

OPTIMAL ADAPTIVE ESTIMATION OF THE RELATIVE DENSITY: SUPPLEMENTARY MATERIAL

GAËLLE CHAGNY^(A) AND CLAIRE LACOUR^(B)

ABSTRACT. This supplementary material provides additional details for the paper "Optimal adaptive estimation of the relative density". We first recall a precise definition of Besov spaces, which are used to compute the rates of convergence of the estimator. Then, we present further simulation results (reconstruction and risk computations), as well as details about the calibration of the penalty constant. Finally, complete proofs which are only sketched in the main article are available.

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

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

1. DEFINITION OF BESOV SPACES

For A a subset of \mathbb{R} , let us recall the definition of $B_{p,\infty}^\alpha(A)$. Let $A_h^k = \{x \in \mathbb{R}, x, x+h, \dots, x+kh \in A\}$. Next, for x in A_h^k , let

$$\Delta_h^k r(x) = \sum_{i=0}^k (-1)^{k-i} \binom{k}{i} r(x+ih)$$

the k th difference operator with step h . For $t > 0$, the moduli of smoothness of function r is given by

$$\omega_k(r, t) = \sup_{|h| \leq t} \left(\int_{A_h^k} |\Delta_h^k r(x)|^p dx \right)^{1/p}.$$

We say that r is in the Besov space $B_{p,\infty}^\alpha(A)$ if $\sup_{t>0} t^{-\alpha} \omega_k(r, t) < \infty$ for k the smaller integer larger than α . The Besov norm $\|\cdot\|_{\alpha,p}$ is defined by:

$$\|r\|_{\alpha,p} = \|r\|_p + \sup_{t>0} t^{-\alpha} \omega_k(r, t).$$

It can be proved that $\mathcal{B}_{p,\infty}^\alpha((0;1)) \subset \mathcal{B}_{2,\infty}^\alpha((0;1))$ for $p \geq 2$. This justifies that we restrict to $\mathcal{B}_{2,\infty}^\alpha((0;1))$.

2. ADDITIONAL SIMULATION RESULTS

This section is devoted to giving additional simulation results.

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.

2.1. Additional reconstructions and risk computations. In this section, we present simulation results for examples of data which complete the three models given in Section 4 of the main article. First, Table 1 gives the details of the MISE which are partly plotted in Figure 2 of the main article, for samples simulated from Example (2) described in Section 4.2. Then, we investigate other examples. The two samples $(X_{0,i_0})_{i_0=1,\dots,n_0}$ and $(X_i)_{i=1,\dots,n}$ comes from random variables X_0 and X respectively, with probability distributions described below.

- (3) The variable X_0 is from the Weibull distribution with parameters (2,3). We denote by W the corresponding c.d.f.. The variable X is built such that $X = W^{-1}(S)$, with S chosen from one of the three following distribution:
 - (a3) a beta distribution $\mathcal{B}(14, 17)$,
 - (b3) a mixture of V_1 with probability 4/5 and V_2 with probability 1/5, where V_1 is from $\mathcal{B}(14, 37)$ and V_2 is from $\mathcal{B}(14, 20)$,
 - (c3) a mixture of V_1 with probability 1/3 and V_2 with probability 2/3, where V_1 is from $\mathcal{B}(34, 15)$ and V_2 is from $\mathcal{B}(15, 30)$.
 In these three cases, the relative density r to recover is the density of the variable S . Example (b3) is the one considered to evaluate the influence of n and n_0 in Section 4.3 (main article).
- (4) The variable X_0 is chosen to have a uniform distribution in the set $(0; 1)$. The variable X fits one of the following models:
 - (a4) $(1/4)(U_1 + U_2 + U_3 + U_4)$ where U_j , $j = 1, \dots, 4$ are independent and uniform on $(0; 1)$,
 - (b4) a mixture of V_1 with probability 1/2 and V_2 with probability 1/2, where $V_1 = V/2$ and $V_2 = (V + 1)/2$, and V as for model (a),
 - (c4) a beta distribution with parameters 4 and 5 (denoted by $\mathcal{B}(4, 5)$),
 - (d4) a mixture of X_j , $j = 1, 2, 3$ with probability 1/3, where the X_j have respective distributions $\mathcal{B}(10, 5)$, $\mathcal{B}(7, 7)$ and $\mathcal{B}(5, 10)$,
 - (e4) a mixture of X_1 with probability 1/2 and X_2 with probability 1/2, where X_1 and X_2 have respective distributions $\mathcal{B}(15, 4)$ and $\mathcal{B}(5, 11)$.

Hence the investigated relative densities are defined as the densities of X , in these five examples.

The third set of examples is borrowed from Molanes-López and Cao (2008) and the fourth from Ćwik and Mielniczuk (1993). The true relative densities associated to each framework of these two sets are plotted in Figure 1: they are quite far from the uniform distribution, since the distributions of X and X_0 are not similar.

Figure 2 illustrates the stability of the method and shows beams of estimates $\hat{r}_{\tilde{m}}$: 10 estimators built from *i.i.d.* samples of data are plotted together with the true functions. The MISE, displayed in Table 2, are of the same order as the ones of Ćwik and Mielniczuk (1993) and of Molanes-López and Cao (2008), as announced in Section 4 of the main article. However, the fully data-driven model selection method we implement is also shown to be optimal in theory, from the non-asymptotic point of view.

2.2. Details about penalty calibration. The choice of the value 1 for the constant c_0 involved in the penalty term $\tilde{V}(m) = c_0(D_m/n + 4\|\hat{r}_{m^*}\|^2 D_m/n_0)$ (see (9), Section 3 of the main article) has not been commented in detail until now. Since it plays an important role for the quality of the adaptive estimation, we propose to briefly shed light on its choice.

As explained in the main article, the numerical value obtained in the proof is rough and useless in practice. We thus experiment the tuning by simulation. More precisely, we choose different models for the data, already used in the paper, and experiment several values for c_0 .

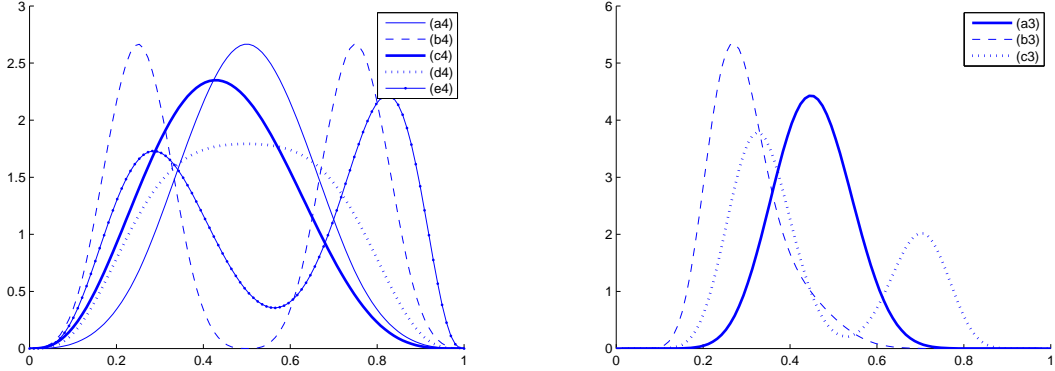


FIGURE 1. Plot of the different investigated relative densities of Examples (3) and (4)

The MISE is plotted with respect to possible values for c_0 , from 0.001 to 4. The results are given in Figure 3 for two sets of examples, Example (2) (a2) described in Section 4.2 of the main article and Example (4) described above (Section 2.1). Considering the graphs, the constant c_0 should not be chosen too small, neither too large. Actually, the adaptive index \tilde{m} is selected as the one which realizes the best trade-off between the estimator of the bias term $\tilde{A}(m)$ and the variance term $\tilde{V}(m)$. Then, heuristically, if c_0 is too large the compromise is made in favour of $\tilde{V}(m)$. As this term increases with m , we select too small models (over-penalization). On the contrary, if c_0 is too small, large models are selected. We decide to choose the value 1 which leads to reasonable risks. Another value in the range $[0.5; 1.5]$ just improves the results for some examples and deteriorates them for some others.

3. ADDITIONAL PROOFS

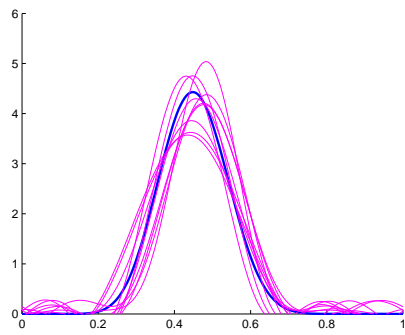
Complete proofs for technical results of the main article are provided in this section. The notations are the same as the one introduced in Section 5.1.1 of the main article. We also use the results established in Section 5.1.2, and more generally, we often refer to the proofs of Section 5.

3.1. Proof of Lemma 12. The aim is to bound $\mathbb{E}[\max_{m' \in \mathcal{M}} (T_2^{p_{m'}} - V_2(p_{m'}))_+]$, where $T_2^{p_{m'}}$ is defined by

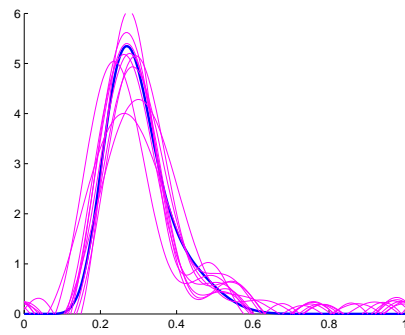
$$T_2^m = \left\| \mathbb{E} \left[\hat{r}_m(\cdot, \hat{F}_0) - \hat{r}_m(\cdot, F_0) \mid (X_0) \right] \right\|^2.$$

The proof is based on the proof of Lemma 9, Section 5.2.2 of the main article. Let us abbreviate $p_{m'}$ by p . We proceed as in this proof and obtain $T_2^p \leq 6T_{2,1,1}^p + 6T_{2,1,2}^p + 3T_{2,2}^p + 3T_{2,3}^p$. Thus,

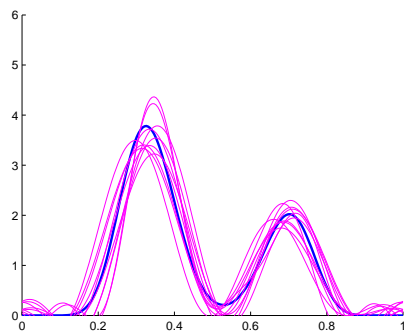
$$\begin{aligned} (1) \quad \mathbb{E} \left[\max_{m' \in \mathcal{M}} (T_2^p - V_2(p))_+ \right] &\leq \mathbb{E} \left[\max_{m' \in \mathcal{M}} \left(6T_{2,1,1}^p - V_2(p)/3 \right)_+ \right] + \mathbb{E} \left[\max_{m' \in \mathcal{M}} 6T_{2,1,2}^p \right] \\ &\quad + \mathbb{E} \left[\max_{m' \in \mathcal{M}} \left(3T_{2,2}^p - V_2(p)/3 \right)_+ \right] \\ &\quad + \mathbb{E} \left[\max_{m' \in \mathcal{M}} \left(3T_{2,3}^p - V_2(p)/3 \right)_+ \right]. \end{aligned}$$



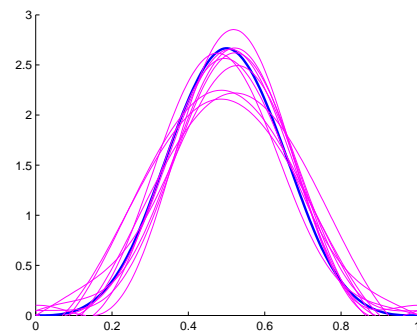
Example (a3)



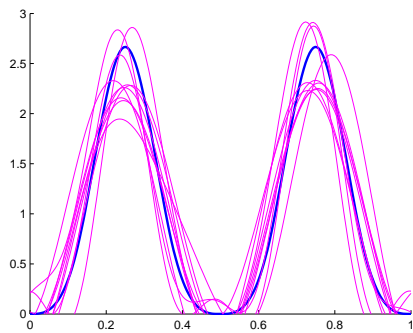
Example (b3)



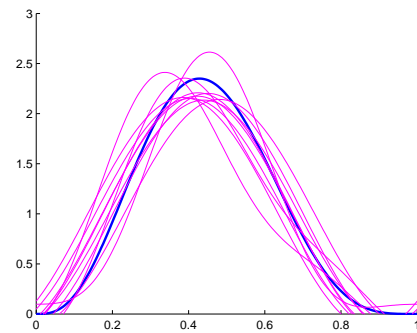
Example (c3)



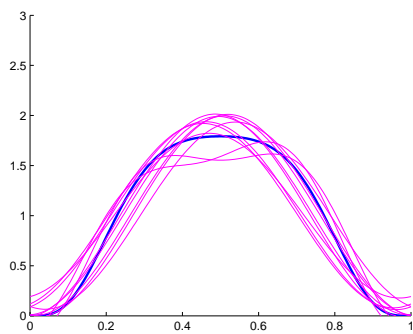
Example (a4)



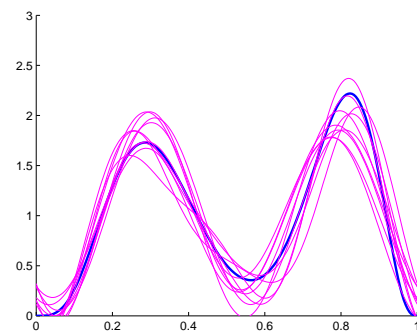
Example (b4)



Example (c4)



Example (d4)



Example (e4)

FIGURE 2. Beams of 10 estimators built from *i.i.d.* samples of size $n = n_0 = 500$ (thin lines) versus true function (thick line) in Examples (3) and (4).

Example (a2)

$n \setminus n_0$	50	100	200	400
50	0.220	0.168	0.162	0.196
100	0.088	0.123	0.137	0.085
200	0.141	0.103	0.083	0.089
400	0.105	0.092	0.057	0.039

$c = 1.01$

$n \setminus n_0$	50	100	200	400
50	0.221	0.212	0.166	0.220
100	0.132	0.137	0.126	0.132
200	0.139	0.134	0.113	0.089
400	0.137	0.146	0.071	0.066

$c = 1.05$

$n \setminus n_0$	50	100	200	400
50	0.288	0.296	0.308	0.275
100	0.236	0.236	0.222	0.215
200	0.211	0.199	0.199	0.162
400	0.227	0.191	0.152	0.150

$c = 1.1$

$n \setminus n_0$	50	100	200	400
50	2.200	1.842	1.496	1.357
100	1.990	1.383	1.126	0.878
200	1.778	1.149	0.918	0.769
400	1.717	1.081	0.786	0.720

$c = 1.5$

Example (b2)

$n \setminus n_0$	50	100	200	400
50	0.245	0.190	0.156	0.184
100	0.129	0.128	0.100	0.099
200	0.165	0.101	0.080	0.067
400	0.115	0.096	0.066	0.044

$a = 2.01$

$n \setminus n_0$	50	100	200	400
50	0.270	0.208	0.228	0.221
100	0.212	0.174	0.150	0.122
200	0.214	0.138	0.109	0.101
400	0.134	0.102	0.090	0.071

$a = 2.05$

$n \setminus n_0$	50	100	200	400
50	0.335	0.311	0.284	0.284
100	0.295	0.227	0.238	0.203
200	0.264	0.259	0.193	0.186
400	0.260	0.196	0.163	0.157

$a = 2.1$

$n \setminus n_0$	50	100	200	400
50	3.253	3.262	3.169	3.170
100	3.320	3.162	3.035	2.984
200	3.249	3.029	2.946	2.848
400	3.236	3.031	2.866	2.728

$a = 2.5$

TABLE 1. Values of MISE $\times 10$ averaged over 500 samples for the estimator $\hat{r}_{\hat{m}}$, in Examples (2) (a2) and (b2).

We do not subtract $V(p)$ to one of the term, since we immediately derive from Section 5.2.2

$$(2) \quad \mathbb{E} \left[\max_{m' \in \mathcal{M}} 6T_{2,1,2}^p \right] \leq \mathbb{E} \left[6T_{2,1,2}^{m_{\max}} \right] \leq 6C_2 \|r'\|^2 / n_0.$$

For the term depending on $T_{2,1,1}^p$, note that $T_{2,1,1}^p = \sum_{j=1}^{D_p} (\nu_{n_0}^b(\varphi_j))^2$, with

$$\nu_{n_0}^b(t) = \frac{1}{n_0} \sum_{i_0=1}^{n_0} \psi_t(X_{0,i_0}) - \mathbb{E}[\psi_t(X_{0,i})], \text{ with } \psi_t(x) = r(F_0(x))t(F_0(x)).$$

$n \setminus n_0$	50	100	200	400
50	10.27	6.117	3.743	2.468
100	9.958	5.394	3.496	2.208
200	9.100	5.651	3.204	1.883
400	9.439	5.303	3.190	1.849

Example (a3)

$n \setminus n_0$	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

Example (b3)

$n \setminus n_0$	50	100	200	400
50	9.041	6.026	3.904	2.855
100	8.179	5.396	3.319	1.929
200	8.113	5.055	2.915	1.659
400	7.886	4.855	2.726	1.589

Example (c3)

$n \setminus n_0$	50	100	200	400
50	2.712	1.870	1.362	1.110
100	2.561	1.765	1.126	0.863
200	2.576	1.647	1.114	0.729
400	2.567	1.630	0.989	0.636

Example (a4)

$n \setminus n_0$	50	100	200	400
50	5.833	3.513	2.288	1.741
100	5.377	3.203	2.008	1.446
200	5.392	3.217	1.925	1.286
400	4.609	3.086	1.927	1.195

Example (b4)

$n \setminus n_0$	50	100	200	400
50	2.172	1.383	0.952	0.799
100	1.966	1.270	0.798	0.544
200	2.064	1.135	0.696	0.481
400	1.928	1.147	0.681	0.424

Example (c4)

$n \setminus n_0$	50	100	200	400
50	2.153	1.305	0.926	0.747
100	1.768	1.064	0.714	0.549
200	1.510	0.929	0.597	0.441
400	1.585	0.911	0.549	0.377

Example (d4)

$n \setminus n_0$	50	100	200	400
50	4.364	2.838	2.043	1.590
100	3.731	2.159	1.372	0.969
200	3.511	1.810	1.109	0.816
400	3.272	1.649	0.982	0.675

Example (e4)

TABLE 2. Values of MISE $\times 10$ averaged over 500 samples for the estimator $\hat{r}_{\hat{m}}$, in Examples (3) and (4).

We proceed as in (18) (see the main article) to write $T_{2,1,1}^p = \sup_{t \in \mathcal{S}(p)} (\nu_{n_0}^b(t))^2$. We anew apply the Talagrand Inequality (Proposition 7, Section 5.1.3). We easily compute $M_1 = \|r\|_\infty \sqrt{D_p}$, and $v = \|r\|_\infty^2$. For H^2 , the same computations as in Lemma 9, Section 5.2.2, give

$$\mathbb{E} \left[\sup_{t \in \mathcal{S}(p)} \left(\nu_{n_0}^b(t) \right)^2 \right] = \mathbb{E} \left[T_{2,1,1}^p \right] \leq \|r\|^2 \frac{D_p}{n_0} := H^2.$$

The result is the following, with $V_{2,1,1}(p) = 6 \times 2(1 + 2\delta) \|r\|^2 D_p / n_0$, $\delta > 0$,

$$(3) \quad \mathbb{E} \left[\max_{m' \in \mathcal{M}} \left(6T_{2,1,1}^p - V_{2,1,1}(p) \right)_+ \right] \leq \frac{C}{n_0}.$$

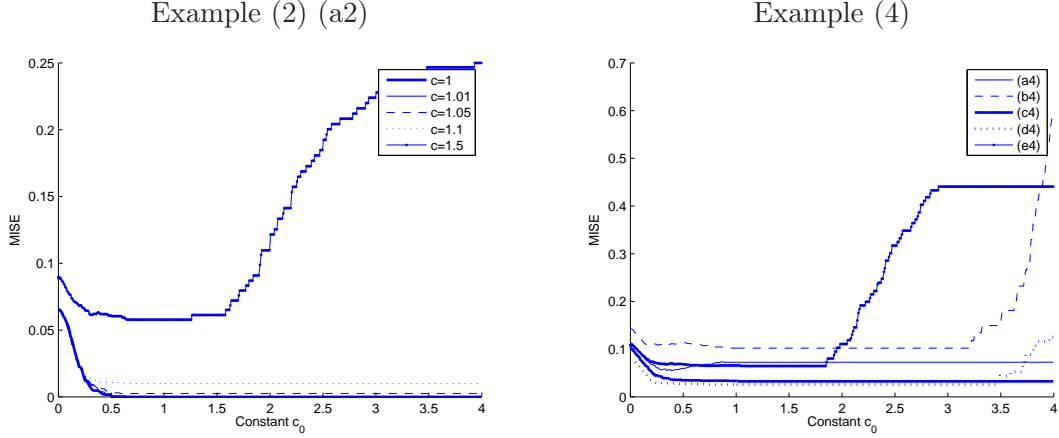


FIGURE 3. Values of the MISE (averaged over 500 samples) for $n = n_0 = 500$ with respect to the constant c_0 .

For the term in which $T_{2,2}^p$ is involved, we begin with

$$\mathbb{E} \left[\max_{m' \in \mathcal{M}} \left(3T_{2,2}^p - \frac{V_2(p)}{3} \right)_+ \right] \leq \sum_{m' \in \mathcal{M}} \mathbb{E} \left[\left(3T_{2,2}^p - \frac{V_2(p)}{3} \right)_+ \right],$$

and compute the right-hand side, for a fixed index p . We prove in Section 5.2.2 that $T_{2,2}^p \leq (\pi^4/4)\|r\|^2 D_p^4 \|\widehat{U}_0 - id\|_\infty^4$. Corollary 3 (Section 5.1.2) with $p = 4$ gives, for all $\kappa > 0$,

$$\mathbb{E} \left[\left(\|\widehat{U}_0 - id\|_\infty^4 - \kappa \frac{\ln^2(n_0)}{n_0^2} \right)_+ \right] \leq C n_0^{-\frac{1}{\sqrt{2}}\kappa^{1/2}}.$$

Therefore, denoting by $V_{2,2}(p) = (3\pi^4/4)\|r\|^2 \kappa \frac{D_p^4 \ln^2(n_0)}{n_0^2}$,

$$(4) \quad \mathbb{E} \left[\max_{m' \in \mathcal{M}} \left(3T_{2,2}^p - V_{2,2}(p) \right)_+ \right] \leq C \sum_{m' \in \mathcal{M}} D_p^4 n_0^{-\frac{1}{\sqrt{2}}\kappa^{1/2}}.$$

But we roughly bound $\sum_{m' \in \mathcal{M}} D_p^4 \leq D_{m_{\max}}^5 \leq n_0^5$. The right-hand side of (4) is thus bounded by $C n_0^{5-\sqrt{\kappa/2}}$, with c a constant, and this last bound is smaller than C/n_0 if κ is large enough. Since we assume $D_p \leq n_0^{1/3}/\ln^{2/3}(n)$, we have $V_{2,2}(p) \leq V_{2,2}^{bis}(p) = (3\pi^4/4)\|r\|^2 \kappa D_p/n_0$, and (4) is still true with $V_{2,2}$ replaced by $V_{2,2}^{bis}$.

We proceed similarly for the term which depends on $T_{2,3}^p$. We see in Section 5.2.2 that $T_{2,3}^p \leq (32\pi^6/9)\|r\|^2 D_p^7 \|\widehat{U}_0 - id\|_\infty^6$, and thanks to Corollary 3 with $p = 6$

$$\mathbb{E} \left[\left(\|\widehat{U}_0 - id\|_\infty^6 - \kappa \frac{\ln^3(n_0)}{n_0^3} \right)_+ \right] \leq C n_0^{-\frac{1}{2^{1/3}}\kappa^{2/3}}.$$

Thus, for $V_{2,3}(p) = (32\pi^6/9)\|r\|^2\kappa D_p^7 \ln^3(n_0)/n_0^3$,

$$(5) \quad \mathbb{E} \left[\max_{m' \in \mathcal{M}} \left(3T_{2,3}^p - V_{2,3}(p) \right)_+ \right] \leq C \sum_{m' \in \mathcal{M}} D_p^4 n_0^{-\frac{1}{2^{1/3}}\kappa^2/3}.$$

If κ is large enough, the right-hand side of (5) is bounded by C/n_0 , and $V_{2,3}$ can be replaced by an upper-bound, if $D_p \leq n_0^{1/3}/\ln^{1/2}(n)$: $V_{2,3}(p) \leq V_{2,3}^{bis}(p) = (32\pi^6/9)\|r\|^2\kappa D_p/n_0$.

We gather (2), (3), (4), and (5) in Inequality (1), and choose $V_2(p)$ with form $c_2\|r\|^2 D_p/n_0$ for c_2 large enough. □

3.2. Proof of Theorem 3. The aim is to prove an oracle-type inequality for the fully data-driven estimator $\hat{r}_{\tilde{m}}(\cdot, \hat{F}_0)$. We introduce the set

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

and split

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

We show in the sequel that the first term gives the order of the upper-bound of Theorem 3, and that the probability of the set Λ^c is negligible compared to $1/n + 1/n_0$.

• **Upper-bound for $\mathbb{E}[\|\hat{r}_{\tilde{m}}(\cdot, \hat{F}_0) - r\|^2 \mathbf{1}_\Lambda]$.** Arguing as in Section 5.3 (main article), we first obtain, for $m \in \mathcal{M}$

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

Moreover, $\tilde{A}(m) \leq A(m) + \max_{m' \in \mathcal{M}} (V(m') - \tilde{V}(m'))_+$. Thus

$$\begin{aligned} \left\| \hat{r}_{\tilde{m}}(\cdot, \hat{F}_0) - r \right\|^2 &\leq 6(A(m) + V(m)) + 3 \left\| \hat{r}_m(\cdot, \hat{F}_0) - r \right\|^2 \\ &\quad + \max_{m' \in \mathcal{M}} \left(V(m') - \tilde{V}(m') \right)_+ + 6 \left(\tilde{V}(m) - V(m) \right). \end{aligned}$$

For every $m \in \mathcal{M}$,

$$\tilde{V}(m) - V(m) = c_0 \frac{D_m}{n_0} \left(4 \|\hat{r}_{m^*}(\cdot, \hat{F}_0)\|^2 - \|r\|^2 \right).$$

On the set Λ , since $\|r\| < 2\|\hat{r}_{m^*}(\cdot, \hat{F}_0)\|$, we thus have $(V(m') - \tilde{V}(m'))_+ = 0$. On this set, we also have : $\|\hat{r}_{m^*}(\cdot, \hat{F}_0)\| \leq (3/2)\|r\|$,

$$\left(\tilde{V}(m) - V(m) \right) \leq c_0 \frac{D_m}{n_0} \left(4 \times \frac{9}{4} \|r\|^2 - \|r\|^2 \right) = 8c_0 \|r\|^2 \frac{D_m}{n_0}.$$

Using also Lemma 10 enables to conclude

$$\mathbb{E} \left[\left\| \hat{r}_{\tilde{m}}(\cdot, \hat{F}_0) - r \right\|^2 \mathbf{1}_\Lambda \right] \leq \min_{m \in \mathcal{M}} \left\{ \frac{D_m}{n} + \|r\|^2 \frac{D_m}{n_0} + 15 \|r_m - r\|^2 \right\} + \frac{C}{n} + \frac{C}{n_0}.$$

• **Upper-bound for $\mathbb{E}[\|\hat{r}_{\tilde{m}}(\cdot, \hat{F}_0) - r\|^2 \mathbf{1}_{\Lambda^c}]$.** First, $\|\hat{r}_{\tilde{m}}(\cdot, \hat{F}_0) - r\|^2 \leq 2\|\hat{r}_{\tilde{m}}(\cdot, \hat{F}_0)\|^2 + 2\|r\|^2$, and

$$\left\| \hat{r}_{\tilde{m}}(\cdot, \hat{F}_0) \right\|^2 = \sum_{j=1}^{D_{\tilde{m}}} \left(\hat{a}_j^{\hat{F}_0} \right)^2 \leq \sum_{j=1}^{D_{\tilde{m}}} \|\varphi_j^2\|_\infty \leq D_{\tilde{m}} \leq \min(n, n_0).$$

Thus

$$\mathbb{E} \left[\|\tilde{r}_{\hat{m}}(\cdot, \hat{F}_0) - r\|^2 \mathbf{1}_{\Lambda^c} \right] \leq \mathbb{P}(\Lambda^c) 2 (\min(n, n_0) + 2\|r\|^2).$$

It remains to bound $\mathbb{P}(\Lambda^c)$. We split

$$\mathbb{P}(\Lambda^c) \leq \mathbf{1}_{\{\|r - r_{m^*}\| \geq \frac{\|r\|}{4}\}} + \mathbb{P} \left(\|r_{m^*} - \hat{r}_{m^*}(\cdot, \hat{F}_0)\| \geq \frac{\|r\|}{4} \right).$$

Recall that r belongs to the Besov ball $B_{2,\infty}^\alpha((0;1), L)$, and that $D_{m^*} \geq \ln(n_0)$. Hence, $\|r - r_{m^*}\| \leq CD_{m^*}^{-\alpha} \leq C(\ln(n_0))^{-\alpha}$. This quantity goes to 0 when n_0 goes to $+\infty$. Therefore, $C(\ln(n_0))^{-\alpha} \leq \|r\|/4$ for n_0 large enough. Consequently,

$$(6) \quad \mathbf{1}_{\{\|r - r_{m^*}\| \geq \frac{\|r\|}{4}\}} = 0.$$

Thanks to (20) (Section 5.3.1, main article), we also have

$$\begin{aligned} \mathbb{P} \left(\|r_{m^*} - \hat{r}_{m^*}(\cdot, \hat{F}_0)\|^2 \geq \frac{\|r\|^2}{16} \right) &\leq \mathbb{P} \left(\sup_{t \in \mathcal{S}(m^*)} (\nu_n(t))^2 \geq \frac{\|r\|^2}{3 \times 16} \right) \\ &\quad + \mathbb{P} \left(4T_1^{m^*} \geq \frac{\|r\|^2}{3 \times 16} \right) + \mathbb{P} \left(4T_2^{m^*} \geq \frac{\|r\|^2}{3 \times 16} \right), \end{aligned}$$

with ν_n defined by (19)(Section 5.3.1) and $T_1^{m^*}$, $T_2^{m^*}$ by (13) (Section 5.2). We use (14) of Section 5.2.2, and the mean-value theorem to obtain, $T_2^{m^*} \leq 8\pi^2 D_{m^*}^3 \|\hat{U}_0 - id\|_\infty^2$. Thus

$$\begin{aligned} \mathbb{P} \left(4T_2^{m^*} \geq \frac{\|r\|^2}{3 \times 16} \right) &\leq \mathbb{P} \left(\|\hat{U}_0 - id\|_\infty^2 \geq \frac{\|r\|^2}{192 D_{m^*}^3 \times 8\pi^2} \right), \\ &\leq C \exp \left(-n_0 \frac{\|r\|^2}{768\pi^2 D_{m^*}^3} \right). \end{aligned}$$

by applying Proposition 5 (main article, Section 5.1.2). Since $D_{m^*}^3 \leq n_0/\ln^2(n_0)$, we have

$$(7) \quad \mathbb{P} \left(4T_2^{m^*} \geq \frac{\|r\|^2}{3 \times 16} \right) \leq C \exp \left(-\frac{\|r\|^2}{768\pi^2} (\ln(n_0))^2 \right).$$

The same arguments permit to bound the term in which $T_1^{m^*}$ is involved. We first note that $T_1^{m^*} \leq 32\pi^2 D_{m^*}^3 \sup_{x \in \mathbb{R}} |\hat{F}_0(x) - F_0(x)|^2$ and conclude with Proposition 5:

$$(8) \quad \mathbb{P} \left(4T_1^{m^*} \geq \frac{\|r\|^2}{48} \right) \leq C \exp \left(-\frac{\|r\|^2}{4 \times 768\pi^2} (\ln(n_0))^2 \right).$$

We go back to the term involving the empirical process:

$$\mathbb{P} \left(\sup_{t \in \mathcal{S}(m^*)} (\nu_n(t))^2 \geq \frac{\|r\|^2}{48} \right) \leq \sum_{j=1}^{D_{m^*}} \mathbb{P} \left(\nu_n^2(\varphi_j) \geq \frac{\|r\|^2}{48 D_{m^*}} \right).$$

Writing $\nu_n(\varphi_j) = (1/n) \sum_{i=1}^n Z_i^j - \mathbb{E}[Z_i^j]$ with $Z_i^j = \varphi_j(F_0(X_i))$ (see (19) of the main article) allows to apply the Bernstein Inequality (see for instance Proposition 2.9 and its comments in Massart (2007)). We compute $b = \sqrt{D_{m^*}}$, and $v = n\|r\|_\infty$. This leads, for $u > 0$

$$\mathbb{P} \left(\nu_n^2(\varphi_j) \geq \sqrt{2\|r\|_\infty} u + u \frac{1}{3} \sqrt{D_{m^*}} \right) \leq e^{-nu}.$$

Choosing $u = a/D_{m^*}^4$, for a constant a and using $D_{m^*} \geq \ln(n_0)$, we can obtain $\sqrt{2\|r\|_\infty u} + u\sqrt{D_{m^*}^4}/3 \leq \|r\|^2/48D_{m^*}$, for n_0 large enough, and

$$(9) \quad \sum_{j=1}^{D_{m^*}} \mathbb{P} \left(\nu_n^2(\varphi_j) \geq \frac{\|r\|^2}{48D_{m^*}} \right) \leq D_{m^*} \exp \left(-n \frac{a}{D_{m^*}^4} \right).$$

Putting together (6), (8), (7) and (9), we have proved

$$\mathbb{E} \left[\|\hat{r}_{\hat{m}}(\cdot, \hat{F}_0) - r\|^2 \mathbf{1}_{\Lambda^c} \right] \leq 2(\min(n, n_0) + 2\|r\|^2) \left\{ C \exp \left(-C(\ln(n_0))^2 \right) + D_{m^*} \exp \left(-\frac{na}{D_{m^*}^4} \right) \right\}.$$

Recall that $D_{m^*} \leq C(n/\ln(n))^{1/4}$. The last term of this upper-bound is thus negligible compared to $1/n$ (if a is large enough). The other term has the order $n_0 \exp(-C(\ln(n_0))^2)$, and are thus smaller than C/n_0 . Finally,

$$\mathbb{E} \left[\|\hat{r}_{\hat{m}}(\cdot, \hat{F}_0) - r\|^2 \mathbf{1}_{\Lambda^c} \right] \leq \frac{C}{n_0} + \frac{C}{n}.$$

□

3.3. Proof of Theorem 4. The aim is to prove the following lower bound for the minimax risk over Besov balls

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

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 separate two cases: $n \leq n_0$ and $n > n_0$. The result comes down to the proof of the two following inequalities:

$$(11) \quad \inf_{\hat{r}_{n,n_0} \in \mathcal{E}_n} \sup_{r \in \mathcal{F}_\alpha} \mathbb{E} \left[\|\hat{r}_{n,n_0} - r\|^2 \right] \geq c\phi_{n,n_0} = cn^{-2\alpha/(2\alpha+1)},$$

$$(12) \quad \inf_{\hat{r}_{n,n_0} \in \mathcal{E}_{n_0}} \sup_{r \in \mathcal{F}_\alpha} \mathbb{E} \left[\|\hat{r}_{n,n_0} - r\|^2 \right] \geq c\phi_{n,n_0} = cn_0^{-2\alpha/(2\alpha+1)},$$

where \mathcal{E}_n is the set of all possible estimators built with $(X_i)_{i \in \{1, \dots, n\}}$ and $(X_{0,i_0})_{i_0 \in \{1, \dots, n_0\}}$ when $n \leq n_0$, and \mathcal{E}_{n_0} the analogous set when $n > n_0$.

The proof of each of the Inequalities (11) and (12) 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}_α to a finite well-chosen subset $\{r_a, r_1, \dots, r_M\}$, $M \geq 2$ such that

- (i) $r_l \in \mathcal{F}_\alpha$, for every $l \in \{a, 1, \dots, M\}$.
- (ii) $\|r_l - r_{l'}\|^2 \geq 2B_0\phi_{n,n_0}$, for every $l, l' \in \{a, 1, \dots, M\}$, $l \neq l'$, for a constant $B_0 > 0$.
- (iii) For every $l \in \{1, \dots, M\}$, $\mathbb{P}_l^{(n,n_0)}$ is absolutely continuous with respect to $\mathbb{P}_a^{(n,n_0)}$ and there exists a constant $\kappa \in (0; 1/8)$, such that

$$\frac{1}{M} \sum_{l=1}^M K \left(P_l^{(n,n_0)}, P_a^{(n,n_0)} \right) \leq \kappa \log(M),$$

with $K(\cdot, \cdot)$ the Kullback-Leibler divergence, $\mathbb{P}_l^{(n,n_0)}$ (resp. $\mathbb{P}_a^{(n,n_0)}$) the probability distribution of a double sample $(X_i)_{i \in \{1, \dots, n\}}$ and $(X_{0,i_0})_{i_0 \in \{1, \dots, n_0\}}$ with relative density r_l (resp. r_a).

If we construct a set of functions which satisfies (i), (ii), and (iii) in each of the cases $n \leq n_0$ and $n > n_0$, Inequalities (11) and (12), and thus Theorem 4, are a consequence from Theorem 2.5 of Tsybakov (2009).

3.3.1. *Construction of the subset $\{r_a, r_1, \dots, r_M\}$.* It is in the spirit of Härdle et al. (1998) (Section 10.4, Chapter 10). First let $r_a = \mathbf{1}_{(0;1)}$. Let also ψ be a regular wavelet with compact support, and $\psi_{j,k} : x \mapsto 2^{j/2}\psi(2^j x - k)$, for $(j, k) \in \mathbb{Z}^2$. We define R_j the maximal subset of \mathbb{Z} such that $\text{supp } \psi_{j,k} \subset [0, 1]$, $k \in R_j$ and that $\text{supp } \psi_{j,k} \cap \text{supp } \psi_{j,k'} = \emptyset$ if $k \neq k'$. The cardinal of R_j is $|R_j| = c2^j$, for j an integer and c a constant, both to be chosen below. For all $\varepsilon = (\varepsilon_k)_{k \in R_j} \in \{0, 1\}^{|R_j|}$, consider

$$r_\varepsilon = r_a + \chi_\varepsilon, \quad \text{with} \quad \chi_\varepsilon = \gamma_{n,n_0} \sum_{k \in R_j} \varepsilon_k \psi_{j,k},$$

for γ_{n,n_0} a nonnegative number (decreasing when $\min(n, n_0)$ goes to ∞) defined below.

Now, from the Varshamov-Gilbert bound (see Lemma 2.9 of Tsybakov 2009), there exist a finite subset $\{\varepsilon^{(0)}, \dots, \varepsilon^{(M)}\}$ of $\{0, 1\}^{|R_j|}$, such that $\varepsilon^{(0)} = (0, \dots, 0)$, and

$$(13) \quad \rho(\varepsilon^{(l)}, \varepsilon^{(l')}) := \sum_{k \in R_j} \mathbf{1}_{\varepsilon_k^{(l)} \neq \varepsilon_k^{(l')}} \geq \frac{|R_j|}{8} = \frac{c2^j}{8}, \quad \text{and} \quad M \geq 2^{|R_j|/8} = c2^{2^j/8}.$$

We set $r_l = r_{\varepsilon^{(l)}}$, $l \in \{1, \dots, M\}$ and remark that $r_a = r_{\varepsilon^{(0)}}$. In the sequel, we establish the conditions to adjust j and γ_{n,n_0} such that (i), (ii), and (iii) are verified for the set $\{r_a, r_l, l \in \{1, 2, \dots, M\}\}$. The computations are mainly the same to prove Inequalities (11) and (12), except to check (iii). Thus, we distinguish the two cases only at the end, to conclude.

3.3.2. *Conditions which guarantees (i).* The function r_a is a relative density with support $(0; 1)$ (density supported by $(0; 1)$), and $\|r_a\|_{\alpha,2} = \|r_a\| = 1 < L$. Moreover, to have $r_\varepsilon(x) \geq 0$ for $x \in (0; 1)$, we must suppose $\|\chi_\varepsilon\|_\infty \leq 1$, that is

$$(14) \quad \gamma_{n,n_0} 2^{j/2} \leq \frac{1}{\|\psi\|_\infty}.$$

Since $\int_{\mathbb{R}} \psi(x) dx = 0$, we have $\int_{\mathbb{R}} \chi_\varepsilon(x) dx = 0$, and thus, $\int_{\mathbb{R}} r_\varepsilon(x) dx = 1$. Therefore, if (14) holds, r_ε is also a relative density, for all $\varepsilon \in \{0, 1\}^{|R_j|}$. According to Hochmuth (2002) (Theorem 3.5), $\|\sum_{k \in R_j} \varepsilon_k \psi_{j,k}\|_{\alpha,2} \leq 2^{j\alpha} \|\sum_{k \in R_j} \varepsilon_k \psi_{j,k}\|$, if ψ smooth enough. Since $\|\sum_{k \in R_j} \varepsilon_k \psi_{j,k}\| \leq |R_j| = 2^j$, we deduce

$$\|r_\varepsilon\|_{\alpha,2} \leq \|r_a\|_{\alpha,2} + \|\chi_\varepsilon\|_{\alpha,2} \leq 1 + \gamma_{n,n_0} 2^{j\alpha} 2^{j/2} \leq L,$$

as soon as the following condition is satisfied:

$$(15) \quad \gamma_{n,n_0} 2^{j\alpha} 2^{j/2} \leq L - 1.$$

3.3.3. *Conditions which guarantees (ii).* Let $l, l' \in \{0, 1, \dots, M\}$. We compute

$$\|r_l - r_{l'}\|^2 = \gamma_{n,n_0}^2 \left\| \sum_{k \in R_j} (\varepsilon_k^{(l)} - \varepsilon_k^{(l')}) \psi_{j,k} \right\|^2 = \gamma_{n,n_0}^2 \sum_{k \in R_j} (\varepsilon_k^{(l)} - \varepsilon_k^{(l')})^2 \|\psi\|^2.$$

Thanks to (13), $\|r_l - r_{l'}\|^2 = \gamma_{n,n_0}^2 \rho(\varepsilon^{(l)}, \varepsilon^{(l')}) \|\psi\|^2 \geq 2^j \gamma_{n,n_0}^2 \|\psi\|^2 / 8$. The condition (ii) is thus fulfilled as soon as

$$(16) \quad \gamma_{n,n_0}^2 2^j \geq \frac{16B_0}{\|\psi\|^2} \phi_{n,n_0}.$$

3.3.4. *Conditions which guarantees (iii) and conclusion.* For $l \in \{a, 1, \dots, M\}$, denote by (X_0^l, X^l) a couple of independent random variables with relative density r_l , probability distribution $\mathbb{P}_l = \mathbb{P}_{X_0^l} \otimes \mathbb{P}_{X^l}$, marginal density f^l and f_0^l . If $l = a$, remark that the definition of r_a requires $f_0^a = f^a$. If $((X_{0,i_0}^l)_{i_0=1,\dots,n_0}, (X_i^l)_{i=1,\dots,n})$ is a double sample distributed as X_0^l and X^l , then its distribution is $\mathbb{P}_l^{(n,n_0)} = \mathbb{P}_{X^l}^{\otimes n} \otimes \mathbb{P}_{X_0^l}^{\otimes n_0}$. Let now $l \in \{1, \dots, M\}$ be fixed. The following decomposition holds

$$(17) \quad K\left(\mathbb{P}_l^{(n,n_0)}, \mathbb{P}_a^{(n,n_0)}\right) = nK\left(\mathbb{P}_{X^l}, \mathbb{P}_{X^a}\right) + n_0K\left(\mathbb{P}_{X_0^l}, \mathbb{P}_{X_0^a}\right).$$

To satisfy (iii), we check that it is possible to obtain $K\left(\mathbb{P}_l^{(n,n_0)}, \mathbb{P}_a^{(n,n_0)}\right) \leq 2^j \kappa \log(2)/8$. This is sufficient since $\log(M) \geq 2^j \log(2)/8$ (see (13)). We have now to distinguish the two cases $n \leq n_0$ and $n > n_0$ to end the proof: we choose the density functions of X_0^a, X_a, X_0^l, X^l to only keep the term depending on n in (17) when $\min(n, n_0) = n$, and to only keep the term depending on n_0 in the other case.

End of the proof of Inequality (12). Assume $n \leq n_0$, such that $\min(n, n_0) = n$. We set $f_0^a = f^a = f_0^l = \mathbf{1}_{(0,1)}$, and $f^l = r_l$. These choices guarantee that (X_0^p, X^p) has the relative density r_p , for $p = a$ and $p = l$, and, using (17), lead to

$$\begin{aligned} K\left(\mathbb{P}_l^{(n,n_0)}, \mathbb{P}_a^{(n,n_0)}\right) &= n \int_{\mathbb{R}} \log\left(\frac{f^l}{f^a}(x)\right) f^l(x) dx = n \int_{\mathbb{R}} \log(r_l(x)) r_l(x) dx \\ &= n \int_{\mathbb{R}} \log(1 + \chi_{\varepsilon(l)}(x)) (1 + \chi_{\varepsilon(l)}(x)) dx. \end{aligned}$$

Noting that $\log(1 + u) \leq u$ for every $u > -1$, and using $\int_{\mathbb{R}} \chi_{\varepsilon(l)}(x) dx = 0$, we obtain

$$K\left(\mathbb{P}_l^{(n,n_0)}, \mathbb{P}_a^{(n,n_0)}\right) \leq n \int_{\mathbb{R}} \chi_{\varepsilon(l)}^2(x) dx \leq n \gamma_{n,n_0}^2 \|\psi\|^2 2^j.$$

Thus to fulfill (iii) it is sufficient to have

$$(18) \quad n \gamma_{n,n_0}^2 \|\psi\|^2 \leq \kappa \frac{\log(2)}{8}.$$

Now the parameters γ_{n,n_0} and 2^j are chosen so that the conditions (14), (15), (16) and (18) are satisfied. We set, for two constants $b, c_0 > 0$,

$$(19) \quad \gamma_{n,n_0} = \sqrt{\frac{b}{\min(n, n_0)}} \quad \text{and} \quad 2^j = c_0 (\min(n, n_0))^{1/(2\alpha+1)}.$$

With the choices $b = \log(2)/(128\|\psi\|^2)$, $c_0 = \min(((L-1)/\sqrt{b})^{2/(2\alpha+1)}, (\|\psi\|_{\infty}/\sqrt{b})^{-2})$, we check that the three conditions are verified for any $B_0 < bc_0\|\psi\|^2/16$, and for $\kappa = 1/16$. This concludes the proof of Inequality (11).

End of the proof of Inequality (11). Assume $n > n_0$. The choices are now $f_0^a = f^a = f^l = r_l$, and $f_0^l = \mathbf{1}_{(0,1)}$, which also lead to the relative density r_p for (X_0^p, X^p) , $p = a$ and $p = l$. Here,

$$K\left(\mathbb{P}_l^{(n,n_0)}, \mathbb{P}_a^{(n,n_0)}\right) = n_0 \int_{\mathbb{R}} \log\left(\frac{f_0^l}{f_0^a}(x)\right) f_0^l(x) dx = n_0 \int_{\mathbb{R}} \log\left(\frac{1}{r_l(x)}\right) dx,$$

We now notice that $\log(1 + u) \geq u - u^2/2$, for $u \geq -1/2$. We can assume that $\chi_{\varepsilon(l)}(x) \geq -1/2$ (even if it means reducing the choice of the constant involved in γ_{n,n_0} , see below), and the same computations as in the case $n \leq n_0$ permits to obtain

$$K\left(\mathbb{P}_l^{(n,n_0)}, \mathbb{P}_a^{(n,n_0)}\right) \leq n_0 \gamma_{n,n_0}^2 \|\psi\|^2 2^j,$$

and the new condition

$$(20) \quad n_0 \gamma_{n, n_0}^2 \|\psi\|^2 \leq \kappa \frac{\log(2)}{8}.$$

The parameters (19) are still suitable to guarantee (15), (16) and (20). In this case, the constant b need to be small enough so that $\chi_{\varepsilon^{(l)}}(x) \geq -1/2$, which is possible. We conclude by checking the three conditions, which leads to Inequality (12), and ends the proof of Theorem 4. \square

REFERENCES

- J. Ćwik and J. Mielniczuk. Data-dependent bandwidth choice for a grade density kernel estimate. *Statist. Probab. Lett.*, 16(5):397–405, 1993.
- W. Härdle, G. Kerkycharian, D. Picard, and A. Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1998.
- R. Hochmuth. Wavelet characterizations for anisotropic Besov spaces. *Appl. Comput. Harmon. Anal.*, 12(2):179–208, 2002.
- P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.
- E. Molanes-López and R. Cao. Plug-in bandwidth selector for the kernel relative density estimator. *Ann. Inst. Statist. Math.*, 60(2):273–300, 2008.
- 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.