) 680–720.

[12] C.R. Genovese and L. Wasserman, Confidence sets for nonparametric wavelet regression. *Ann. Statist.* **33** (2005) 698–729.

[13] C. Genovese and L. Wasserman, Adaptive confidence bands. *Ann. Statist.* **36** (2008) 875–905.

[14] M. Hoffmann and O. Lepski, Random rates in anisotropic regression. *Ann. Statist.* **30** (2002) 325–396. With discussions and a rejoinder by the authors.

[15] C. Houdré and P. Reynaud-Bouret, Exponential inequalities, with constants, for U-statistics of order two, in Stochastic inequalities and applications. *Progr. Probab.* **56** (2003) 55–69.

[16] Y.I. Ingster, Asymptotically minimax hypothesis testing for nonparametric alternatives. I. *Math. Methods Stat.* **2** (1993) 85–114.

[17] Y.I. Ingster, Asymptotically minimax hypothesis testing for nonparametric alternatives. II. *Math. Methods Stat.* **2** (1993) 171–189.

[18] Y.I. Ingster, Asymptotically minimax hypothesis testing for nonparametric alternatives. III. *Math. Methods Stat.* **2** (1993) 249–268.

[19] A. Juditsky and S. Lambert-Lacroix, Nonparametric confidence set estimation. *Math. Methods Stat.* **12** (2003) 410–428.

[20] A. Juditsky and O. Lepski, Evaluation of the accuracy of nonparametric estimators. *Math. Methods Stat.* **10** (2001) 422–445. *Meeting on Mathematical Statistics,* Marseille (2000).

[21] B. Laurent, Estimation of integral functionnals of a density. *Ann. Statist.* **24** (1996) 659–681.

[22] B. Laurent, Adaptive estimation of a quadratic functional of a density by model selection. *ESAIM: PS* **9** (2005) 1–18 (electronic).

[23] O.V. Lepski, How to improve the accuracy of estimation. *Math. Methods Stat.* **8** (1999) 441–486.

[24] M. Lerasle, *Optimal model selection in density estimation.* Preprint (2009).

[25] K.C. Li, Honest confidence regions for nonparametric regression. *Ann. Statist.* **17** (1989) 1001–1008.

[26] M.G. Low, On nonparametric confidence intervals. *Ann. Statist.* **25** (1997) 2547–2554.

[27] P. Massart, Concentration inequalities and model selection. Springer, Berlin. *Lect. Notes Math.* **1896** (2007). *Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour* (2003). With a foreword by Jean Picard.

[28] J. Robins and A. van der Vaart, Adaptive nonparametric confidence sets. *Ann. Statist.* **34** (2006) 229–253.