OPTIMAL ADAPTIVE ESTIMATION OF THE RELATIVE DENSITY

GAËLLE CHAGNY$^{(A)}$ AND CLAIRE LACOUR$^{(B)}$

Abstract. This paper deals with the classical statistical problem of comparing the probability distributions of two real random variables X and X_0, from a double independent sample. While most of the usual tools are based on the cumulative distribution function F and F_0 of the variables, we focus on the relative density, a function recently used in two-sample problems, and defined as the density of the variable $F_0(X)$. We provide a nonparametric adaptive strategy to estimate the target function. We first define a collection of estimates using a projection on the trigonometric basis and a preliminar estimator of F_0. An estimator is selected among this collection of projection estimates, with a criterion in the spirit of the Goldenshluger-Lepski methodology. We show the optimality of the procedure both in the oracle and the minimax sense: the convergence rate for the risk computed from an oracle inequality matches with the lower bound, that we also derived. Finally, some simulations illustrate the method.

Keywords: Nonparametric estimation, model selection, relative density, two-sample problem.

AMS Subject Classification 2010: 62G05; 62G07; 62G30.

1. Introduction

1.1. **Statistical model.**

The study of differences among groups is the main challenge of two-sample problems, and statistical methods are required to do this in various fields (biology or social research for example). Nonparametric inference procedures are well-developed for comparing samples coming from two populations, modeled by two real random variables X_0 and X. Most of the methods are based on the comparison of the cumulative distribution functions (c.d.f. in the sequel) F_0 and F of X_0 and X respectively. The study of the *relative density* r of X with respect to X_0 is quite recent. Assume that f_0, the density of X_0, is defined on an interval A_0 and does not vanish on it. Denote by F_0^{-1} the inverse of F_0. The relative density is defined as the density of the variable $F_0(X)$ and can be expressed as

$$r(x) = \frac{f \circ F_0^{-1}(x)}{f_0 \circ F_0^{-1}(x)}, \quad x \in F_0(A), \tag{1}$$

where $\circ$ is the composition symbol, f is a density of X, defined on an interval $A \subset \mathbb{R}$. In the present work, we focus on the optimal adaptive estimation of this function (in the oracle and minimax senses), from two independent samples $(X_i)_{i\in\{1,\dots,n\}}$ and $(X_{0,i_0})_{i_0\in\{1,\dots,n_0\}}$ of variables X and X_0.

1.2. **Motivation.**

The most classical nonparametric methods to tackle the initial issue of the comparison of F and F_0 are statistical tests such as Kolmogorov and Smirnov (Kolmogorov, 1933, 1941; Smirnov, 1939, 1944), Wilcoxon (Wilcoxon, 1945), or Mann and Whitney tests (Mann and Whitney, 1947), which all propose to check the null hypothesis of equal c.d.f.. We refer to Gibbons and Chakraborti (2011) for a detailed review of these tests. Probability plotting tools such as

Date: November 14, 2014.

(A) LMRS, UMR CNRS 6085, Université de Rouen, France. gaelle.chagny@gmail.com

(B) Laboratoire de Mathématiques d'Orsay, Université Paris-Sud, France. claire.lacour@u-psud.fr.

quantile-quantile plots, whose functional form is $x \mapsto F_0^{-1}(F(x))$, are also commonly considered. However, the representation of the quantiles of one distribution versus the quantiles of the other may be questionable. For example, Holmgren (1995) showed that it does not enable scale-invariant comparisons of treatment effects and that it depends on outliers. Some authors have thus been interested by an alternative, the probability-probability plot, a graph of the percentiles of one distribution versus the percentiles of the other (see among all Li et al. 1996). The functional form can be written $x \mapsto F(F_0^{-1}(x))$, which defines the *relative c.d.f.*, a function closely related to the Receiver Operating Characteristic (ROC) curve: the latter is $x \mapsto 1 - F(F_0^{-1}(1-x))$. This curve is well-known in fields such as signal detection and diagnostic test for example. Both the relative c.d.f. and the ROC curve are based on the following transformation of the data: to compare X to X_0, consider $F_0(X)$, a variable known in the literature as the *grade transformation* or most commonly as the *relative transformation*. Its c.d.f. is the relative c.d.f. defined above. The basic idea is to look at the rank that a comparison value (that is a value of X) would have in the reference group (that is in the values of the sample of X_0). To recover from a double-sample the ROC curve or the relative c.d.f. in a nonparametric way, two types of strategies have mainly been studied: estimators based on the empirical c.d.f. of X and X_0 (see Hsieh and Turnbull 1996a,b and references therein), as well as kernel smoothers (see among all Lloyd 1998; Lloyd and Yong 1999; Hall and Hyndman 2003 for the ROC curve, Gastwirth 1968; Hsieh 1995; Handcock and Morris 1999 for the relative c.d.f.). Conditional version of the previous strategies have also been studied (see the review provided by Pardo-Fernández et al. 2013). These two functions are based on the c.d.f. F and F_0 of the two variables to compare.

Nevertheless, focusing on their densities is likely to provide more precised and visual details. That is why the present work addresses the problem of comparison through the estimation of the relative density (1), which is the derivative of the relative c.d.f., and thus a density of the variable $F_0(X)$. Graphically more informative than the ROC curve (see the introduction of Molanes-López and Cao 2008b), another reason for the choice of the relative density is that an estimate of this function is required to study the asymptotic variance of any ROC curve estimator, and thus to build confidence regions based on it (see the references above, and also Claeskens et al. 2003). Moreover, some summary measures for the comparison of X and X_0 are based on the relative density r: the most classical example is the Kullback-Leibler divergence (Kullback and Leibler, 1951) which can be recovered by the plug-in of an estimate of r (Mielniczuk, 1992; Handcock and Morris, 1999). But there exist other measures that can pertain to the relative density, such as the Gini separation measurement and some discriminant rules (Gijbels and Mielniczuk, 1995), Lorenz curves and the median polarization index (Handcock and Morris, 1999). It is also possible to build goodness-of-fit tests from the relative density, see Kim (2000).

However, not many investigations are concerned with theoretical results for the estimation of the relative density, and most of the references are sociological ones. A clear account is provided by Handcock and Janssen (2002). Early mathematical references for the relative density are Bell and Doksum (1966) and Silverman (1978), who approached the problem with the maximum likelihood point of view. A kernel estimate was first proposed by Ćwik and Mielniczuk (1993), and modified by Molanes-López and Cao (2008a) who proved asymptotic developments for the Mean Integrated Squared Error (MISE), under the assumption that r is twice continuously derivable. The problem of bandwidth selection is also addressed, but few theoretical results are proved for the estimators with the selected parameters, to the best of our knowledge. The question has also been studied in a semiparametric setting (see Cheng and Chu 2004 and references therein). If the relative density can also be brought closer to the density-ratio, for which numerous studies are available (see Sugiyama et al. 2012 for a review), some authors have noticed that the relative distribution leads to smoother and more stable results (Yamada et al., 2013). Our work is the first to study a nonparametric projection method in this setting, and provide a detailed optimal study of an adaptive estimator.

1.3. **Contribution and overview.** Our main contribution is a theoretical one. The novelty of our work is to provide a theoretically-justified adaptive estimator with optimal rate of convergence. A collection of projection estimators on linear models is built in Section 2, and the quadratic risk is studied: the upper-bound is non-trivial, and requires non-straightforward splittings. We obtain a bias-variance decomposition which permits to understand what we can expect at best from adaptive estimation, which is the subject of Section 3: the model selection is automatically performed in the spirit of the Goldenshluger-Lepski method in a data-driven way (Goldenshluger and Lepski, 2011). The resulting estimator is shown to be optimal in the collection, but also, from an asymptotic point of view, among all possible estimators for a large class of regular relative density. To be more precise, an oracle-type inequality first proves that adaptation has no cost (Section 3.2): the estimator achieves the same performance as the one which would have been selected if the regularity index of the target function has been known. The choice of the quadratic risk permits to use the Hilbert structure and thus the standard model selection tools (mainly concentration inequalities) even if our selection criterion is based on the Goldenshluger-Lepski methodology. Rates of convergence are deduced, for functions r belonging to Besov balls: we obtain the nonparametric rate $(n^{-1}+n_0^{-1})^{2\alpha/(2\alpha+1)}$, where α is the smoothness index of r. These rates are also shown to be optimal: a lower bound for the minimax risk is established (Section 3.3). Such results are new for this estimation problem. Especially, no assumption about a link between the sample sizes n and n_0 is required, and the regularity assumptions are not restrictive. Section 4 provides a brief discussion of some practical issues via simulations. Finally, the proofs are gathered in Section 5. The supplementary material Chagny and Lacour (2014) is available with further simulation results (reconstructions and risk computations), as well as further details about technical definitions and proofs.

2. The collection of projection estimators

For the sake of clarity, we assume that the variables X and X_0 have the same support: $A = A_0$. Hence, $F_0(A) = (0;1)$ is the estimation interval. This assumption is natural to compare the distribution of X to the one of X_0.

2.1. **Approximation spaces.** We denote by $L^2((0;1))$, the space of square integrable functions on $(0;1)$, equipped with its usual Hilbert structure: $\langle .,. \rangle$ is the scalar product, and $\|.\|$ the associated norm. The relative density r, defined by (1) and estimated on its definition set $(0;1)$ is assumed to belong to $L^2((0;1))$. Our estimation method is based on this device: we consider a family S_m, $m \in \mathcal{M}$ of finite dimensional subspaces of $L^2((0;1))$ and compute a collection of estimators $(\hat{r}_m)_{m\in\mathcal{M}}$, where, for all m, $\hat{r}_m$ belongs to S_m. In a second step a data driven procedure chooses among the collection the final estimator $\hat{r}_{\hat{m}}$.

Here, simple projection trigonometric spaces are considered: the set S_m is linearly spanned by $\varphi_1, \ldots, \varphi_{2m+1}$, with

$$\varphi_1(x) = 1, \;\; \varphi_{2j}(x) = \sqrt{2}\cos(2\pi jx), \;\; \varphi_{2j+1}(x) = \sqrt{2}\sin(2\pi jx), \; x \in (0;1).$$

We set $D_m = 2m+1$, the dimension of S_m, and $\mathcal{M} = \{1, 2, \ldots, \lfloor \min(n, n_0)/2 \rfloor - 1\}$, the collection of indices, whose cardinality depends on the two sample-sizes. The largest space in the collection has maximal dimension $D_{m_{\max}}$, which is subject to constraints appearing later. We focus on the trigonometric basis mainly for its simplicity to be handled. It is also used for a lot of other nonparametric estimation problems, by several authors (see e.g. Efromovich 1999 among all). Moreover, the presence of a constant function (namely φ_1) in the basis is perfectly well-adapted to the relative density estimation context, see Section 4.2 below. The method may however probably be extended to other projection spaces, thanks to different "tricks" in the computations.

2.2. **Estimation on a fixed model.** For each index $m \in \mathcal{M}$, we define an estimator for the orthogonal projection $r_m = \sum_{j=1}^{D_m} a_j \varphi_j$ of r onto the model S_m, where $a_j = \langle \varphi_j, r \rangle$. First notice that

$$\mathbb{E}\left[\varphi_j(F_0(X))\right] = \int_A \varphi_j \circ F_0(x) f(x) dx = \int_{F_0(A)} \varphi_j(u) \frac{f \circ F_0^{-1}(u)}{f_0 \circ F_0^{-1}(u)} du = \langle \varphi_j, r \rangle = a_j, \tag{2}$$

with the change of variables $u = F_0(x)$, and keeping in mind that $F_0(A) = (0;1)$. Thus, the following function suits well to estimate r_m:

$$\hat{r}_m(x) = \sum_{j=1}^{D_m} \hat{a}_j \varphi_j(x), \text{ with } \hat{a}_j = \frac{1}{n} \sum_{i=1}^{n} \varphi_j\left(\hat{F}_0(X_i)\right), \tag{3}$$

and where $\hat{F}_0$ is the empirical c.d.f. of the sample $(X_{0,i_0})_{i_0=1,\ldots,n_0}$, that is

$$\hat{F}_0 : x \mapsto \frac{1}{n_0} \sum_{i_0=1}^{n_0} \mathbf{1}_{X_{0,i_0} \leq x}.$$

Note that in the "toy" case of known c.d.f. F_0, the procedure amounts to estimate a density : $\hat{r}_m$ is the classical density projection estimator (adapted to the estimation of the density of $F_0(X)$).

Remark 1. Comparison with other estimation methods.

(1) The estimator $\hat{r}_m$ defined in (3) can also be seen as a minimum of contrast estimate: $\hat{r}_m = \arg\inf_{t \in S_m} \gamma_n(t, \hat{F}_0)$, $m \in \mathcal{M}$, with

$$\gamma_n(t, \hat{F}_0) = \|t\|^2 - \frac{2}{n} \sum_{i=1}^{n} t \circ \hat{F}_0(X_i).$$

(2) It is worthwhile to draw a parallel between the projection method and the kernel estimator of Ćwik and Mielniczuk (1993) or Molanes-López and Cao (2008a). Thanks to the properties of the sine-cosine basis,

$$\hat{r}_m(x) = \frac{2}{n} \sum_{i=1}^{n} \sum_{j=0}^{(D_m-1)/2} \cos\left(2\pi j \left(\hat{F}_0(X_i) - x\right)\right).$$

Heuristically, by setting $(D_m - 1)/2 = \lfloor 1/(2\pi h) \rfloor - 1$, $h > 0$, the previous expression shows that $\hat{r}_m$ can be seen as an approximation of

$$\begin{aligned} \tilde{r}_h(x) &= \frac{2}{n} \sum_{i=1}^{n} \int_0^{1/(2\pi h)} \cos\left(2\pi u \left(\hat{F}_0(X_i) - x\right)\right) du, \\ &= \frac{1}{2\pi n} \sum_{i=1}^{n} \int_{-1/h}^{1/h} \cos\left(u \left(\hat{F}_0(X_i) - x\right)\right) du, \\ &= \frac{1}{2\pi n} \sum_{i=1}^{n} \int_{-1/h}^{1/h} \exp\left(-iu\left(x - \hat{F}_0(X_i)\right)\right) du = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - \hat{F}_0(X_i)}{h}\right), \end{aligned}$$

with K the sinus cardinal kernel defined by its Fourier transform: $\mathcal{F}(K)(x) = 1$ if $x \in (0;1)$, $\mathcal{F}(K)(x) = 0$ otherwise. Our strategy thus seems to be close to the kernel estimators of Ćwik and Mielniczuk (1993) and Molanes-López and Cao (2008a). But contrary to them, the projection method makes possible to obtain an unbiased estimate when the target function belongs to one of the approximation spaces. In the relative density estimation setting, this can occur if the two variables X and X_0 have the same distribution, and if the constant functions are included in one of the models, which is the case.

2.3. **Risk of a projection estimator.** The global squared error is the natural criterion associated to the projection estimation procedure. First consider the toy case of known c.d.f. F_0. The Pythagoras theorem simply leads to the classical bias-variance decomposition:

$$\|r - \hat{r}_m\|^2 = \|r - r_m\|^2 + \|\hat{r}_m - r_m\|^2 . \tag{4}$$

Moreover, the variance term can be easily bounded, still with known F_0, and using the property of the trigonometric basis:

$$\mathbb{E}\left[\|\hat{r}_m - r_m\|^2\right] = \sum_{j=1}^{D_m} \mathrm{Var}\left(\hat{a}_j\right) \leq \frac{1}{n}\sum_{j=1}^{D_m} \mathbb{E}\left[\varphi_j^2(F_0(X_1))\right] = \frac{D_m}{n}. \tag{5}$$

The challenge in the general case comes from the plug-in of the empirical $\hat{F}_0$. It seems natural but involves non straightforward computations. This is why the proof of the following upper-bound for the risk is postponed to Section 5.

Proposition 1. *Assume that the relative density r is continuously differentiable on $(0;1)$. Assume also that $D_m \leq \kappa n_0^{1/3}$, for a constant $\kappa > 0$. Then, there exist two constants c_1 and c_2 such that,*

$$\mathbb{E}\left[\|\hat{r}_m - r\|^2\right] \;\leq\; 3\,\|r - r_m\|^2 + \left(3\frac{D_m}{n} + c_1\|r\|^2\frac{D_m}{n_0}\right) + c_2\left(\frac{1}{n} + \frac{1}{n_0}\right). \tag{6}$$

The constants c_1 and c_2 do not depend on n, n_0 and m. Moreover, c_1 also does not depend on r.

The assumption on the model dimension D_m comes from the control of the deviations between $\hat{F}_0$ and F_0. Proposition 1 shows that the risk is divided into three terms: a squared-bias term, a variance term (proportional to $D_m(n^{-1} + n_0^{-1})$) and a remainder (proportional to $(n^{-1} + n_0^{-1})$). The upper bound of (1) is non trivial, and the proof requires tricky approximations (see Section 5.2 e.g.).

2.4. **Rates of convergence on Besov balls.** The result (6) also gives the asymptotic rate for an estimator if we consider that r has smoothness $\alpha > 0$. Indeed, in this case, it is possible to calculate the approximation error $\|r - r_m\|$. A space of functions with smoothness α which has good approximation properties is the Besov space $\mathcal{B}_{2,\infty}^{\alpha}$, where index 2 refers to the L^2 norm. This space is somehow a generalization of the Sobolev space and is known to be optimal for nonparametric estimation (Kerkyacharian and Picard, 1993). More precisely, we assume that the relative density r belongs to a Besov ball $B_{2,\infty}^{\alpha}((0;1), L)$ of radius L, for the Besov norm $\|.\|_{\alpha,2}$ on the Besov space $\mathcal{B}_{2,\infty}^{\alpha}((0;1))$. A precise definition is recalled in the supplementary material (Section 1 of Chagny and Lacour 2014), see also DeVore and Lorentz (1993). The following rate is then obtained.

Corollary 1. *Assume that the relative density r belongs to the Besov ball $B_{2,\infty}^{\alpha}((0;1), L)$, for $L > 0$, and $\alpha \geq 1$. Choose a model m_{n,n_0} such that $D_{m_{n,n_0}} = C(n^{-1} + n_0^{-1})^{-1/(2\alpha+1)}$, for $C > 0$.*

Then, under the assumptions of Proposition 1, there exists a numerical constant C' such that

$$\mathbb{E}\left[\left\|\hat{r}_{m_{n,n_0}} - r\right\|^2\right] \leq C'\left(\frac{1}{n} + \frac{1}{n_0}\right)^{\frac{2\alpha}{2\alpha+1}}.$$

This inequality is a straightforward consequence of the result of DeVore and Lorentz (1993) and of Lemma 12 of Barron et al. (1999), which imply that the bias term $\|r - r_m\|^2$ is of order $D_m^{-2\alpha}$. The minimum of the right-hand side term of (6) can thus be computed, leading to Corollary 1. Nevertheless, it is worth noticing that the rate depends on the two sample sizes n and n_0. Heuristically, it is $(\min(n, n_0))^{-2\alpha/(2\alpha+1)}$. The rate we obtain is new in nonparametric estimation, but it is not surprising. Actually, it looks like the Kolmogorov-Smirnov two-sample test convergence result: it is well-known that the test statistic rate is $\sqrt{nn_0/(n+n_0)}$ (see for example Doob 1949). More recently, similar rates have been obtained in adaptive minimax testing (see e.g. Butucea and Tribouley 2006).

Remark 2. The regularity condition $\alpha \geq 1$ ensures that there exists a dimension $D_{m_{n,n_0}}$ which satisfies $D_m \leq Cn_0^{1/3}$ while being of order $(n^{-1} + n_0^{-1})^{-1/(2\alpha+1)}$. When $\alpha < 1$, this choice remains possible and the convergence rate is preserved under the additional assumption $n \leq n_0/(n_0^{(2-2\alpha)/3} - 1)$. Roughly, this condition means that $n \leq n_0^{(2\alpha+1)/3} < n_0$, and thus n and n_0 must be put in order to handle this case.

It follows from Corollary 1 that the optimal dimension depends on the unknown regularity α of the function to be estimated. The aim is to perform an adaptive selection only based on the data.

3. Adaptive optimal estimation

3.1. **Model selection.** Consider the collection $(S_m)_{m\in\mathcal{M}}$ of models defined in Section 2.1 and the collection $(\hat{r}_m)_{m\in\mathcal{M}}$ of estimators defined by (3). The aim is to propose a data driven choice of m leading to an estimator with risk near of the squared-bias/variance compromise (see (6)). The selection combines two strategies: the model selection device performed with a penalization of the contrast (see e.g. Barron et al. 1999), and the recent Goldenshluger-Lepski method (Goldenshluger and Lepski, 2011). A similar device has already been used in Comte and Johannes (2012), Bertin et al. (2013) and Chagny (2013). We set, for every index m,

$$\begin{aligned} V(m) &= c_0\left(\frac{D_m}{n} + \|r\|^2\frac{D_m}{n_0}\right), \\ A(m) &= \max_{m'\in\mathcal{M}}\left(\|\hat{r}_{m'} - \hat{r}_{m\wedge m'}\|^2 - V(m')\right)_+, \end{aligned} \tag{7}$$

where $m \wedge m'$ is the minimum between m and m', $(x)_+$ the maximum between x and 0 (for a real number x), and c_0 a tuning parameter. The quantity V must be understood as a penalty term, and A is an approximation of the squared-bias term (see Lemma 10). The estimator of r is now given by $\hat{r}_{\hat{m}}$, with

$$\hat{m} = \operatorname{argmin}_{m\in\mathcal{M}}\{A(m) + V(m)\}.$$

By construction, the choice of the index m, and hence the estimator $\hat{r}_{\hat{m}}$ does not depend on the regularity assumption on the relative density r.

3.2. **Optimality in the oracle sense.** A non-asymptotic upper-bound is derived for the risk of the estimator $\hat{r}_{\hat{m}}$.

Theorem 2. *Assume that the relative density r is continuously differentiable on $(0;1)$. Assume also that $D_m \leq \kappa n_0^{1/3}/\ln^{2/3}(n_0)$, for a constant $\kappa > 0$. Then, there exist two constants c and C such that*

$$\mathbb{E}\left[\|\hat{r}_{\hat{m}} - r\|^2\right] \quad \leq \quad c \min_{m\in\mathcal{M}}\left\{\left(\frac{D_m}{n} + \|r\|^2\frac{D_m}{n_0}\right) + \|r_m - r\|^2\right\} + C\left(\frac{1}{n} + \frac{1}{n_0}\right). \tag{8}$$

The constant c is purely numerical, while C depends on r, but neither on n nor n_0.

Theorem 2 establishes the optimality of the selection rule in the oracle sense. For every index $m \in \mathcal{M}$, $\{(D_m/n+\|r\|^2 D_m/n_0)+\|r_m-r\|^2\}$ has the same order as $\mathbb{E}\left[\|\hat{r}_m - r\|^2\right]$ (see Proposition 1). Thus, Inequality (8) indicates that up to a multiplicative constant, the estimator $\hat{r}_{\hat{m}}$ converges as fast as the best estimator in the collection. The proof of such result is based on the following scheme: we first come down to the case of a known c.d.f. F_0, by using deviation results for the empirical c.d.f. Then, we use concentration results for empirical processes to prove that $A(m)$ defined in (7) is a good estimate of the bias term.

The following corollary states the convergence rate of the risk over Besov balls. Since the regularity parameter defining the functional class is not supposed to be known to select the estimator $\hat{r}_{\hat{m}}$, it is an adaptation result: the estimator adapts to the unknown regularity α of the function r.

Corollary 2. *Assume that the relative density r belongs to $B^{\alpha}_{2,\infty}((0;1),L)$, for $L > 0$, and $\alpha \geq 1$. Under the assumptions of Theorem 2,*

$$\mathbb{E}\left[\|\hat{r}_{\hat{m}} - r\|^2\right] \leq C\left(\frac{1}{n} + \frac{1}{n_0}\right)^{\frac{2\alpha}{2\alpha+1}}.$$

It is worth noticing that the rate of convergence computed above (that is the one of the best estimator among the collection, see Corollary 1), is automatically achieved by the estimator $\hat{r}_{\hat{m}}$. Corollary 2 is established with regularity assumptions stated on the target function r only. To the best of our knowledge, in the previous works, convergence results for selected relative density estimators (among a family of kernel ones) depended on strong assumptions on r ($r \in \mathcal{C}^6((0;1))$ *e.g.*) but also on the regularity of f_0.

The penalty term V given in (7) cannot be used in practice, since it depends on the unknown quantity $\|r\|^2$. A solution is to replace it by an estimator, and to prove that the estimator of r built with this random penalty keeps the adaptation property. To that aim, set, for an index $m^* \in \mathcal{M}$,

$$\begin{aligned} \widetilde{V}(m) &= c_0\left(\frac{D_m}{n} + 4\|\hat{r}_{m^*}\|^2\frac{D_m}{n_0}\right), \\ \widetilde{A}(m) &= \max_{m'\in\mathcal{M}}\left(\|\hat{r}_{m'} - \hat{r}_{m\wedge m'}\|^2 - \widetilde{V}(m')\right)_+, \end{aligned} \tag{9}$$

and $\tilde{m} = \operatorname{argmin}_{m\in\mathcal{M}}\{\widetilde{A}(m) + \widetilde{V}(m)\}$. The result for $\hat{r}_{\tilde{m}}$ is described in the following theorem.

Theorem 3. *Assume that the assumptions of Theorem 2 are satisfied, and that r belongs to $B^{\alpha}_{2,\infty}((0;1),L)$, for $L > 0$, and $\alpha \geq 1$. Choose m^* in the definition of $\widetilde{V}$ such that $D_{m^*} \geq \ln(n_0)$ and $D_{m^*} = O(n^{1/4}/\ln^{1/4}(n))$. Then, for n_0 large enough, there exist two constants c and C such that,*

$$\mathbb{E}\left[\|\hat{r}_{\tilde{m}} - r\|^2\right] \quad \leq \quad c \min_{m\in\mathcal{M}}\left\{\left(\frac{D_m}{n} + \|r\|^2\frac{D_m}{n_0}\right) + \|r_m - r\|^2\right\} + C\left(\frac{1}{n} + \frac{1}{n_0}\right).$$

As for Theorem 2, the result proves that the selection rule leads to the best trade-off between a bias and a variance term. Our estimation procedure is thus optimal in the oracle sense. The convergence rates derived in Corollary 2 remain valid for $\hat{r}_{\tilde{m}}$. Now, the only remaining parameter to tune is the constant c_0 involved in the definition of $\widetilde{V}$. A value is obtained in the proof, but it is quite rough and useless in practice. A sharp bound seems difficult to obtain from a theoretical point of view: obtaining minimal penalties is still a difficult problem (see *e.g.* Birgé and Massart 2007), and this question could be the subject of a full paper. Therefore, we experiment the tuning by a simulation study over various models.

3.3. **Optimality in the minimax sense.** Until now, we have drawn conclusions about the performance of the selected estimators $\hat{r}_{\hat{m}}$ or $\hat{r}_{\tilde{m}}$ within the collection $(\hat{r}_m)_{m\in\mathcal{M}}$ of projection estimators. A natural question follows: is the convergence rate obtained in Corollary 2 optimal among all the possible estimation strategies? We prove that the answer is yes by establishing the following lower bound for the minimax risk of the relative density estimation problem, without making any assumption.

Theorem 4. *Let $\mathcal{F}_\alpha$ be the set of relative density functions on $(0;1)$ which belong to the Besov ball $B^\alpha_{2,\infty}((0;1),L)$, for a fixed radius $L>1$, and for $\alpha\geq 1$. Then there exists a constant $c>0$ which depends on (α,L) such that*

$$\inf_{\hat{r}_{n,n_0}}\sup_{r\in\mathcal{F}_\alpha}\mathbb{E}\left[\|\hat{r}_{n,n_0}-r\|^2\right]\geq c\left(\frac{1}{n}+\frac{1}{n_0}\right)^{2\alpha/(2\alpha+1)}, \tag{10}$$

where the infimum is taken over all possible estimators $\hat{r}_{n,n_0}$ obtained with the two data samples $(X_i)_{i\in\{1,\dots,n\}}$ and $(X_{0,i_0})_{i_0\in\{1,\dots,n_0\}}$.

The optimal convergence rate is thus $(n^{-1}+n_0^{-1})^{2\alpha/(2\alpha+1)}$. The upper-bound of Corollary 2 and the lower bound (10) match, up to constants. This proves that our estimation procedure achieves the minimax rate and is thus also optimal in the minimax sense. The result is not straightforward: the proof requires specific constructions, since it captures the influence of both sample sizes, n and n_0. Although it is a lower bound for a kind of density function, we think it can not be easily deduced from the minimax rate of density estimation over Besov ball (see for example Kerkyacharian and Picard 1992), since the two samples do not have symmetric roles.

4. Simulation

In this section, we present the performance of the adaptive estimator $\hat{r}_{\tilde{m}}$ on simulated data. We have carried out an intensive simulation study (with the computing environment MATLAB) which shows that the results are equivalent to the ones of Ćwik and Mielniczuk (1993) and Molanes-López and Cao (2008a). Here, we thus prefer to discuss two types of questions, to evaluate the specific robustness of our method. After describing the way we compute the estimator, we first focus on the quality of estimation when the variable X is close (in distribution) to X_0. Second, we investigate the role of the two sample sizes, n and n_0. For additional reconstructions, risk computations and details about calibration, the reader may refer to the supplementary material (Chagny and Lacour, 2014, Section 2).

4.1. **Implementation.** The implementation of the estimator is very simple, and follows the steps below.

- For each $m\in\mathcal{M}$, compute $(\hat{r}_m(x_k))_{k=1,\dots,K}$ defined by (3) for grid points $(x_k)_{k=1,\dots,K}$ evenly distributed across $(0;1)$, with $K=50$.
- For each $m\in\mathcal{M}$, compute $\widetilde{V}(m)$ and $\widetilde{A}(m)$, defined by (9).

- For $\widetilde{V}(h)$. We choose $c_0 = 1$, but the estimation results seem quite robust to slight changes. This value has been obtained by a numerical calibration on various examples (see Section 2.2 of the supplementary material Chagny and Lacour 2014 for more details). The index m^* of the estimator $\hat{r}_{m^*}$ used in $\widetilde{V}$ is the smallest integer greater than $\ln(n_0) - 1$.
- For $\widetilde{A}(h)$. We approximate the L^2 norms by the corresponding Riemann sums computed over the grid points $(x_k)_k$:

$$\|\hat{r}_{m'} - \hat{r}_{m\wedge m'}\|^2 \approx \frac{1}{K}\sum_{k=1}^{K}\left(\hat{r}_{m'}(x_k) - \hat{r}_{m\wedge m'}(x_k)\right)^2.$$

- Select the argmin $\tilde{m}$ of $\widetilde{A}(m) + \widetilde{V}(m)$, and choose $\hat{r}_{\tilde{m}}$.

The risks $\mathbb{E}[\|(\hat{r}_{\tilde{m}})_+ - r\|^2]$ are also computed: it is not difficult to see that the choice of the positive part of the estimator can only make its risk decreases. To compute the expectation, we average the integrated squared error (ISE) computed with $N = 500$ replications of the samples $(X_{0,i_0})_{i_0}$ and $(X_i)_i$. Notice that the grid size ($K = 50$), and the number of replications ($N = 500$) are the same as Ćwik and Mielniczuk (1993).

4.2. **Experiment 1 - Two samples with close distributions.** The trigonometric basis suits well to recover relative densities. Indeed, the first function of the basis is $\varphi_1 : x \in (0;1) \mapsto 1$, and thus the first estimated coefficient $\hat{a}_1$ in (3) also equals 1. But we know that the relative density is constant equal to 1 over $(0;1)$ when X and X_0 have the same distribution. Consequently, our procedure permits to obtain an exact estimation in this case, provided that the data driven criterion leads to the choice of the first model in the collection. We hope to select $D_{\hat{m}} = 1$, that is $\hat{m} = 0$. In this section, we check that the estimation procedure actually easily handles this case.

First, we generate two samples $(X_{0,i_0})_{i_0=1,\ldots,n_0}$ and $(X_i)_{i=1,\ldots,n}$ coming from random variables X_0 and X respectively, with one of the following common probability distributions (Example (1) in the sequel): (a1) a uniform distribution in the set $(0;1)$, (b1) a beta distribution $\mathcal{B}(2,5)$, (c1) a Gaussian distribution with mean 0 and variance 1, (d1) an exponential distribution with mean 5. As explained, the estimator is expected to be constant equal to 1: the selected index m must thus be 0. This is the case for most of the samples we simulate: for example, only 1% of the 500 estimators computed with 50 *i.i.d.* Gaussian pairs (X, X_0) are not identically equal to 1. The medians of the ISE over 500 replicated samples are always equal to 0, whatever the distribution of X and X_0, chosen among the examples (uniform, beta, Gaussian, or exponential). The MISE are dispayed in Table 1, for different possible sample sizes. We can also check that they are much more smaller than the MISE obtained with two different distributions for X and X_0 (see Table 2 in the supplementary material, Chagny and Lacour 2014, Section 2.2).

Then, we investigate what happens when X is close to X_0 but slightly different, with samples simulated from the set of Example (2).

(a2) The variable X_0 is from the uniform distribution on $(0;1)$, and the variable X has the density $f(x) = c\mathbf{1}_{(0;0.5)}(x) + (2-c)\mathbf{1}_{(0.5;1)}(x)$, with $c \in \{1.01, 1.05, 1.1, 1.3, 1.5\}$ (the case $c = 1$ is the case of the uniform distribution on $(0;1)$).

(b2) The variable X_0 is from the beta distribution $\mathcal{B}(2,5)$, and the variable X from a beta distribution $\mathcal{B}(a,5)$ with $a \in \{2.01, 2.05, 2.1, 2.3, 2.5\}$. For this example, the risks are computed over a regular grid of the interval $[F_0(0.01); F_0(0.99)]$.

Figure 1 shows the true relative densities for these two examples.

$n\backslash n_0$	50	100	200	400
50	0.213	0.206	0.156	0.185
100	0.114	0.159	0.115	0.096
200	0.125	0.109	0.058	0.056
400	0.089	0.078	0.054	0.036

Example (a1)

$n\backslash n_0$	50	100	200	400
50	0.180	0.163	0.165	0.157
100	0.140	0.153	0.105	0.105
200	0.110	0.095	0.075	0.069
400	0.099	0.076	0.047	0.035

Example (b1)

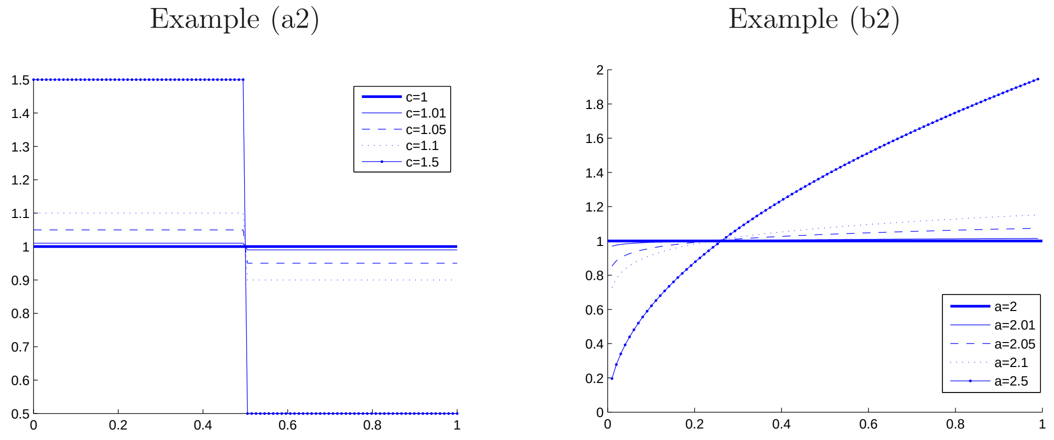
$n\backslash n_0$	50	100	200	400
50	0.245	0.162	0.202	0.119
100	0.125	0.131	0.110	0.099
200	0.141	0.103	0.077	0.055
400	0.132	0.086	0.051	0.039

Example (c1)

$n\backslash n_0$	50	100	200	400
50	0.177	0.186	0.147	0.165
100	0.117	0.119	0.092	0.094
200	0.095	0.099	0.081	0.073
400	0.084	0.105	0.056	0.041

Example (d1)

TABLE 1. Values of MISE ×10 averaged over 500 samples for the estimator $\hat{r}_{\tilde{m}}$, in Example (1) ((a1) to (d1)).


Example (a2)

Example (b2)

FIGURE 1. Plot of the different investigated relative densities of Examples (2), (a2) and (b2).

The MISEs in Examples (2) (a2) and (b2) are plotted in Figure 2 with respect to the sample sizes $n = n_0$. Details are also given in Table 1 of the supplementary material (Chagny and Lacour, 2014, Section 2.2). The larger c (resp. a), the further X from X_0 the larger the MISE. The results are thus better especially when the two variable distributions are close.

4.3. **Experiment 2 - Influence of the two sample sizes.** We now study the influence of the two sample sizes. Recall that the theoretical results we obtain do not require any link between n and n_0. On the contrary, they are often supposed to be proportional in the literature. But we obtain a convergence rate in which n and n_0 play symmetric roles (see Corollary 2). What happens in practice? To briefly discuss this question, let us consider the observations

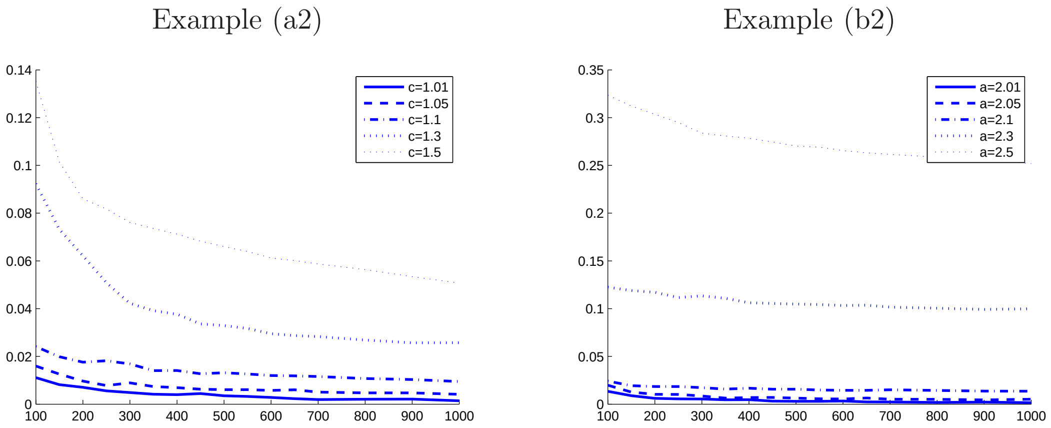


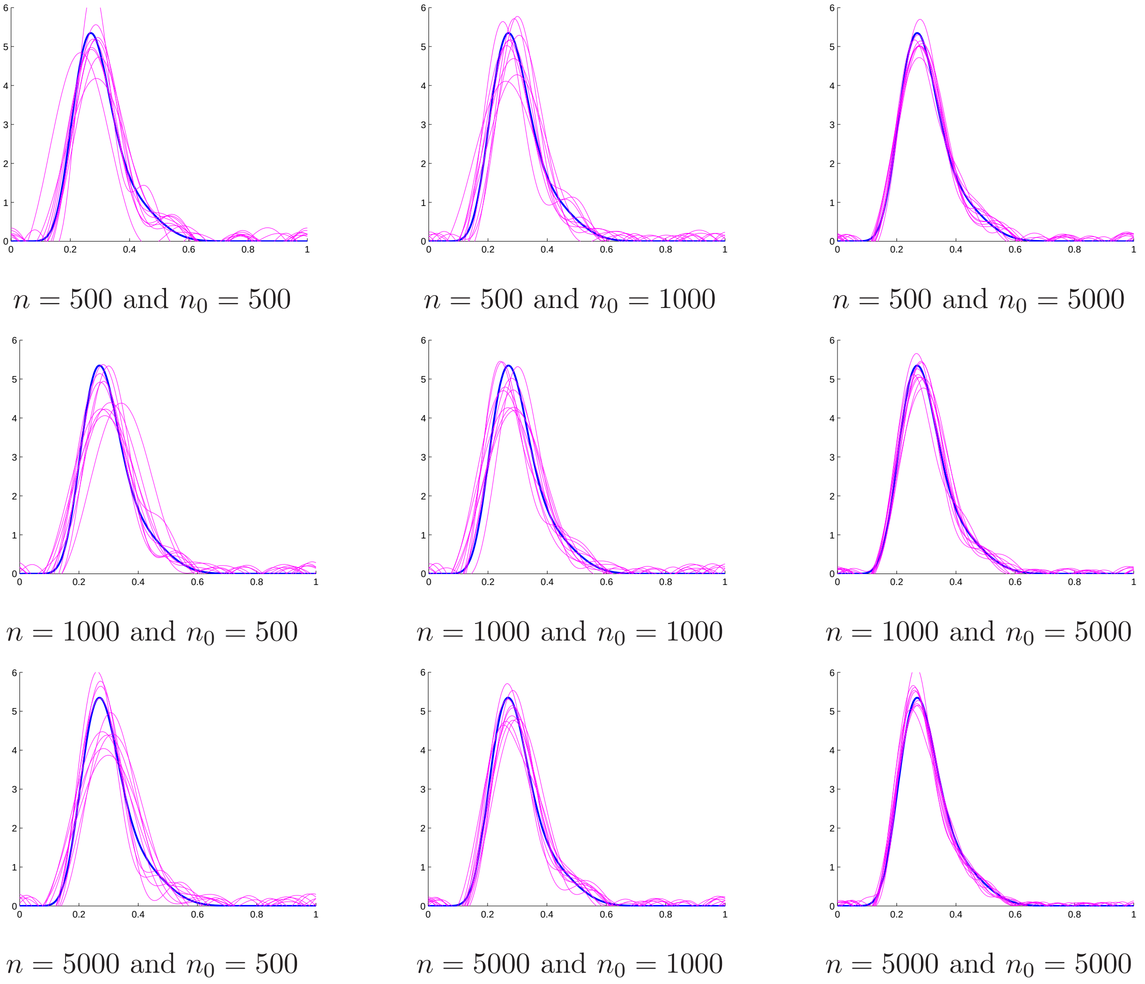
Figure 2. Values of MISE (averaged over 500 samples) for the estimator $\hat{r}_{\tilde{m}}$ with respect to the sample sizes $n = n_0$ in Examples (2) (a2) and (b2).

of $(X_i)_{i\in\{1,\ldots,n\}}$ and $(X_{0,i_0})_{i_0\in\{1,\ldots,n_0\}}$ fitting the following model (Example (3)). The variable X_0 is from the Weibull distribution with parameters (2,3) (we denote by W the corresponding c.d.f.) and X is built such that $X = W^{-1}(S)$, with S a mixture of two beta distributions: $\mathcal{B}(14, 37)$ with probability 4/5 and $\mathcal{B}(14, 20)$ with probability 1/5. The example is borrowed from Molanes-López and Cao (2008a). Let us look at the beams of estimates $\hat{r}_{\tilde{m}}$: in Figure 3, 10 estimators built from *i.i.d.* samples of data are plotted together with the true functions. This illustrates that increasing n_0 for fixed n seems to improve more substantially the risk than the other way round (the improvement when n_0 increases appears horizontally in Figure 3). Such a phenomenon also appears when a more quantitative criterion is considered: the MISE in Table 2 are not symmetric with respect to n and n_0, even if, as expected, they all get smaller when the sample sizes n and n_0 increase. Even if this can be suprising when compared with the theory, recall that the relative density of X with respect to X_0 is not the same as the relative density of X_0 with respect to X. The role of the reference variable is coherently more important, even if it is not clear in the convergence rate of Corollary 2. The details of the computation in the proofs also show that n and n_0 do not play similar roles (see *e.g.* Lemma 9). An explanation may be the following: in the method, the sample $(X_i)_{i\in\{1,\ldots,n\}}$ is used in a nonparametric way, like in classical density estimation, while the other, that is $(X_{0,i_0})_{i_0\in\{1,\ldots,n_0\}}$ is usefull through the empirical c.d.f. which is known to be convergent at a parametric rate, faster than the previous one. Notice finally that the same results are obtained for estimators computed from the sets of observations described in the supplementary material (see Table 2 in Chagny and Lacour 2014). In any case, such results might be used by a practitioner, when the choice of the reference sample is not natural: a judicious way to decide which of the sample which play the role of (X_{0,i_0}) is to choose the larger one.

5. Proofs

Detailed proofs of Proposition 1 and Theorem 2 are gathered in this section. The proofs of Theorems 3 and 4 are only sketched. Complete proofs are available in Section 3 of the supplementary material (Chagny and Lacour, 2014).

5.1. **Preliminary notations and results.**



$n = 500$ and $n_0 = 500$ $n = 500$ and $n_0 = 1000$ $n = 500$ and $n_0 = 5000$

$n = 1000$ and $n_0 = 500$ $n = 1000$ and $n_0 = 1000$ $n = 1000$ and $n_0 = 5000$

$n = 5000$ and $n_0 = 500$ $n = 5000$ and $n_0 = 1000$ $n = 5000$ and $n_0 = 5000$

FIGURE 3. Beams of 10 estimators built from *i.i.d.* samples of various sizes $(n; n_0)$ (thin lines) versus true function (thick line) in Example (3).

$n\backslash n_0$	50	100	200	400
50	12.05	7.977	5.631	3.745
100	11.68	7.596	4.789	3.297
200	12.57	7.557	4.831	2.731
400	11.26	7.445	4.429	2.729

TABLE 2. Values of MISE ×10 averaged over 500 samples for the estimator $\hat{r}_{\tilde{m}}$, in Example (3).

5.1.1. *Notations.* We need additional notations in this section. First, we specify the definition of the procedure. The estimators $\hat{r}_m$, $m \in \mathcal{M}$ defined by (3) are now denoted by $\hat{r}_m(., \hat{F}_0)$. Its coefficients in the Fourier basis are $\hat{a}_j^{\hat{F}_0}$. When we plug F_0 in (3), we denote it by $\hat{r}_m(., F_0)$, and the coefficients by $\hat{a}_j^{F_0}$. Then, we set $U_{0,i_0} = F_0(X_{0,i_0})$ ($i_0 = 1, \dots, n_0$), and let $\widehat{U}_0$ be the empirical c.d.f. associated to the sample $(U_{0,i_0})_{i_0=1,\dots,n_0}$. We also denote by $\mathbb{E}[.|(X_0)]$ the conditional expectation given the sample $(X_{0,i_0})_{i_0=1,\dots,n_0}$ (the conditional variance will be coherently denoted by $\mathrm{Var}(.|(X_0)))$.

Finally, for any measurable function t defined on $(0;1)$, we denote by $\|t\|_\infty$ the quantity $\sup_{x\in(0;1)}|t(x)|$, and id is the function such that $u\mapsto u$, on the interval $(0;1)$.

5.1.2. *Useful tools.* Key arguments for the proofs are the deviations properties of the empirical c.d.f. $\hat{F}_0$ of the sample $(X_{0,i_0})_{i_0}$.

First, recall that U_{0,i_0} is a uniform variable on $(0;1)$ and that $\hat{F}_0(F_0^{-1}(u)) = \widehat{U}_0(u)$, for all $u\in(0;1)$. Keep in mind that the random variable $\sup_{x\in A_0}|\hat{F}_0(x)-F_0(x)|$ has the same distribution as $\|\widehat{U}_0-id\|_\infty$. The following inequalities are used several times to control the deviations of the empirical c.d.f. $\hat{U}_n$. Dvoretzky et al. (1956) established the first one.

Proposition 5. *(Dvoretzky-Kiefer-Wolfowitz's Inequality) There exists a constant $C>0$, such that, for any integer $n_0\geq 1$ and any $\lambda>0$,*

$$\mathbb{P}\left(\left\|\widehat{U}_0-id\right\|_\infty\geq\lambda\right)\leq C\exp\left(-2n_0\lambda^2\right).$$

By integration, we then deduce a first other bound:

Proposition 6. *For any integer $p>0$, there exists a constant $C_p>0$ such that*

$$\mathbb{E}\left[\left\|\widehat{U}_0-id\right\|_\infty^p\right]\leq\frac{C_p}{n_0^{p/2}}.$$

More precise bounds are also required:

Corollary 3. *For any $\kappa>0$, for any integer $p\geq 2$, there exists also a constant C such that*

$$\mathbb{E}\left[\left(\left\|\widehat{U}_0-id\right\|_\infty^p-\kappa\frac{\ln^{p/2}(n_0)}{n_0^{p/2}}\right)_+\right]\leq Cn_0^{-2^{\frac{2-p}{p}}\kappa^{2/p}}. \tag{11}$$

5.1.3. *The Talagrand Inequality.* The proofs of the main results (Theorems 2 and 3) are based on the use of concentration inequalities. The first one is the classical Bernstein Inequality, and the second one is the following version of the Talagrand Inequality.

Proposition 7. *Let $\xi_1,\dots,\xi_n$ be i.i.d. random variables, and define $\nu_n(s)=\frac{1}{n}\sum_{i=1}^n s(\xi_i)-\mathbb{E}[s(\xi_i)]$, for s belonging to a countable class $\mathcal{S}$ of real-valued measurable functions. Then, for $\delta>0$, there exist three constants c_l, $l=1,2,3$, such that*

$$\begin{aligned}\mathbb{E}\left[\left(\sup_{s\in\mathcal{S}}(\nu_n(s))^2-c(\delta)H^2\right)_+\right] \quad \leq \quad & c_1\left\{\frac{v}{n}\exp\left(-c_2\delta\frac{nH^2}{v}\right)\right. \\ & \left.+\frac{M_1^2}{C^2(\delta)n^2}\exp\left(-c_3C(\delta)\sqrt{\delta}\frac{nH}{M_1}\right)\right\},\end{aligned} \tag{12}$$

with $C(\delta)=(\sqrt{1+\delta}-1)\wedge 1$, $c(\delta)=2(1+2\delta)$ and

$$\sup_{s\in\mathcal{S}}\|s\|_\infty\leq M_1,\ \mathbb{E}\left[\sup_{s\in\mathcal{S}}|\nu_n(s)|\right]\leq H,\ \textit{and}\ \sup_{s\in\mathcal{S}}\mathit{Var}(s(\xi_1))\leq v.$$

Inequality (12) is a classical consequence of Talagrand's Inequality given in Klein and Rio (2005): see for example Lemma 5 (page 812) in Lacour (2008). Using density arguments, we can apply it to the unit sphere of a finite dimensional linear space.

5.2. **Proof of Proposition 1.** A key point is the following decomposition which holds for any index m

$$\left\|\hat{r}_m(.,\hat{F}_0)-r\right\|^2 \leq 3T_1^m+3T_2^m+3\left\|\hat{r}_m(.,F_0)-r\right\|^2,$$

with

$$\begin{aligned} T_1^m &= \left\|\hat{r}_m(.,\hat{F}_0)-\hat{r}_m(.,F_0)-\mathbb{E}\left[\hat{r}_m(.,\hat{F}_0)-\hat{r}_m(.,F_0)\,|(X_0)\right]\right\|^2, \\ T_2^m &= \left\|\mathbb{E}\left[\hat{r}_m(.,\hat{F}_0)-\hat{r}_m(.,F_0)\,|(X_0)\right]\right\|^2. \end{aligned} \tag{13}$$

We have already proved (see (4) and (5)) that $\|\hat{r}_m(.,F_0)-r\|^2 \leq D_m/n+\|r_m-r\|^2$. Therefore, it remains to apply the two following lemmas, proved in the two following sections.

Lemma 8. *Under the assumptions of Proposition 1,*

$$\mathbb{E}\left[T_1^m\right] \leq 2\pi^2\frac{D_m^3}{nn_0}$$

Lemma 9. *Under the assumptions of Proposition 1, ,*

$$\mathbb{E}\left[T_2^m\right] \leq 3\|r\|^2\frac{D_m}{n_0}+3\frac{\pi^4}{4}C_4\|r\|^2\frac{D_m^4}{n_0^2}+\frac{32\pi^6C_6}{3}\|r\|^2\frac{D_m^7}{n_0^3}+3\frac{\|r'\|^2}{n_0}.$$

The result follows if $D_m \leq \kappa n_0^{1/3}$.

□

5.2.1. *Proof of Lemma 8.* The decompositions of the estimator in the orthogonal basis $(\varphi_j)_j$ yields

$$T_1^m = \sum_{j=1}^{D_m}\left(\hat{a}_j^{\hat{F}_0}-\hat{a}_j^{F_0}-\mathbb{E}\left[\hat{a}_j^{\hat{F}_0}-\hat{a}_j^{F_0}\,|(X_0)\right]\right)^2,$$

and therefore, $\mathbb{E}[T_1^m|(X_0)]=\sum_{j=1}^{D_m}\mathrm{Var}(\hat{a}_j^{\hat{F}_0}-\hat{a}_j^{F_0}|(X_0))$. Moreover, for any index j,

$$\begin{aligned} \mathrm{Var}\left(\hat{a}_j^{\hat{F}_0}-\hat{a}_j^{F_0}\,|(X_0)\right) &\leq \frac{1}{n}\mathbb{E}\left[\left(\varphi_j\circ\hat{F}_0(X_1)-\varphi_j\circ F_0(X_1)\right)^2|(X_0)\right], \\ &\leq \frac{1}{n}\left\|\varphi_j'\right\|_\infty^2\int_A\left(\hat{F}_0(x)-F_0(x)\right)^2f(x)dx, \end{aligned}$$

by using the mean-value theorem. Since $\|\varphi_j'\|_\infty^2 \leq 8\pi^2D_m^2$ in the Fourier basis, this leads to

$$\mathbb{E}\left[T_1^m\right] \leq \frac{8\pi^2}{n}D_m^3\int_A\mathbb{E}\left[\left(\hat{F}_0(x)-F_0(x)\right)^2\right]f(x)dx.$$

Notice finally that $\mathbb{E}[(\hat{F}_0(x)-F_0(x))^2]=\mathrm{Var}(\hat{F}_0(x))=(F_0(x)(1-F_0(x)))/n_0 \leq 1/(4n_0)$. This permits to conclude the proof of Lemma 8.

□

5.2.2. *Proof of Lemma 9.* Arguing as in the beginning of the proof of Lemma 8 yields

$$T_2^m = \sum_{j=1}^{D_m} \left(\int_A \left(\varphi_j \circ \hat{F}_0(x) - \varphi_j \circ F_0(x) \right) f(x)dx \right)^2. \tag{14}$$

We apply the Taylor formula to the function φ_j, with the Lagrange form for the remainder. There exists a random number $\hat{\alpha}_{j,n_0,x}$ such that the following decomposition holds: $T_2^m \leq 3T_{2,1}^m + 3T_{2,2}^m + 3T_{2,3}^m$, where

$$T_{2,1}^m = \sum_{j=1}^{D_m} \left(\int_A \varphi_j'(F_0(x)) \left(\hat{F}_0(x) - F_0(x) \right) f(x)dx \right)^2,$$

$$T_{2,2}^m = \sum_{j=1}^{D_m} \left(\int_A \varphi_j''(F_0(x)) \frac{\left(\hat{F}_0(x) - F_0(x) \right)^2}{2} f(x)dx \right)^2,$$

$$T_{2,3}^m = \sum_{j=1}^{D_m} \left(\int_A \varphi_j^{(3)}(\hat{\alpha}_{j,n_0,x}) \frac{\left(\hat{F}_0(x) - F_0(x) \right)^3}{6} f(x)dx \right)^2.$$

We now bound each of these three terms. Let us begin with $T_{2,1}^m$. The change of variables $u = F_0(x)$ permits to obtain first

$$T_{2,1}^m = \sum_{j=1}^{D_m} \left(\int_{(0;1)} \varphi_j'(u) \left(\widehat{U}_0(u) - u \right) r(u)du \right)^2,$$

and, with the definition of $\widehat{U}_0(u)$, we get

$$T_{2,1}^m = \sum_{j=1}^{D_m} \left(\frac{1}{n_0} \sum_{i=1}^{n_0} B_{i,j} - \mathbb{E}[B_{i,j}] \right)^2, \quad \text{with } B_{i,j} = \int_{U_{0,i}}^1 r(u)\varphi_j'(u)du.$$

An integration by parts for $B_{i,j}$ leads to another splitting $T_{2,1}^m \leq 2T_{2,1,1}^m + 2T_{2,1,2}^m$, with notations

$$T_{2,1,1}^m = \sum_{j=1}^{D_m} \left\{ \frac{1}{n_0} \sum_{i=1}^{n_0} r(U_{0,i})\varphi_j(U_{0,i}) - \mathbb{E}\left[r(U_{0,i})\varphi_j(U_{0,i}) \right] \right\}^2,$$

$$T_{2,1,2}^m = \sum_{j=1}^{D_m} \left\{ \int_{(0;1)} r'(u) \left(\widehat{U}_0(u) - u \right) \varphi_j(u)du \right\}^2.$$

The expectation of the first term is a variance and is bounded as follows:

$$\mathbb{E}\left[T_{2,1,1}^m \right] \leq \frac{1}{n_0} \sum_{j=1}^{D_m} \mathbb{E}\left[\left(r(U_{0,1})\varphi_j(U_{0,1}) \right)^2 \right] \leq \int_0^1 r(u)^2 du \frac{D_m}{n_0}.$$

For $T_{2,1,2}^m$, we use the definitions and properties of the orthogonal projection operator Π_{S_m} on the space S_m:

$$\begin{aligned} T_{2,1,2}^m &= \sum_{j=1}^{D_m} \left(\langle r'(\widehat{U}_0 - id), \varphi_j \rangle_{(0;1)} \right)^2 = \left\| \Pi_{S_m}(r'(\widehat{U}_0 - id)) \right\|^2, \\ &\leq \left\| r'(\widehat{U}_0 - id) \right\|^2 \leq \|r'\|^2 \|\widehat{U}_0 - id\|_\infty^2. \end{aligned}$$

Applying Proposition 6 proves that $\mathbb{E}[T_{2,1,2}^m] \leq C_2 \|r'\|^2/n_0$. Therefore,

$$\mathbb{E}\left[T_{2,1}^m\right] \leq \|r\|^2 \frac{D_m}{n_0} + C_2 \|r'\|^2 \frac{1}{n_0}. \tag{15}$$

Consider now $T_{2,2}^m$. The trigonometric basis satisfies $\varphi_j'' = -(\pi\mu_j)^2\varphi_j$, with $\mu_j = j$ for even $j \geq 2$, and $\mu_j = j-1$ for odd $j \geq 2$. We thus have,

$$\begin{aligned} \mathbb{E}\left[T_{2,2}^m\right] &= (\pi^4/4)\mathbb{E}\left[\sum_{j=1}^{D_m}\left\{\int_{(0;1)} r(u)\left(\widehat{U}_0(u)-u\right)^2 \mu_j^2\varphi_j(u)du\right\}^2\right], \\ &\leq (\pi^4/4)D_m^4\mathbb{E}\left[\sum_{j=1}^{D_m}\left\{\langle r\left(\widehat{U}_0 - id\right)^2, \varphi_j\rangle_{(0;1)}\right\}^2\right], \\ &\leq (\pi^4/4)D_m^4\mathbb{E}\left[\left\|r\left(\widehat{U}_0-id\right)^2\right\|^2\right] \leq (\pi^4/4)D_m^4\mathbb{E}\left[\left\|\widehat{U}_0-id\right\|_\infty^4\right]\int_{(0;1)} r^2(u)du. \end{aligned}$$

Thanks to Proposition 6, we obtain

$$\mathbb{E}\left[T_{2,1}^m\right] \leq C_4(\pi^4/4)\|r\|^2\frac{D_m^4}{n_0^2}. \tag{16}$$

The last term is then easily controlled, using also Proposition 6:

$$\mathbb{E}\left[T_{2,3}^m\right] \leq \frac{32\pi^6}{9}\sum_{j=1}^{D_m}\|r\|^2\mathbb{E}\left[\left\|\widehat{U}_0-id\right\|_\infty^6\right] \leq \frac{32\pi^6 C_6}{9}\|r\|^2\frac{D_m^7}{n_0^3}. \tag{17}$$

Lemma 9 is proved by gathering (15), (16) and (17).

□

5.3. **Proof of Theorem 2.** In the proof, C is a constant which may change from line to line, and is independent of all $m \in \mathcal{M}$, n, and n_0. Let $m \in \mathcal{M}$ be fixed. The following decomposition holds:

$$\begin{aligned} \left\|\hat{r}_{\hat{m}}\left(.,\hat{F}_0\right) - r\right\|^2 &\leq 3\left\|\hat{r}_{\hat{m}}\left(.,\hat{F}_0\right) - \hat{r}_{m\wedge\hat{m}}\left(.,\hat{F}_0\right)\right\|^2 \\ &\quad +3\left\|\hat{r}_{m\wedge\hat{m}}\left(.,\hat{F}_0\right) - \hat{r}_m\left(.,\hat{F}_0\right)\right\|^2 + 3\left\|\hat{r}_m\left(.,\hat{F}_0\right) - r\right\|^2. \end{aligned}$$

We use successively the definition of $A(\hat{m})$, $A(m)$, and $\hat{m}$ to obtain

$$\left\|\hat{r}_{\hat{m}}\left(.,\hat{F}_0\right) - r\right\|^2 \leq 6\left(A(m)+V(m)\right) + 3\left\|\hat{r}_m\left(.,\hat{F}_0\right) - r\right\|^2.$$

Keeping in mind that we can split $\|\hat{r}_m(.,\hat{F}_0) - r\|^2 \leq 3T_1^m + 3T_2^m + 3\|\hat{r}_m(.,F_0) - r\|^2$ with the notations of Section 5.2, we derive from (4) and (5):

$$\left\|\hat{r}_{\hat{m}}\left(.,\hat{F}_0\right) - r\right\|^2 \leq 6\left(A(m)+V(m)\right) + 9T_1^m + 9T_2^m + 9\frac{D_m}{n} + 9\left\|r_m - r\right\|^2.$$

We also apply Lemmas 8 and 9. Taking into account that $D_m \leq \kappa n_0^{1/3}$, we thus have

$$\begin{aligned} \mathbb{E}\left[\left\|\hat{r}_{\hat{m}}\left(.,\hat{F}_0\right) - r\right\|^2\right] &\leq 6\mathbb{E}\left[A(m)\right] + 6V(m) + C\frac{D_m}{n} + C\|r\|^2\frac{D_m}{n_0} \\ &\quad +9\left\|r_m - r\right\|^2 + \frac{C}{n_0} + \frac{C}{n}. \end{aligned}$$

Therefore, the conclusion of Theorem 2 is the result of the following lemma.

Lemma 10. *Under the assumptions of Theorem 2, there exists a constant $C>0$ such that, for any $m\in\mathcal{M}$,*

$$\mathbb{E}\left[A(m)\right]\leq C\left(\frac{1}{n}+\frac{1}{n_0}\right)+12\left\|r_m-r\right\|^2.$$

□

5.3.1. *Proof of Lemma 10.* To study $A(m,\hat{F}_0)$, we write, for $m'\in\mathcal{M}$.

$$\begin{aligned}\left\|\hat{r}_{m'}\left(.,\hat{F}_0\right)-\hat{r}_{m\wedge m'}\left(.,\hat{F}_0\right)\right\|^2 &\leq 3\left\|\hat{r}_{m'}\left(.,\hat{F}_0\right)-r_{m'}\right\|^2+3\left\|r_{m'}-r_{m\wedge m'}\right\|^2\\ &\quad +3\left\|r_{m\wedge m'}-\hat{r}_{m\wedge m'}\left(.,\hat{F}_0\right)\right\|^2.\end{aligned}$$

Let $\mathcal{S}(p_{m'})$ be the set $\{t\in S_{p_{m'}},\ \|t\|=1\}$, for $p_{m'}=m'$ or $p_{m'}=m\wedge m'$. We note that

$$\left\|r_{p_{m'}}-\hat{r}_{p_{m'}}(.,\hat{F}_0)\right\|^2=\sum_{j=1}^{D_{p_{m'}}}\left(\tilde{\nu}_n(\varphi_j)\right)^2=\sup_{t\in\mathcal{S}(p_{m'})}\tilde{\nu}_n(t)^2, \tag{18}$$

with $\tilde{\nu}_n(t)=n^{-1}\sum_{i=1}^n t\circ\hat{F}_0(X_i)-\mathbb{E}[t\circ F_0(X_i)]$. Since the empirical process $\tilde{\nu}_n$ is not centered, we consider the following splitting: $(\tilde{\nu}_n(t))^2\leq 2\nu_n^2(t)+2((1/n)\sum_{i=1}^n(t\circ\hat{F}_0(X_i)-t\circ F_0(X_i)))^2$, with

$$\nu_n(t)=\frac{1}{n}\sum_{i=1}^n\left(t\circ F_0(X_i)-\mathbb{E}\left[t\circ F_0(X_i)\right]\right). \tag{19}$$

But we also have

$$\sup_{t\in\mathcal{S}(p_{m'})}\left(\frac{1}{n}\sum_{i=1}^n\left(t\circ\hat{F}_0(X_i)-t\circ F_0(X_i)\right)\right)^2 = \sum_{j=1}^{D_{p_{m'}}}\left(\hat{a}_j^{\hat{F}_0}-\hat{a}_j^{F_0}\right)^2\leq 2T_1^{p_{m'}}+2T_2^{p_{m'}},$$

with the notations of Section 5.2. This shows that

$$\left\|r_{p_{m'}}-\hat{r}_{p_{m'}}(.,\hat{F}_0)\right\|^2\leq 2\sup_{t\in\mathcal{S}(p_{m'})}(\nu_n(t))^2+4T_1^{p_{m'}}+4T_2^{p_{m'}}. \tag{20}$$

We thus have

$$\begin{aligned}\left\|\hat{r}_{m'}\left(.,\hat{F}_0\right)-\hat{r}_{m\wedge m'}\left(.,\hat{F}_0\right)\right\|^2 &\leq 6\sup_{t\in\mathcal{S}(m')}(\nu_n(t))^2+6\sup_{t\in\mathcal{S}(m\wedge m')}(\nu_n(t))^2+12T_2^{m'}+12T_2^{m\wedge m'}\\ &\quad +12T_1^{m'}+12T_1^{m\wedge m'}+3\left\|r_{m'}-r_{m\wedge m'}\right\|^2.\end{aligned}$$

We get back to the definition of $A(m)$. To do so, we subtract $V(m')$. For convenience, we split it into two terms: $V(m')=V^{(1)}(m')+V^{(2)}(m')$, with $V^{(1)}(m')=c_0D_m/n$, and $V^{(2)}(m')=c_0\|r\|^2D_m/n_0$. Thus,

$$\begin{aligned}\mathbb{E}\left[A\left(m\right)\right] &\leq 6\mathbb{E}\left[\max_{m'\in\mathcal{M}}\left(\sup_{t\in\mathcal{S}(m')}(\nu_n(t))^2-\frac{V^{(1)}(m')}{12}\right)_+\right]+3\max_{m'\in\mathcal{M}}\left\|r_{m'}-r_{m\wedge m'}\right\|^2\\ &+6\mathbb{E}\left[\max_{m'\in\mathcal{M}}\left(\sup_{t\in\mathcal{S}(m\wedge m')}(\nu_n(t))^2-\frac{V^{(1)}(m')}{12}\right)_+\right]+12\mathbb{E}\left[\max_{m'\in\mathcal{M}}\left(T_2^{m'}-\frac{V^{(2)}(m')}{24}\right)_+\right]\\ &+12\mathbb{E}\left[\max_{m'\in\mathcal{M}}\left(T_2^{m\wedge m'}-\frac{V^{(2)}(m')}{24}\right)_+\right]+12\mathbb{E}\left[\max_{m'\in\mathcal{M}}T_1^{m'}\right]+12\mathbb{E}\left[\max_{m'\in\mathcal{M}}T_1^{m\wedge m'}\right].\end{aligned}$$

For the deterministic term, we notice that

$$\max_{m'\in\mathcal{M}} \|r_{m'} - r_{m\wedge m'}\|^2 \quad \leq \quad 2 \max_{\substack{m'\in\mathcal{M}\\ m\leq m'}} \|r_{m'} - r\|^2 + 2\,\|r - r_m\|^2 .$$

If $m \leq m'$, the spaces are nested $S_m \subset S_{m'}$, thus the orthogonal projections r_m and $r_{m'}$ of r onto S_m and S_m' respectively satisfy $\|r_{m'} - r\|^2 \leq \|r_m - r\|^2$. Thus,

$$\max_{m'\in\mathcal{M}} \|r_{m'} - r_{m\wedge m'}\|^2 \leq 4\,\|r_m - r\|^2 . \tag{21}$$

Moreover, for $p_{m'} = m'$ or $p_{m'} = m\wedge m'$, $T_1^{p_{m'}} \leq T_1^{m_{\max}}$ (recall that $m_{\max}$ is the largest index in the collection $\mathcal{M}$). Therefore,

$$12\mathbb{E}\left[\max_{m'\in\mathcal{M}} T_1^{m'}\right] + 12\mathbb{E}\left[\max_{m'\in\mathcal{M}} T_1^{m\wedge m'}\right] \leq 24\mathbb{E}\left[T_1^{m_{\max}}\right] \leq C\frac{D_{m_{\max}}^3}{nn_0} \leq \frac{C}{n}.$$

Consequently, we have at this stage

$$\begin{aligned}\mathbb{E}\left[A\left(m\right)\right] \quad \leq \quad & \frac{C}{n} + 12\,\|r_m - r\|^2 + 6\mathbb{E}\left[\max_{m'\in\mathcal{M}}\left(\sup_{t\in\mathcal{S}(m')} (\nu_n(t))^2 - \frac{V^{(1)}(m')}{12}\right)_+\right] \\ & +6\mathbb{E}\left[\max_{m'\in\mathcal{M}}\left(\sup_{t\in\mathcal{S}(m\wedge m')} (\nu_n(t))^2 - \frac{V^{(1)}(m')}{12}\right)_+\right] \\ & +12\mathbb{E}\left[\max_{m'\in\mathcal{M}}\left(T_2^{m'} - \frac{V^{(2)}(m')}{24}\right)_+\right] + 12\mathbb{E}\left[\max_{m'\in\mathcal{M}}\left(T_2^{m\wedge m'} - \frac{V^{(2)}(m')}{24}\right)_+\right].\end{aligned}$$

Since $V^{(l)}(m') \geq V^{(l)}(m'\wedge m)$ it remains to bound the two following terms:

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}}\left(\sup_{t\in\mathcal{S}(p_{m'})} (\nu_n(t))^2 - \frac{V^{(1)}(p_{m'})}{12}\right)_+\right] \text{ and } \mathbb{E}\left[\max_{m'\in\mathcal{M}}\left(T_2^{p_{m'}} - \frac{V^{(2)}(p_{m'})}{24}\right)_+\right]$$

We use the two following lemmas. The first one is proved below, the second one is proved in Section 3.1 of the supplementary material (Chagny and Lacour, 2014).

Lemma 11. *Assume that r is bounded on $(0;1)$. The deviations of the empirical process ν_n defined by (19) can be controlled as follows,*

$$\forall\delta > 0,\ \ \mathbb{E}\left[\max_{m'\in\mathcal{M}}\left\{\sup_{t\in\mathcal{S}(p_{m'})} \nu_n^2(t) - \bar{V}_\delta(p_{m'})\right\}_+\right] \leq \frac{C(\delta)}{n},$$

where $\bar{V}_\delta(p_{m'}) = 2(1+2\delta)D_{p_{m'}}/n$, and $C(\delta)$ a constant which depends on δ.

We fix a $\delta > 0$ (e.g. $\delta = 1/2$). We choose c_0 in the definition of V (see (7)) large enough to have $V^{(1)}(p_{m'})/12 \geq \bar{V}_\delta(p_{m'})$, for every m'. The inequality of Lemma 11 with $V^{(1)}(p_{m'})$ as a replacement for $\bar{V}_\delta(p_{m'})$.

Lemma 12. *Under the assumptions of Theorem 2,*

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}}\left(T_2^{p_{m'}} - V_2(p_{m'})\right)_+\right] \leq \frac{C}{n_0},$$

with $V_2(p_{m'}) = c_2\|r\|^2 D_{p'_m}/n_0$, c_2 a positive constant large enough, and C depending on the basis, on r, and on the constants C_p of Proposition 6.

We choose c_0 in the definition of V (see (7)) large enough to have $V^{(2)}(p_{m'})/24 \geq V_2(p_{m'})$, for every m'. This enables to apply Lemma 12 with $V^{(2)}(p_{m'})$ as a replacement for $V_2(p_{m'})$.

The proof of Lemma 10 is completed.

□

5.3.2. *Proof of Lemma 11.* We roughly bound

$$\mathbb{E}\left[\max_{m'\in\mathcal{M}}\left\{\sup_{t\in\mathcal{S}(p_{m'})}\nu_n^2(t)-\bar{V}_\delta(p_{m'})\right\}_+\right] \leq \sum_{m'\in\mathcal{M}}\mathbb{E}\left[\left\{\sup_{t\in\mathcal{S}(p_{m'})}\nu_n^2(t)-\bar{V}_\delta(p_{m'})\right\}_+\right].$$

We apply the Talagrand Inequality recalled in Proposition 7. To this aim, we compute M_1, H^2 and v. Write for a moment $\nu_n(t)=(1/n)\sum_{i=1}^n \psi_t(X_i)-\mathbb{E}[\psi_t(X_i)]$, with $\psi_t(x)=t\circ F_0(x)$.

- First, for $t\in\mathcal{S}(p_{m'})$, $\sup_{x\in A}|\psi_t(x)|\leq\|t\|_\infty\leq\sqrt{D_{p_{m'}}}\|t\|=\sqrt{D_{p_{m'}}}=: M_1$.
- Next, we develop $t\in\mathcal{S}(p_{m'})$ in the orthogonal basis $(\varphi_j)_{j=1,\ldots,D_{p_{m'}}}$. This leads to

$$\mathbb{E}\left[\sup_{t\in\mathcal{S}(p_{m'})}\nu_n^2(t)\right]\leq\sum_{j=1}^{D_{p_{m'}}}\mathbb{E}\left[\nu_n(\varphi_j^2)\right]=\sum_{j=1}^{D_{p_{m'}}}\mathbb{E}\left[\left(\hat{a}_j^{F_0}-a_j\right)^2\right]\leq\frac{D_{p_{m'}}}{n}=: H^2,$$

  thanks to the upper-bound for the variance term (see (5)).
- Last, for $t\in\mathcal{S}(p_{m'})$, $\mathrm{Var}(\psi_t(X_1))\leq\int_A t^2(F_0(x))f(x)dx=\int_{(0;1)}t^2(u)r(u)du\leq\|r\|_\infty\|t\|^2=\|r\|_\infty =: v$.

Inequality (12) gives, for $\delta>0$,

$$\begin{aligned}\sum_{m'\in\mathcal{M}}\mathbb{E}\left[\left(\sup_{t\in\mathcal{S}(p_{m'})}\nu_n^2(t)-c(\delta)H^2\right)_+\right] \quad &\leq \quad c_1\sum_{m'\in\mathcal{M}}\left\{\frac{1}{n}\exp\left(-c_2\delta D_{p_{m'}}\right)\right.\\ &\left.\qquad +\frac{D_{p_{m'}}}{C^2(\delta)n^2}\exp\left(-c_3C(\delta)\sqrt{\delta}\sqrt{n}\right)\right\},\end{aligned}$$

where c_l, $l=1,2,3$ are three constants. Now, it is sufficient to use that $D_{p'_m}=2p_{m'}+1$, and that the cardinal of $\mathcal{M}$ is bounded by n to end the proof of Lemma 11.

5.4. **Sketch of the proof of Theorem 3.** The main idea is to introduce the set

$$\Lambda=\left\{\left|\frac{\|\hat{r}_{m^*}(.,\hat{F}_0)\|}{\|r\|}-1\right|<\frac{1}{2}\right\},$$

and to split

$$\mathbb{E}\left[\|\hat{r}_{\tilde{m}}(.,\hat{F}_0)-r\|^2\right]=\mathbb{E}\left[\|\hat{r}_{\tilde{m}}(.,\hat{F}_0)-r\|^2\mathbf{1}_\Lambda\right]+\mathbb{E}\left[\|\hat{r}_{\tilde{m}}(.,\hat{F}_0)-r\|^2\mathbf{1}_{\Lambda^c}\right].$$

Then, the aim is to show that the first term give the order of the upper-bound of Theorem 3, and that the probability of the set Λ^c is negligible compared to $1/n+1/n_0$. See the supplementary material Chagny and Lacour (2014, Section 3.2).

5.5. **Sketch of the proof of Theorem 4.** Denote by $\phi_{n,n_0} = (\min(n, n_0))^{-2\alpha/(2\alpha+1)}$. Since there exists a constant $c' > 0$ (depending on α) such that $(n^{-1} + n_0^{-1})^{2\alpha/(2\alpha+1)} \leq c'\phi_{n,n_0}$, it is sufficient to prove Inequality (10) with the lower bound ϕ_{n,n_0}. We also separate two cases: $n \leq n_0$ and $n > n_0$. Then the result comes down to the proof of two inequalities. For each of these inequalities, the proof is based on the general reduction scheme which can be found in Section 2.6 of Tsybakov (2009): the main idea is to reduce the class of functions $\mathcal{F}_\alpha$ to a finite well-chosen subset $\{r_a, r_1, \dots, r_M\}$, $M \geq 2$. All the technical details are provided in the supplementary material Chagny and Lacour (2014, Section 3.3).

Acknowledgements

We are very grateful to Fabienne Comte for constructive discussions and careful reading of the paper. We thank an anonymous referee for pointing us to the link between our estimator and a kernel one.

References

A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

C. Bell and K. Doksum. "Optimal" one-sample distribution-free tests and their two-sample extensions. *Ann. Math. Statist.*, 37:120–132, 1966.

K. Bertin, L. Lacour, and V. Rivoirard. Adaptive estimation of conditional density function. arXiv:1312.7402v1, 2013.

L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

C. Butucea and K. Tribouley. Nonparametric homogeneity tests. *J. Statist. Plann. Inference*, 136(3):597–639, 2006.

G. Chagny. Penalization versus Goldenshluger-Lepski strategies in warped bases regression. *ESAIM Probab. Statist.*, 17:328–358 (electronic), 2013.

G. Chagny and C. Lacour. Optimal adaptive estimation of the relative density: Supplementary material. Technical report, 2014. URL `https://hal.archives-ouvertes.fr/hal-00955161`.

K. F. Cheng and C. K. Chu. Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, 10(4):583–604, 2004.

G. Claeskens, B.-Y. Jing, L. Peng, and W. Zhou. Empirical likelihood confidence regions for comparison distributions and ROC curves. *Canad. J. Statist.*, 31(2):173–190, 2003.

F. Comte and J. Johannes. Adaptive functional linear regression. *Ann. Statist.*, 40(6):2765–2797, 2012.

J. Ćwik and J. Mielniczuk. Data-dependent bandwidth choice for a grade density kernel estimate. *Statist. Probab. Lett.*, 16(5):397–405, 1993.

R. DeVore and G. Lorentz. *Constructive approximation*, volume 303 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1993.

J. Doob. Heuristic approach to the Kolmogorov-Smirnov theorems. *The Annals of Mathematical Statistics*, 20(3):393–403, 1949.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27:642–669, 1956.

S. Efromovich. *Nonparametric curve estimation.* Springer Series in Statistics. Springer-Verlag, New York, 1999. Methods, theory, and applications.

J. L. Gastwirth. The first-median test: A two-sided version of the control median test. *J. Amer. Statist. Assoc.*, 63:692–706, 1968.

J. D. Gibbons and S. Chakraborti. *Nonparametric statistical inference*. Statistics: Textbooks and Monographs. CRC Press, Boca Raton, FL, fifth edition, 2011. ISBN 978-1-4200-7761-2.

I. Gijbels and J. Mielniczuk. Asymptotic properties of kernel estimators of the Radon-Nikodým derivative with applications to discriminant analysis. *Statist. Sinica*, 5(1):261–278, 1995.

A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3):1608–1632, 2011.

P. Hall and R. Hyndman. Improved methods for bandwidth selection when estimating ROC curves. *Statist. Probab. Lett.*, 64(2):181–189, 2003.

M. Handcock and P. Janssen. Statistical inference for the relative density. *Sociol. Methods Res.*, 30(3):394–424, 2002.

M. Handcock and M. Morris. *Relative distribution methods in the social sciences*. Statistics for Social Science and Public Policy. Springer-Verlag, New York, 1999.

E. B. Holmgren. The p-p plot as a method for comparing treatment effects. *J. Amer. Statist. Assoc.*, 90:360–365, 1995.

F. Hsieh. The empirical process approach for semiparametric two-sample models with heterogeneous treatment effect. *J. Roy. Statist. Soc. Ser. B*, 57(4):735–748, 1995.

F. Hsieh and B. Turnbull. Nonparametric methods for evaluating diagnostic tests. *Statist. Sinica*, 6(1):47–62, 1996a.

F. Hsieh and B. Turnbull. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Ann. Statist.*, 24(1):25–40, 1996b.

G. Kerkyacharian and D. Picard. Density estimation in Besov spaces. *Statist. Probab. Lett.*, 13 (1):15–24, 1992.

G. Kerkyacharian and D. Picard. Density estimation by kernel and wavelets methods: optimality of Besov spaces. *Statist. Probab. Lett.*, 18(4):327–336, 1993.

J.-T. Kim. An order selection criterion for testing goodness of fit. *J. Amer. Statist. Assoc.*, 95 (451):829–835, 2000.

T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *Ann. Probab.*, 33(3):1060–1077, 2005.

A. Kolmogorov. Sulla determinazione empirica di une legge di distribuzione. *Giorn. dell'Istit. degli att.*, 4:83–91, 1933.

A. Kolmogorov. Confidence limits for an unknown distribution function. *Ann. Math. Stat.*, 12: 461–463, 1941.

S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statistics*, 22:79–86, 1951. ISSN 0003-4851.

C. Lacour. Adaptive estimation of the transition density of a particular hidden Markov chain. *J. Multivariate Anal.*, 99(5):787–814, 2008.

G. Li, R. Tiwari, and M. Wells. Quantile comparison functions in two-sample problems, with application to comparisons of diagnostic markers. *J. Amer. Statist. Assoc.*, 91(434):689–698, 1996.

C. Lloyd. Using smoothed ROC curves to summarize and compare diagnostic systems. *J. Amer. Statist. Assoc.*, 93(444):1356–228, 1998.

C. Lloyd and Z. Yong. Kernel estimators of the ROC curve are better than empirical. *Statist. Probab. Lett.*, 44(3):221–228, 1999.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statistics*, 18:50–60, 1947. ISSN 0003-4851.

J. Mielniczuk. Grade estimation of Kullback-Leibler information number. *Probab. Math. Statist.*, 13(1):139–147, 1992.

E. Molanes-López and R. Cao. Plug-in bandwidth selector for the kernel relative density estimator. *Ann. Inst. Statist. Math.*, 60(2):273–300, 2008a.

E. Molanes-López and R. Cao. Relative density estimation for left truncated and right censored data. *J. Nonparametr. Stat.*, 20(8):693–720, 2008b.

J.-C. Pardo-Fernández, M.-X. Rodríguez-Álvarez, and I. van Keilegom. A review on ROC curves in the presence of covariates. preprint, 2013. URL `http://www.uclouvain.be/cps/ucl/doc/stat/documents/DP2013_50.pdf`.

B. Silverman. Density ratios, empirical likelihood and cot death. *J. Roy. Statist. Soc. Ser. B*, 27 (1):26–33, 1978.

N. Smirnov. On the estimation of t.he discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscow*, 2(2):3–14, 1939.

N. Smirnov. Approximate laws of distribution of random variables from empirical data. *Uspehi Matem. Nauk*, 10:179–206, 1944.

M. Sugiyama, T. Suzuki, and T. Kanamori. Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation. *Ann. Inst. Statist. Math.*, 64(5): 1009–1044, 2012.

A. Tsybakov. *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

F. Wilcoxon. Individual comparisons by rank methods. *Biometrics*, 1:80–83, 1945.

M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural Comput.*, 25(5):1324–1370, 2013.