

Aggregation of density estimators and dimension reduction

Samarov, Alexander *

University of Massachusetts-Lowell and MIT,
77 Massachusetts Avenue, Cambridge, MA 02139-4307, U.S.A.
samarov@mit.edu

Tsybakov, Alexandre

Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VI
tsybakov@ccr.jussieu.fr

November 18, 2005

Abstract

We consider the problem of model-selection-type aggregation of arbitrary density estimators using MISE risk. Given a collection of arbitrary density estimators, we propose a data-based selector of the best estimator in the collection and prove a general ready-to-use oracle inequality for the selected aggregate estimator. We then apply this inequality to the adaptive estimation of a multivariate density in a “multiple index” model. We show that the proposed aggregate estimator adapts to the unknown index space of unknown dimension in the sense that it allows us to estimate the density with the optimal rate attainable when the index space is known.

AMS 2000 subject classifications: 62G07, 62G20

Key words and phrases: Nonparametric density estimation, aggregation of estimators, dimensionality reduction models

*The research of this author was supported by the NSF Grant DMS 0505561

1 Introduction

The problem of aggregation of M arbitrary estimators has been recently studied by many authors (see, e.g., Nemirovski (2000), Yang (2000), Devroye and Lugosi (2000), Catoni (2004), Wegkamp (2003), Tsybakov (2003), Birgé (2003), Bunea, Tsybakov and Wegkamp (2004), Rigollet and Tsybakov (2004) and the references cited therein). A motivating factor is that in frequently used statistical models (such as regression or density estimation) there exists a great variety of possible competing estimators, and it is often difficult to decide which estimator to choose. Assume that a Statistician is given a list of size M of such estimators: p_1, \dots, p_M . A natural idea is then to look for a new, improved, estimator constructed by combining p_1, \dots, p_M in a suitable way. A combined “super-estimator” obtained from p_1, \dots, p_M is usually called *aggregate* and its construction is called aggregation.

One can distinguish between three main types of aggregation: model selection (MS) aggregation, convex (C) aggregation and linear (L) aggregation. The objective of (MS) is to select the optimal single estimator from the list; that of (C) is to select the optimal convex combination of the given estimators; and that of (L) is to select the optimal linear combination of the given estimators. The notion of optimality mentioned here is defined with respect to a given risk function, and it can be formalized in a minimax sense leading to the concept of optimal rates of aggregation (Tsybakov (2003)). A standard approach to establishing this kind of optimality is to show that the aggregate satisfies a sufficiently precise oracle inequality.

Most of the currently available results on aggregation were obtained for the regression model (see a recent overview in Bunea, Tsybakov and Wegkamp (2004)). The literature on aggregation of density estimators is not as large: Catoni (2004) and Yang (2000) investigated the (MS) aggregation with the Kullback-Leibler divergence as a loss function; Devroye and Lugosi (2000) developed a method of (MS) aggregation of density estimators under the L_1 loss. Another approach to density aggregation under the L_1 loss was proposed by Birgé (2003). Finally, we mention the recent paper of Rigollet and Tsybakov (2004) on optimal convex (C) and linear (L) aggregation of density estimators under the L_2 loss, and the work of Juditsky, Nazin, Tsybakov and Vayatis (2005a,b) where a recursive aggregation procedure is proposed for various statistical contexts, including density estimation, classification and regression.

In this paper we consider the (MS) aggregation of arbitrary density estimators under the L_2 loss (MISE). The main precursor of our study is the paper of Wegkamp (1999) who treated a more particular problem of bandwidth selection for kernel density estimation, but some of his results can be interpreted in general aggregation framework. For instance, some oracle inequalities can be deduced from Wegkamp's work, although he does not derive them explicitly. Our first aim is to obtain a ready-to-use oracle inequality for the L_2 (MS) aggregation using techniques that are somewhat different from those of Wegkamp (1999). Then we consider an example of application of this inequality, namely, to the adaptive estimation of a multivariate density in a *multiple index model*. We show that the proposed aggregate adapts to the unknown index matrix B in the sense that it allows to estimate the density with the optimal rate attainable when B is known.

2 A density aggregation theorem

Let X_1, \dots, X_n be i.i.d. random vectors with common probability density p on \mathbf{R}^d . Suppose that we are given M candidate estimators p_1, \dots, p_M of the density p based on the sample X_1, \dots, X_n . Our goal here is the model selection (MS) aggregation, that is, we would like to choose $\tilde{N} \in \{1, \dots, M\}$, a random index based on the data, such that the aggregate $p_{\tilde{N}}$ satisfies an oracle inequality of the form

$$\mathbf{E}\|p_{\tilde{N}} - p\|^2 \leq (1 + \delta_n) \min_{1 \leq N \leq M} \mathbf{E}\|p_N - p\|^2 + r_n, \quad (1)$$

where the value $\delta_n = \delta_{n,M} > 0$ and the remainder term $r_n = r_{n,M} > 0$ are small enough (they tend to 0, as $n \rightarrow \infty$), and

$$\|p\| = \left(\int p^2 \right)^{1/2} = \left(\int_{\mathbf{R}^d} p^2(x) dx \right)^{1/2}.$$

We interpret the inequality (1) as the fact that the aggregate $p_{\tilde{N}}$ mimics asymptotically the best among the estimators p_1, \dots, p_M (in the sense of MISE), up to a small remainder term. Note that here p_1, \dots, p_M are arbitrary estimators, not necessarily belonging to a specific family of nonparametric estimators. In particular, some estimators in the list can be parametric and others can be nonparametric of different nature (kernel, spline, wavelet etc.). To apply the inequality (1) in the nonparametric

density estimation context, it is usually sufficient that the remainder r_n were smaller in order than the standard nonparametric MISE rates, for example, $r_n = (\log n)^a/n$ for some $a > 0$. This will be the case in the result that we prove below.

In order to define a specific aggregation algorithm, we split the sample X_1, \dots, X_n into two parts: I_1 , used for constructing “base” estimators p_N , and I_2 , used for their aggregation. Let $n_1 = \text{Card}(I_1)$, $n_2 = \text{Card}(I_2)$, $n = n_1 + n_2$. We select \tilde{N} using the rule:

$$\tilde{N} = \arg \min_{1 \leq N \leq M} J_N, \quad (2)$$

where

$$J_N = -\frac{2}{n_2} \sum_{I_2} p_N(X_i) + \int p_N^2. \quad (3)$$

Here and later we abbreviate $\sum_{X_i \in I_2} = \sum_{I_2}$. Note that, because sub-samples I_1 and I_2 are independent,

$$\mathbf{E} \left(\frac{1}{n_2} \sum_{I_2} p_N(X_i) \right) = \mathbf{E} \left(\int p_N(x)p(x)dx \right). \quad (4)$$

Therefore, J_N is such that

$$\mathbf{E}(J_N) = \mathbf{E}\|p - p_N\|^2 - \|p\|^2, \quad N = 1, \dots, M,$$

i.e. J_N is an unbiased estimator of the MISE of p_N , up to the summand $\|p\|^2$ free from N .

To state the aggregation theorem, we need the following assumptions.

Assumption 1. *There exist finite positive constants a_1, a_2 , and C_1, C_2 such that*

$$\sum_{N=1}^M \mathbf{E}\|p_N - p\| \leq C_1 n^{a_1} \quad (5)$$

with $M \geq 2$ satisfying

$$M \geq C_2 n^{a_2}. \quad (6)$$

Assumption 2. *There exists a finite constant C_3 and a constant $\gamma_0 \leq 1/12$ such that*

$$\sum_{N=1}^M \mathbf{E} \left[\|p_N - p\|_\infty \exp \left(-\frac{\gamma_0 \log^{7/4} M}{\|p_N - p\|_\infty} \right) \right] \leq C_3 \log^2 M, \quad (7)$$

where $\|f\|_\infty = \sup_{x \in D} |f(x)|$ and $D \in \mathbf{R}^d$ is the support of the density $p(\cdot)$.

Assumption 3. *The density p is uniformly bounded: there exists a constant $p_{max} < \infty$ such that $\|p\|_\infty \leq p_{max}$.*

Remark 1. Assumptions 1 – 3 are not very restrictive. First of all, note that the (MS) aggregation has the largest oracle risk and the smallest order of the remainder term among the three types of aggregation mentioned in the introduction (Tsybakov (2003), see also Bunea, Tsybakov and Wegkamp (2004), where these issues are discussed for the regression model). Therefore, it is not crucial to use (MS) aggregation when the number M of base estimators is small, for example, when M grows as a power of $\log n$. In this case one can efficiently mimic more powerful convex or linear oracles (Rigollet and Tsybakov (2004)). However, if the number M of estimators to aggregate is polynomial in n or bigger, the remainder terms of convex and linear aggregation become too large as compared to the typical nonparametric MISE rates. This does not happen for the (MS) aggregation remainder term. Therefore, the (MS) aggregation is the type of aggregation which is especially important for polynomial M , explaining why assumption (6) is natural.

Given (6), the assumption (5) is almost trivially satisfied: it suffices to have the risks $\mathbf{E}\|p_N - p\|$ uniformly bounded and M bounded by a power of n . Typically p_N are consistent with rates, and we have even a stronger bound.

Finally, Assumption 2 looks rather technical, but it is also quite a mild one. For example, it is satisfied if

$$\max_{N=1, \dots, M} \mathbf{E} \left[\|p_N - p\|_\infty I(\|p_N - p\|_\infty > \gamma_0 \log^{3/4} M) \right] \leq \frac{\log^2 M}{M}, \quad (8)$$

where $I(\cdot)$ denotes the indicator function. Below we give examples showing that (8) is not a restrictive condition in density estimation. For instance, a sufficient condition for (8) is that the probability $\mathbf{P}(\|p_N - p\|_\infty > t)$ decreases exponentially in t , as $t \rightarrow \infty$ (an example is given in Section 3), but often it suffices to check a weaker and quite natural condition that the deviation of the stochastic part of the estimator $\mathbf{P}(\|p_N - \mathbf{E}p_N\|_\infty > t)$ is exponentially small (see the example below).

To show that (8) implies (7), define the event $W = \{\|p_N - p\|_\infty \leq \gamma_0 \log^{3/4} M\}$

and write

$$\begin{aligned}
& \mathbf{E} \left[\left\| p_N - p \right\|_\infty \exp \left(-\frac{\gamma_0 \log^{7/4} M}{\left\| p_N - p \right\|_\infty} \right) \right] \\
& \leq \frac{\gamma_0}{M} \log^{3/4} M + \mathbf{E} \left[\left\| p_N - p \right\|_\infty \exp \left(-\frac{\gamma_0 \log^{7/4} M}{\left\| p_N - p \right\|_\infty} \right) I(W^c) \right] \\
& \leq \frac{\gamma_0}{M} \log^{3/4} M + \mathbf{E} \left[\left\| p_N - p \right\|_\infty I(\left\| p_N - p \right\|_\infty > \gamma_0 \log^{3/4} M) \right].
\end{aligned}$$

Consider a simple example illustrating that (8) is indeed a mild assumption: let p be supported on $[0, 1]$ and let p_N be a kernel density estimator with bandwidth $h_N > 0$ and with a bounded Lipschitz continuous kernel $K \geq 0$ such that $\int K = 1$:

$$p_N(x) = \frac{1}{nh_N} \sum_{i=1}^n K \left(\frac{X_i - x}{h_N} \right), \quad N = 1, \dots, M.$$

Then, clearly, $\|\mathbf{E}p_N\|_\infty \leq p_{\max}$ and $\|p_N\|_\infty \leq D_1/h_{\min}$ where $D_1 > 0$ is a constant and $h_{\min} = \min\{h_1, \dots, h_M\}$. Hence $\|p_N - p\|_\infty \leq 2p_{\max} + \|p_N - \mathbf{E}p_N\|_\infty$ and $\|p_N - p\|_\infty \leq D_1/h_{\min} + p_{\max}$, so that we get $\mathbf{E} \left[\|p_N - p\|_\infty I(\|p_N - p\|_\infty > \gamma_0 \log^{3/4} M) \right] \leq (D_1/h_{\min} + p_{\max}) \mathbf{P}(\|p_N - \mathbf{E}p_N\|_\infty > D_2 \log^{3/4} M)$ with some constant $D_2 > 0$. Now, using Bernstein's inequality, the Lipschitz condition on K and bounding $\|p_N - \mathbf{E}p_N\|_\infty$ by the maximum over a fine enough grid on $[0, 1]$ with step $n^{-\alpha}$ for some large enough $\alpha > 0$ we get the bound on the probability $\mathbf{P}(\|p_N - \mathbf{E}p_N\|_\infty > D_2 \log^{3/4} M) \leq D_3 n^\alpha \exp(-D_4 nh_N \log^{3/4} M) \leq D_3 n^\alpha \exp(-D_4 nh_{\min} \log^{3/4} M)$ with some constants $D_3, D_4 > 0$. Finally, if $M \asymp n^a$ with $a > 0$ and if the bandwidths are such that $h_{\min} \geq n^{-1} \log^{3/4} n$ we get the bound $\mathbf{E} \left[\|p_N - p\|_\infty I(\|p_N - p\|_\infty > \gamma_0 \log^{3/4} M) \right] \leq D_5 M^{a'} \exp(-D_6 \log^{3/2} M)$ with some constants $D_5, D_6, a' > 0$, which implies (8) for n large enough. Thus, Assumption 2 holds under quite standard conditions on the kernel K and on the bandwidths h_N .

Theorem 1 *If $n_2 = \lfloor \frac{cn}{\log M} \rfloor$ for some constant $c > 0$ such that $1 \leq n_2 < n$, then, under Assumptions 1 – 3, we have*

$$\mathbf{E} \|p_{\tilde{N}} - p\|^2 \leq \left(1 + \frac{C^*}{\log^{1/4} M} \right) \min_{1 \leq N \leq M} \mathbf{E} \|p_N - p\|^2 + C^* \frac{\log^3 M}{n}, \quad (9)$$

where $C^* > 0$ is a constant which depends only on $p_{\max}, a_1, a_2, C_1, C_2, C_3, c$.

Proof. Note first that, by definition, $J_{\tilde{N}} \leq J_N$ for all $1 \leq N \leq M$. Using this and (4), we have

$$\begin{aligned}
\mathbf{E}\|p_{\tilde{N}} - p\|^2 - \mathbf{E}\|p_N - p\|^2 &= \mathbf{E}\left(-2 \int pp_{\tilde{N}} + \int p_{\tilde{N}}^2\right) - \mathbf{E}\left(-\frac{2}{n_2} \sum_{I_2} p_N(X_i) + \int p_N^2\right) \\
&= \mathbf{E}(J_{\tilde{N}}) - \mathbf{E}(J_N) + \mathbf{E}\left(\frac{2}{n_2} \sum_{I_2} p_{\tilde{N}}(X_i) - 2 \int pp_{\tilde{N}}\right) \\
&\leq 2\mathbf{E}\left(\frac{1}{n_2} \sum_{I_2} p_{\tilde{N}}(X_i) - \int pp_{\tilde{N}}\right) \\
&= 2\mathbf{E}[Z_{\tilde{N}}], \tag{10}
\end{aligned}$$

where

$$Z_N \triangleq \frac{1}{n_2} \sum_{I_2} (p_N(X_i) - p(X_i)) - \left(\int pp_N - \int p^2\right).$$

Set $W_N = \gamma(\|p_N - p\|^2 + r)$, where $r = (\log M)^2/n_2$ and $\gamma > 0$ will be chosen later. Denoting by $I(A)$ the indicator of a set A , we have

$$\begin{aligned}
\mathbf{E}(|Z_N|) &\leq \mathbf{E}(|Z_N|I(|Z_N| < W_N)) + \mathbf{E}(|Z_N|I(|Z_N| \geq W_N)) \\
&\leq \gamma\mathbf{E}[\|p_{\tilde{N}} - p\|^2 + r] + \mathbf{E}(|Z_N|I(|Z_N| \geq W_N)) \\
&\leq \gamma\mathbf{E}\|p_{\tilde{N}} - p\|^2 + \gamma r + \sum_{N=1}^M \mathbf{E}(|Z_N|I(|Z_N| \geq W_N)). \tag{11}
\end{aligned}$$

Now,

$$\mathbf{E}(|Z_N|I(|Z_N| \geq W_N)) = \mathbf{E}\{\mathbf{E}[|Z_N|I(|Z_N| \geq W_N)|I_1]\}. \tag{12}$$

Note that $Z_N = n_2^{-1} \sum_{I_2} [\zeta_{iN} - \mathbf{E}(\zeta_{iN}|I_1)]$ where, for fixed subsample I_1 , the random variables $\zeta_{iN} = p_N(X_i) - p(X_i)$, $X_i \in I_2$, are i.i.d., and

$$\begin{aligned}
\mathbf{E}(\zeta_{iN}|I_1) &= \int pp_N - \int p^2, \\
\mathbf{E}(\zeta_{iN}^2|I_1) &= \int (p_N(x) - p(x))^2 p(x) dx \leq p_{max} \|p_N - p\|^2,
\end{aligned}$$

by Assumption 3. To evaluate (12) we will use Bernstein's inequality (see, e.g., Serfling (1980)):

$$\mathbf{P}(|Z_N| \geq t|I_1) \leq 2\rho(t) \quad \text{for all } t > 0,$$

where

$$\rho(t) = \exp\left(-\frac{n_2 t^2}{2p_{max}\|p_N - p\|^2 + 2t\|p_N - p\|_\infty/3}\right).$$

We have

$$\begin{aligned} \mathbf{E}[|Z_N|I(|Z_N| \geq W_N)|I_1] &= W_N \mathbf{P}(|Z_N| \geq W_N|I_1) \\ &\quad + \int_{W_N}^{\infty} \mathbf{P}(|Z_N| \geq t|I_1)dt \\ &\leq A_0 + A_1, \end{aligned} \tag{13}$$

where

$$A_0 = 2W_N \rho(W_N) \quad \text{and} \quad A_1 = 2 \int_{W_N}^{\infty} \rho(t)dt.$$

We first bound from above the integral A_1 . Consider the following two sets:

$$\begin{aligned} T_1 &= \{t > 0 : t\|p_N - p\|_\infty \leq 3p_{max}\|p_N - p\|^2\}, \\ T_2 &= \{t > 0 : t\|p_N - p\|_\infty > 3p_{max}\|p_N - p\|^2\}. \end{aligned}$$

On T_1 we evaluate:

$$\rho(t) \leq \exp\left(-\frac{n_2 t^2}{4p_{max}\|p_N - p\|^2}\right) \quad \text{for all } t \in T_1, \tag{14}$$

while on T_2 :

$$\rho(t) \leq \exp\left(-\frac{3n_2 t}{4\|p_N - p\|_\infty}\right), \quad \text{for all } t \in T_2. \tag{15}$$

Consider first the set T_1 . Setting $u = t\sqrt{n_2}/(\sqrt{2p_{max}}\|p_N - p\|)$ and $W'_N = W_N\sqrt{n_2}/(\sqrt{2p_{max}}\|p_N - p\|)$, we get

$$\begin{aligned} A_{11} &\triangleq \int_{W_N}^{\infty} \exp\left(-\frac{n_2 t^2}{4p_{max}\|p_N - p\|^2}\right) I(t \in T_1)dt \\ &\leq \frac{\sqrt{2p_{max}}\|p_N - p\|}{\sqrt{n_2}} \int_{W'_N}^{\infty} e^{-u^2/2} du \\ &\leq C \frac{\sqrt{p_{max}}\|p_N - p\|}{\sqrt{n_2}} \exp(-(W'_N)^2/2) \\ &= C \frac{\sqrt{p_{max}}\|p_N - p\|}{\sqrt{n_2}} \exp\left(-\frac{n_2 W_N^2}{4p_{max}\|p_N - p\|^2}\right) \\ &\leq C \sqrt{\frac{p_{max}}{n_2}} \|p_N - p\| \exp\left(-\frac{\gamma^2 \log^2 M}{p_{max}}\right), \end{aligned}$$

where we have used $W_N \geq 2\gamma(\log M)\|p_N - p\|/\sqrt{n_2}$, and C , here and later, denotes a positive constant, not always the same.

Consider now the set T_2 . Setting $W_N'' = 3n_2W_N/(4\|p_N - p\|_\infty)$, we find

$$\begin{aligned}
A_{12} &\triangleq \int_{W_N}^{\infty} \exp\left(-\frac{3n_2t}{4\|p_N - p\|_\infty}\right) I(t \in T_2) dt \\
&\leq \frac{4\|p_N - p\|_\infty}{3n_2} \int_{W_N''}^{\infty} e^{-u} du \\
&= \frac{4\|p_N - p\|_\infty}{3n_2} \exp\left(-\frac{3n_2W_N}{4\|p_N - p\|_\infty}\right) \\
&\leq \frac{4\|p_N - p\|_\infty}{3n_2} \exp\left(-\frac{3\gamma \log^2 M}{4\|p_N - p\|_\infty}\right),
\end{aligned}$$

where we have used $W_N \geq \gamma(\log M)^2/n_2$. Therefore we have

$$\begin{aligned}
A_1 \leq 2(A_{11} + A_{12}) &\leq C \sqrt{\frac{p_{max}}{n_2}} \|p_N - p\| \exp\left(-\frac{\gamma^2 \log^2 M}{p_{max}}\right) \\
&\quad + \frac{8\|p_N - p\|_\infty}{3n_2} \exp\left(-\frac{3\gamma \log^2 M}{4\|p_N - p\|_\infty}\right). \tag{16}
\end{aligned}$$

We turn now to the evaluation of A_0 . The argument here is similar to that used above. If $W_N \in T_1$, then using (14) and the inequality $x \exp(-x^2) \leq \exp(-x^2/2)$, for all $x > 0$, we get

$$\begin{aligned}
A_0 &\leq 2W_N \exp\left(-\frac{n_2W_N^2}{4p_{max}\|p_N - p\|^2}\right) \\
&\leq 4\sqrt{\frac{p_{max}}{n_2}} \|p_N - p\| \exp\left(-\frac{n_2W_N^2}{8p_{max}\|p_N - p\|^2}\right) \\
&\leq 4\sqrt{\frac{p_{max}}{n_2}} \|p_N - p\| \exp\left(-\frac{\gamma^2 \log^2 M}{2p_{max}}\right). \tag{17}
\end{aligned}$$

Similarly, if $W_N \in T_2$, then using (15) and the inequality $x \exp(-x) \leq \exp(-x/2)$, for all $x > 0$, we find

$$\begin{aligned}
A_0 &\leq 2W_N \exp\left(-\frac{3n_2W_N}{4\|p_N - p\|_\infty}\right) \\
&\leq \frac{8\|p_N - p\|_\infty}{3n_2} \exp\left(-\frac{3n_2W_N}{8\|p_N - p\|_\infty}\right) \\
&\leq \frac{8\|p_N - p\|_\infty}{3n_2} \exp\left(-\frac{3\gamma \log^2 M}{8\|p_N - p\|_\infty}\right). \tag{18}
\end{aligned}$$

Returning now to (12) and (13) and using (16) – (18), we obtain

$$\begin{aligned} \mathbf{E}(|Z_N|I(|Z_N| \geq W_N)) &\leq \mathbf{E}(A_0) + \mathbf{E}(A_1) \\ &\leq \frac{C}{\sqrt{n_2}} \exp(-C^{-1}\gamma^2 \log^2 M) \mathbf{E}\|p_N - p\| \\ &\quad + \frac{C}{n_2} \mathbf{E} \left[\|p_N - p\|_\infty \exp\left(-\frac{3\gamma \log^2 M}{8\|p_N - p\|_\infty}\right) \right]. \end{aligned}$$

This together with (11) gives

$$\begin{aligned} 2\mathbf{E}(|Z_{\tilde{N}}|) &\leq 2\gamma \left(\mathbf{E}\|p_{\tilde{N}} - p\|^2 + \frac{\log^2 M}{n_2} \right) \\ &\quad + \frac{C}{\sqrt{n_2}} \exp(-C^{-1}\gamma^2 \log^2 M) \sum_{N=1}^M \mathbf{E}\|p_N - p\| \\ &\quad + \frac{C}{n_2} \sum_{N=1}^M \mathbf{E} \left[\|p_N - p\|_\infty \exp\left(-\frac{3\gamma \log^2 M}{8\|p_N - p\|_\infty}\right) \right] \\ &\triangleq 2\gamma \mathbf{E}\|p_{\tilde{N}} - p\|^2 + R. \end{aligned} \tag{19}$$

From (19) and (10) we get

$$(1 - 2\gamma) \mathbf{E}\|p_{\tilde{N}} - p\|^2 \leq \mathbf{E}\|p_N - p\|^2 + R,$$

and, with $0 < \gamma < 1/4$,

$$\mathbf{E}\|p_{\tilde{N}} - p\|^2 \leq (1 + 4\gamma) \mathbf{E}\|p_N - p\|^2 + (1 + 4\gamma)R.$$

Set now $\gamma = (8\gamma_0/3)(\log M)^{-1/4}$ where $\gamma_0 \leq 1/12$ is the constant in Assumption 2. Then $0 < \gamma \leq 2(\log 2)^{-1/4}/9 < 1/4$ for all $M \geq 2$, and we have the following bound on the remainder term R defined in (19):

$$\begin{aligned} R &\leq C \left\{ \frac{\log^{7/4} M}{n_2} + \frac{1}{\sqrt{n_2}} \exp(-C^{-1} \log^{3/2} M) \sum_{N=1}^M \mathbf{E}\|p_N - p\| \right. \\ &\quad \left. + \frac{1}{n_2} \sum_{N=1}^M \mathbf{E} \left[\|p_N - p\|_\infty \exp\left(-\frac{\gamma_0 \log^{7/4} M}{\|p_N - p\|_\infty}\right) \right] \right\}. \end{aligned}$$

The theorem follows from the last two displays by applying Assumptions 1 and 2.

Remark 2. Inspection of the proof shows that Assumption 2 can be slightly generalized and the remainder term $(\log M)^3/n$ in (9) can be reduced to $(\log M)^{1+\varepsilon}/n$

for an arbitrarily small $\varepsilon > 0$. To obtain this, it suffices to fix an arbitrarily small $\nu > 0$, to replace $\log^2 M$ by $(\log M)^{1+\nu}$ in the definition of r , and to take $\gamma \asymp (\log M)^{-\nu'}$ with $\nu' < \nu/2$, $n_2 = \lfloor cn/(\log M)^\nu \rfloor$. Then $\log^{7/4} M$ and $\log^2 M$ in (7) can be replaced by $(\log M)^{1+\nu-\nu'}$ and $(\log M)^{1+2\nu-\nu'}$, respectively. We did not include these extensions in Theorem 1, because they require more notation but seem not to be crucial for application of the result.

3 Application to a dimensionality reduction model

Let X_1, \dots, X_n be i.i.d. random vectors with common probability density p on \mathbf{R}^d , $d \geq 2$. We consider the problem of nonparametric estimation of the density p assuming that it has the form

$$p(x) \equiv f_B(x) \triangleq \phi_d(x)g(B^T x), \quad x \in \mathbf{R}^d, \quad (20)$$

where B is an unknown $d \times m$ matrix with orthonormal columns, $1 \leq m \leq d$, the function $g : \mathbf{R}^m \rightarrow [0, \infty)$ is unknown, and $\phi_d(\cdot)$ is the density of the standard d -variate normal distribution. Our goal is to show, using Theorem 1, that one can estimate the density (20), without knowing B and m , with the same rate as the optimal rate attainable when B and m are known.

Note that the representation (20) is not unique. In particular, if Q_m is an $m \times m$ orthogonal matrix, the density p in (20) can be rewritten as $p(x) = \phi_d(x)g_1(B_1^T x)$ with $g_1(y) = g(Q_m y)$ and $B_1 = BQ_m$. However, the linear subspace \mathcal{M} spanned by the columns of B is uniquely defined by (20). By analogy with regression models, e.g. Li (1991), Hristache, et al. (2001), we will call \mathcal{M} the *index space*. In particular, if the dimension of \mathcal{M} is 1, (20) can be viewed as a density analog of the single index model in regression. In general, if the dimension of \mathcal{M} is arbitrary, we call (20) the *multiple index model*. The directions where the density of projections of X_i is standard normal are interpreted as non-interesting (“pure noise” directions).

The model (20) can be viewed as a modification of the projection pursuit density estimation (PPDE) model, e.g. Huber (1985). A common PPDE model corresponds to the special case of (20) where the function g can be represented as a product of densities corresponding to one-dimensional projections. In this case, the density can be estimated with one-dimensional rate (Samarov and Tsybakov (2004)), and

thus the dimension reduction principle is realized. Models similar to (20) also arise in biased, or weighted, sampling, where a direct sampling from a density f is, for some reason, impossible, and an observation $X = x$ from f may be available with a relative probability proportional to a so-called biasing function $w(x)$. The biased observations have the density $p(x) = f(x)w(x) / \int w(x)f(x)dx$, and a typical problem in biased estimation is: having observations from p , estimate f , when $w(\cdot)$ is known, e.g. Cox (1969), Patil and Rao (1977). In our setting, $f = \phi_d$ is known while the biasing function has the form $g(B^T x)$ and is unknown, and our goal is to estimate $p(\cdot)$.

When the dimension m and an index matrix B (i.e. any of the matrices, equivalent up to an orthogonal transformation, that define the index space \mathcal{M}) are specified, the density (20) can be estimated using a kernel estimator

$$\hat{p}_{m,B}(x) = \frac{\phi_d(x)}{\phi_m(B^T x)} \frac{1}{nh^m} \sum_{i=1}^n K\left(\frac{B^T(X_i - x)}{h}\right), \quad (21)$$

with appropriately chosen bandwidth $h > 0$ and kernel $K : \mathbf{R}^m \rightarrow \mathbf{R}^1$. We will assume the following.

Assumption 4. *The function $g : \mathbf{R}^m \rightarrow [0, \infty)$ in (20) is bounded on \mathbf{R}^m with its gradient ∇g and Hessian $\nabla^2 g$, so that $\max\{g(z), |\nabla g(z)|_m, \|\nabla^2 g(z)\|_2\} \leq L_g$, for all $z \in \mathbf{R}^m$, where L_g is a constant, $|\cdot|_m$ denotes the Euclidean norm in \mathbf{R}^m and $\|A\|_2 = \text{Tr}^{1/2}(AA^T)$ denotes the Frobenius norm of the matrix A .*

Assumption 5. *The kernel $K : \mathbf{R}^m \rightarrow \mathbf{R}^1$ is a bounded function supported on $[-1, 1]^m$ and such that $\int_{\mathbf{R}^m} K(t)dt = 1$ and $\int_{\mathbf{R}^m} K(t)t_j dt = 0$, $j = 1, \dots, m$, where t_j is the j th component of $t \in \mathbf{R}^m$.*

Kernels satisfying Assumption 5 can be easily constructed as products of m one-dimensional kernels.

We first suppose that the dimension m and an index matrix B are known and establish the rate of convergence of the estimator (21).

Proposition 1 *Let the density p be of the form (20) with g satisfying Assumption 4. Then, for the estimator (21) with kernel K satisfying Assumption 5, we have the*

following bounds on the L_2 -bias and variance terms

$$\|\mathbf{E}(\hat{p}_{m,B}) - p\|^2 \leq C_4 h^4, \quad (22)$$

$$\mathbf{E}(\|\mathbf{E}(\hat{p}_{m,B}) - \hat{p}_{m,B}\|^2) \leq \frac{C_5}{nh^m}. \quad (23)$$

Here $0 < h \leq h_0$ with some $h_0 < \infty$ and any integer $n \geq 1$ and C_4 and C_5 are constants depending only on d, L_g, h_0 and on $K_{\max} \triangleq \sup_{z \in \mathbf{R}^m} |K(z)|$.

Proof. For every $x \in \mathbf{R}^d$, the expectation of $\hat{p}_{m,B}(x)$ can be written as follows:

$$\begin{aligned} \mathbf{E}(\hat{p}_{m,B}(x)) &= \frac{\phi_d(x)}{h^m \phi_m(B^T x)} \int_{\mathbf{R}^d} K\left(\frac{B^T(y-x)}{h}\right) \phi_d(y) g(B^T y) dy \\ &= \frac{\phi_d(x)}{h^m \phi_m(B^T x)} \int_{\mathbf{R}^{d-m}} \left[\int_{\mathbf{R}^m} K\left(\frac{u - B^T x}{h}\right) \phi_m(u) g(u) du \right] \phi_{d-m}(v) dv \end{aligned} \quad (24)$$

with new variables $u = B^T y$ and $v = \tilde{B}^T y$, where \tilde{B} is a $d \times (d-m)$ matrix with orthonormal columns such that $(B|\tilde{B})$ is a $d \times d$ orthogonal matrix. Making in (24) the change of variables $t = (u - B^T x)/h$, we find that the bias of $\hat{p}_{m,B}(x)$ equals

$$\begin{aligned} \mathbf{E}(\hat{p}_{m,B}(x)) - p(x) &= \frac{\phi_d(x)}{\phi_m(B^T x)} \int_{\mathbf{R}^m} K(t) \phi_m(B^T x + th) g(B^T x + th) dt \\ &\quad - \phi_d(x) g(B^T x), \end{aligned} \quad (25)$$

and, under the above assumptions about g and K , the standard Taylor expansion argument gives

$$\begin{aligned} \|\mathbf{E}(\hat{p}_{m,B}) - p\|^2 &= \int_{\mathbf{R}^d} (\mathbf{E}(\hat{p}_{m,B}(x)) - p(x))^2 dx = \\ &= \int_{\mathbf{R}^d} \left(\frac{\phi_d(x)}{\phi_m(B^T x)} \int_{\mathbf{R}^m} K(t) \frac{h^2}{2} t^T D(B^T x + a^* t) t dt \right)^2 dx, \end{aligned} \quad (26)$$

where $0 \leq a^* \leq h$ and $D(z) = \nabla^2(\phi_m(z)g(z)) = [(zz^T - \mathbf{I}_m)g(z) - \nabla g(z)z^T - z\nabla^T g(z) + \nabla^2 g(z)]\phi_m(z)$. Here and in what follows \mathbf{I}_m stands for the identity matrix of dimension m . Using Assumption 4 and the fact that $a^* \leq h_0$, we get

$$\begin{aligned} t^T D(B^T x + a^* t) t &\leq CL_g |t|_m^2 (1 + h_0^2 |t|_m^2 + |B^T x|_m^2) \phi_m(B^T x + a^* t) \\ &\leq CL_g |t|_m^2 (1 + h_0^2 |t|_m^2 + |B^T x|_m^2) \exp(|B^T x|_m h_0 |t|_m) \phi_m(B^T x), \end{aligned}$$

with some constant $C > 0$. Because $K(t)$ has bounded support, (22) follows from (26). For the variance term, we have

$$\begin{aligned} \text{Var}(\hat{p}_{m,B}(x)) &= \frac{\phi_d^2(x)}{nh^{2m}\phi_m^2(B^T x)} \text{Var} \left(K \left(\frac{B^T(X-x)}{h} \right) \right) \\ &\leq \frac{1}{(2\pi)^{d-m}nh^{2m}} \exp(-x^T(\mathbf{I}_d - BB^T)x) \int_{\mathbf{R}^d} K^2 \left(\frac{B^T(y-x)}{h} \right) \phi_d(y)g(B^T y)dy \\ &\leq \frac{L_g}{(2\pi)^{d-m}nh^{2m}} \int_{\mathbf{R}^d} K^2 \left(\frac{B^T(y-x)}{h} \right) \phi_d(y)dy, \end{aligned}$$

and after making the same changes of variables as for the bias, we obtain

$$\mathbf{E} (\|\mathbf{E}(\hat{p}_{m,B}) - \hat{p}_{m,B}\|^2) = \int_{\mathbf{R}^d} \text{Var}(\hat{p}_{m,B}(x))dx = O(n^{-1}h^{-m}).$$

■

Consider the mean integrated mean squared error (MISE) of the estimator $\hat{p}_{m,B}$:

$$MISE(\hat{p}_{m,B}, p) \triangleq \mathbf{E}\|\hat{p}_{m,B} - p\|^2 \equiv \mathbf{E}\|\hat{p}_{m,B} - f_B\|^2. \quad (27)$$

Proposition 1 implies that, under Assumptions 4 and 5,

$$MISE(\hat{p}_{m,B}, p) = O(n^{-4/(m+4)}), \quad (28)$$

if the bandwidth h is chosen of the order $h \asymp n^{-1/(m+4)}$. Using the standard techniques of the minimax lower bounds (e.g. Tsybakov (2004)), it is easy to show that the rate $n^{-4/(m+4)}$ given in (28) is the optimal MISE rate for the model (20) on the class of densities p defined by Assumption 4, and thus the estimator $\hat{p}_{m,B}$ with $h \asymp n^{-1/(m+4)}$ has the optimal rate for this class of densities.

Consider now the case where the dimension m and the index matrix B are unknown. We will use the procedure of Section 2 to aggregate estimators of the type (20) corresponding to candidate pairs $(m, B) = (k, A)$ with $k = 1, \dots, d$ and with A that runs over a finite net on the set of all admissible $d \times k$ index matrices. The latter is the set \mathcal{B}_k of all $d \times k$ matrices A with orthonormal columns. This set is bounded in the Frobenius norm $\|A\|_2 = \text{Tr}^{1/2}(AA^T)$. Consider an ϵ -net Q_k on \mathcal{B}_k constructed using the Frobenius norm. Note that orthogonal transformations preserve the norm, so that both estimators (21) and the ϵ -net Q_k are invariant under orthogonal transformations, and thus are not affected by the non-uniqueness of representation (20).

The set \mathcal{B}_k is bounded and can be imbedded in \mathbf{R}^s with $s = k(d - (k + 1)/2)$, and therefore we can construct an ϵ -net Q_k with cardinality

$$\text{Card}(Q_k) = O(\epsilon^{-k(d-(k+1)/2)}), \quad (29)$$

e.g. Wellner and van der Vaart (1996). Doing this for $k = 1, \dots, d$, we obtain a collection Q_1, \dots, Q_k of ϵ -nets with the property (29) each, and in what follows we set $\epsilon = n^{-a}$ with $a > 2/5$ for all $k = 1, \dots, d$.

We can now define the aggregate. As in Section 2, we split the sample X_1, \dots, X_n into two parts, I_1 and I_2 with $n_1 = \text{Card}(I_1)$, $n_2 = \text{Card}(I_2)$, $n = n_1 + n_2$. From the first subsample we construct estimators

$$\hat{p}_{k,A}(x) = \frac{\phi_d(x)}{\phi_k(A^T x)} \frac{1}{n_1 h_k^k} \sum_{I_1} K\left(\frac{A^T(X_i - x)}{h_k}\right), \quad k = 1, \dots, d, \quad A \in Q_k, \quad (30)$$

where $h_k \asymp n^{-1/(k+4)}$. These estimators are of the form (21), but here we plug in k and A that are not necessarily equal to the true unknown values m and B and we use only the first subsample I_1 . Nevertheless, we preserve the same notation as in (21) since this will not cause ambiguity.

Let now $p_{\tilde{N}}$ be the aggregate defined as in (2) and (3) using as $\{p_1, \dots, p_M\}$ the collection of estimators $\{\hat{p}_{k,A}, k = 1, \dots, d, A \in Q_k\}$ of the form (30) with bandwidths $h_k \asymp n^{-\frac{1}{k+4}}$ and ϵ -nets Q_k such that $\epsilon = n^{-a}$, $a > 2/5$. In view of (29), the cardinality M of this set of estimators is

$$M \asymp \sum_{k=1}^d n^{ak(d-(k+1)/2)} \asymp n^{ad(d-1)/2}. \quad (31)$$

In this case, the aggregate $p_{\tilde{N}}$ of (2) and (3) can be written in the form $\hat{p}_{\tilde{k}, \tilde{A}}$ where (\tilde{k}, \tilde{A}) are given by

$$(\tilde{k}, \tilde{A}) = \arg \min_{k=1, \dots, d, A \in Q_k} \left(-\frac{2}{n_2} \sum_{I_2} \hat{p}_{k,A}(X_i) + \int \hat{p}_{k,A}^2 \right). \quad (32)$$

We can now state the main result of this section.

Theorem 2 *Let Assumptions 4 and 5 hold and let $n_2 = \lfloor \frac{cn}{\log n} \rfloor$ for some constant $c > 0$ such that $1 \leq n_2 < n$. Assume in addition that the kernel $K(\cdot)$ is Lipschitz continuous. Then for the aggregate $\hat{p}_{\tilde{k}, \tilde{A}}$ we have*

$$\mathbf{E} \|\hat{p}_{\tilde{k}, \tilde{A}} - p\|^2 = O(n^{-4/(m+4)}), \quad (33)$$

as $n \rightarrow \infty$, so that $\hat{p}_{\tilde{k}, \tilde{A}}$ estimates p with the best rate attainable when dimension m and matrix B are known.

Proof. We first verify the assumptions of Theorem 1. Clearly, Assumption 4 implies Assumption 3. With a bounded kernel K , $\|\hat{p}_{k,A} - p\| \leq Ch_k^{-k} = O(n^{k/(k+4)})$, so that Assumption 1 holds with $a_1 = d/(d+4) + ad(d-1)/2$ and $a_2 = ad(d-1)/2$.

In order to verify Assumption 2, we will show that (8) holds for estimators $p_N = \hat{p}_{k,A}$ with $k = 1, \dots, d$ and $A \in Q_k$.

Proof of (8). For any estimator $\hat{p}_{k,A}$ we have $\|\hat{p}_{k,A} - p\|_\infty \leq \|\hat{p}_{k,A} - \mathbf{E}(\hat{p}_{k,A})\|_\infty + \|\mathbf{E}(\hat{p}_{k,A}) - p\|_\infty$, so that (25), written with $\hat{p}_{k,A}$ instead of $\hat{p}_{m,B}$, implies that $\|\hat{p}_{k,A} - p\|_\infty \leq \|\hat{p}_{k,A} - \mathbf{E}(\hat{p}_{k,A})\|_\infty + C$ for some constant $C > 0$ which depends on g but not on k and A . Therefore we have

$$\begin{aligned} & \mathbf{E}[\|\hat{p}_{k,A} - p\|_\infty I(\|\hat{p}_{k,A} - p\|_\infty > \gamma_0 \log^{3/4} M)] \\ & \leq \mathbf{E}[(\|\hat{p}_{k,A} - \mathbf{E}(\hat{p}_{k,A})\|_\infty + C) I(\|\hat{p}_{k,A} - \mathbf{E}(\hat{p}_{k,A})\|_\infty > \gamma_0 \log^{3/4} M - C)]. \end{aligned} \quad (34)$$

Note that, for any $x \in \mathbf{R}^d$,

$$\hat{p}_{k,A}(x) - \mathbf{E}(\hat{p}_{k,A}(x)) = (2\pi)^{-(d-k)/2} \exp(-x^T(\mathbf{I}_d - AA^T)x/2) \sum_{I_1} \zeta_{i,n}(z), \quad (35)$$

where $z = A^T x$ and

$$\zeta_{i,n}(z) = \zeta'_{i,n}(z) - \mathbf{E}(\zeta'_{i,n}(z)), \quad \zeta'_{i,n}(z) = \frac{1}{n_1 h_k^k} K\left(\frac{A^T X_i - z}{h_k}\right).$$

Introduce the truncated variables

$$\xi_{i,n}(z) = \xi'_{i,n}(z) - \mathbf{E}(\xi'_{i,n}(z)), \quad \xi'_{i,n}(z) = \frac{1}{n_1 h_k^k} K\left(\frac{A^T X_i - z}{h_k}\right) I(|X_i|_d \leq \log n),$$

and note that

$$\mathbf{P}(|X_1|_d > \log n) \leq C(\log n)^d \exp\left(-\frac{\log^2 n}{2}\right), \quad (36)$$

where the constant C depends only on g_{\max} and d . This follows from the relations

$$\mathbf{P}(|X_1|_d > \log n) = \int_{\mathbf{R}^d} I(|x|_d > \log n) \phi_d(x) g(B^T x) dx \leq g_{\max} \int_{|x|_d > \log n} \phi_d(x) dx,$$

followed by evaluation of the tail of d -dimensional standard normal distribution. Consider the random event $\mathcal{A} = \{|X_i|_d \leq \log n, i = 1, \dots, n\}$. In view of (36), the probability of the complementary event satisfies

$$\mathbf{P}(\mathcal{A}^c) \leq Cn(\log n)^d \exp\left(-\frac{\log^2 n}{2}\right). \quad (37)$$

Using (35) and the fact that $\mathbf{I}_d - AA^T \geq 0$ for all matrices $A \in \mathcal{B}_k$, we get

$$\|\hat{p}_{k,A} - \mathbf{E}(\hat{p}_{k,A})\|_\infty \leq (2\pi)^{-(d-k)/2} \sup_{z \in E^k} \left| \sum_{I_1} \zeta_{i,n}(z) \right|,$$

where E^k is the linear subspace of \mathbf{R}^d spanned by the columns of A . Now, (36) implies that for any $D > 0$ there exists a constant C depending only on g_{\max} , K_{\max} and d , such that $\mathbf{E}|\zeta'_{i,n}(z) - \xi'_{i,n}(z)| \leq Cn^{-D}$. Therefore,

$$\|\hat{p}_{k,A} - \mathbf{E}(\hat{p}_{k,A})\|_\infty \leq (2\pi)^{-(d-k)/2} \sup_{z \in E^k} \left| \sum_{I_1} [\zeta'_{i,n}(z) - \mathbf{E}(\xi'_{i,n}(z))] \right| + Cn^{-D}. \quad (38)$$

Setting

$$\eta \triangleq \sup_{z \in E^k} \left| \sum_{I_1} [\zeta'_{i,n}(z) - \mathbf{E}(\xi'_{i,n}(z))] \right|,$$

we note that, in view of the inequalities (31), (34) and (38), to prove (8) it is enough to show that

$$\mathbf{P}(\eta > C \log^{3/4} n) + \mathbf{E}[\eta I(\eta > C \log^{3/4} n)] \leq \frac{\log^2 M}{M}. \quad (39)$$

We will in fact prove a stronger result, namely, that the left-hand side of (39) decreases faster than any power of n . Since on the event \mathcal{A} it holds that $\zeta'_{i,n}(z) = \xi'_{i,n}(z)$ for all $z \in \mathbf{R}^k$, we obtain

$$\begin{aligned} \mathbf{P}(\eta > s) &\leq \mathbf{P}(\mathcal{A}^c) + \mathbf{P}\left(\sup_{z \in E^k} \left| \sum_{I_1} \xi_{i,n}(z) \right| > s\right) \\ &= \mathbf{P}(\mathcal{A}^c) + \mathbf{P}\left(\sup_{z \in S \cap E^k} \left| \sum_{I_1} \xi_{i,n}(z) \right| > s\right), \end{aligned} \quad (40)$$

where the last equality is due to the fact that $\xi'_{i,n}(z) = 0$ for all $z \notin S$, with $S = \{x \in \mathbf{R}^d : |x|_d \leq 1 + \log n\}$.

As kernel K is Lipschitz continuous, we get

$$\left| \sum_{I_1} (\xi_{i,n}(z) - \xi_{i,n}(y)) \right| \leq C_L h_k^{-(k+1)} |z - y|_d, \quad \forall z, y \in E^k, \quad (41)$$

where C_L is a constant. Next, fix some $\delta > 0$, and let z_1, \dots, z_L be a δ -net in Euclidean metric on the bounded set $S \cap E^k$ such that $L \leq C(\frac{\log n}{\delta})^d$. Clearly, a δ -net of cardinality L satisfying the latter inequality exists, since the cardinality of the minimal δ -net on the larger set S is of the order $(\frac{\log n}{\delta})^d$. In view of (41), we have, for $s > 2C_L \delta h_k^{-(k+1)}$,

$$\begin{aligned} \mathbf{P} \left(\sup_{z \in S \cap E^k} \left| \sum_{I_1} \xi_{i,n}(z) \right| > s \right) &\leq \mathbf{P} \left(\max_{1 \leq j \leq L} \left| \sum_{I_1} \xi_{i,n}(z_j) \right| > s/2 \right) \\ &\leq L \sup_{z \in S \cap E^k} \mathbf{P} \left(\left| \sum_{I_1} \xi_{i,n}(z) \right| > s/2 \right). \end{aligned} \quad (42)$$

We have $\mathbf{E}(\xi_{i,n}(z)) = 0$ and $\sup_{z \in S \cap E^k} |\xi_{i,n}(z)| \leq c_1 n_1^{-1} h_k^{-k}$, for some constant $c_1 > 0$. Also, using (20) and Assumption 4, we find

$$\begin{aligned} \text{Var}(\xi_{i,n}(z)) &\leq \mathbf{E} \zeta'_{i,n}(z)^2 = \frac{1}{n_1^2 h_k^{2k}} \int_{\mathbf{R}^d} K^2 \left(\frac{A^T y - z}{h_k} \right) f_B(y) dy \\ &\leq \frac{L_g}{n_1^2 h_k^{2k}} \int_{\mathbf{R}^d} K^2 \left(\frac{A^T y - z}{h_k} \right) \phi_d(y) dy \\ &= \frac{L_g}{n_1^2 h_k^{2k}} \int_{\mathbf{R}^k} K^2(t) \phi_k(th_k + z) \left[\int_{\mathbf{R}^{d-k}} \phi_{d-k}(u) du \right] dt \end{aligned}$$

with new variables $t = (A^T y - z)/h_k$ and $u = \tilde{A}^T y$, where \tilde{A} is a $d \times (d - k)$ matrix with orthonormal columns such that $(A|\tilde{A})$ is a $d \times d$ orthogonal matrix. Therefore we have $\sup_{z \in S \cap E^k} \text{Var}(\xi_{i,n}(z)) \leq c_2 n_1^{-2} h_k^{-k}$, for some constant $c_2 > 0$.

Choosing now $\delta = h_k^{k+1}$, applying in (42) the Bernstein inequality and recalling

that $h_k \asymp n^{-\frac{1}{k+4}}$, $n_1 = n - n_2 = n(1 + o(1))$ we get, for $s > 2C_L$,

$$\begin{aligned}
\mathbf{P} \left(\sup_{z \in S \cap E^k} \left| \sum_{I_1} \xi_{i,n}(z) \right| > s \right) &\leq 2L \exp \left(-\frac{(s/2)^2}{2c_2 n_1^{-1} h_k^{-k} + c_1 n_1^{-1} h_k^{-k} s/3} \right) \\
&\leq C \left(\frac{\log n}{\delta} \right)^d \exp \left(-\frac{s^2 n_1 h_k^k}{8c_2 + 2c_1 s} \right) \\
&\leq C (\log n)^d n^{\frac{d(k+1)}{k+4}} \exp \left(-\frac{s^2 n^{4/(k+4)} (1 + o(1))}{8c_2 + 2c_1 s} \right) \\
&\leq C (\log n)^d n^{\frac{d(d+1)}{d+4}} \exp \left(-\frac{s^2 n^{4/(d+4)}}{C(1+s)} \right), \tag{43}
\end{aligned}$$

where the last inequality is valid for n large enough. From (40), (37) and (43) we deduce that, for n large enough,

$$\begin{aligned}
\mathbf{P}(\eta > C \log^{3/4} n) &\leq C (\log n)^d \left[n \exp \left(-\frac{\log^2 n}{2} \right) \right. \\
&\quad \left. + n^{\frac{d(d+1)}{d+4}} \exp \left(-\frac{n^{4/(d+4)} \log^{3/4} n}{C} \right) \right]. \tag{44}
\end{aligned}$$

On the other hand, $\eta \leq 2K_{\max} h_k^{-k} = O(n^{k/(k+4)}) = O(n^{d/(d+4)})$, and therefore $\mathbf{E}[\eta I(\eta > C \log^{3/4} n)] \leq O(n^{d/(d+4)}) \mathbf{P}(\eta > C \log^{3/4} n)$. This inequality and (44) combined with (31) prove that (39) holds for n large enough. The proof of (8) is thus complete.

All the assumptions of Theorem 1 are therefore satisfied. Applying Theorem 1 we get the oracle inequality

$$\mathbf{E} \|\hat{p}_{\tilde{k}, \tilde{A}} - p\|^2 \leq \left(1 + \frac{C^*}{\log^{1/4} n} \right) \min_{k=1, \dots, d} \min_{A \in Q_k} MISE(\hat{p}_{k,A}, p) + C^* \frac{\log^3 n}{n}. \tag{45}$$

To complete the proof of Theorem 2, we now show that

$$\min_{k=1, \dots, d} \min_{A \in Q_k} MISE(\hat{p}_{k,A}, p) = O(n^{-4/(m+4)}). \tag{46}$$

In fact,

$$\min_{k=1, \dots, d} \min_{A \in Q_k} MISE(\hat{p}_{k,A}, p) \leq MISE(\hat{p}_{m, B^*}, p), \tag{47}$$

where B^* is a matrix in Q_m closest to B in the Frobenius norm, and thus satisfying $\|B^* - B\|_2 \leq \epsilon$. We have (recall that $p \equiv f_B$)

$$\|\hat{p}_{m, B^*} - p\|^2 \leq 2(\|\hat{p}_{m, B^*} - f_{B^*}\|^2 + \|f_{B^*} - p\|^2) = 2(\|\hat{p}_{m, B^*} - f_{B^*}\|^2 + \|f_{B^*} - f_B\|^2). \tag{48}$$

It follows from (27) and (28) that

$$\mathbf{E}\|\hat{p}_{m,B^*} - f_{B^*}\|^2 = O(n^{-4/(m+4)}). \quad (49)$$

(Note that we proved (28) for the estimator (21), while here the estimator \hat{p}_{m,B^*} is defined by (30) and based on the sample of size n_1 ; nevertheless the result remains valid, since $n_1 = n(1 + o(1))$.) Using (49) and applying Assumption 4 to bound from above the last summand in (48), we obtain

$$MISE(\hat{p}_{m,B^*}, p) \leq b_1 n^{-4/(m+4)} + b_2 \epsilon^2 \quad (50)$$

with some constants b_1, b_2 . Since $\epsilon = n^{-a}$ with $a > 2/5 \geq 2/(m+4)$ we get $MISE(\hat{p}_{m,B^*}, p) = O(n^{-4/(m+4)})$. Together with (47) this implies (46). ■

Remark 3. The aggregate estimator for model (20) suggested here automatically accomplishes dimension reduction. In fact, if the unknown true dimension m is small, it achieves the rate $O(n^{-4/(m+4)})$ that can be much faster than the best attainable rate $O(n^{-4/(d+4)})$ for a model of full dimension. The aggregate can be interpreted as an adaptive estimator, but in contrast to adaptation to unknown smoothness usually considered in nonparametrics, here we deal with adaptation to unknown dimension m and to the index space \mathcal{M} determined by a matrix B . The procedure provides explicit estimates (\tilde{k}, \tilde{A}) of (m, B) that are optimal in the sense of Theorem 2. The tools of this paper do not allow us, however, to evaluate how close is (\tilde{k}, \tilde{A}) to (m, B) (or, equivalently, how close is the estimated index space to the true one \mathcal{M}).

References

- [1] Birgé, L. (2003). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. Preprint n.862, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7 (available at <http://www.proba.jussieu.fr/mathdoc/preprints>).
- [2] Bunea, F., Tsybakov, A. and Wegkamp, M. (2004). Aggregation for regression learning. Preprint n.948, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7 (available at arXiv:math.ST/0410214, 8 Oct. 2004.)

- [3] Catoni, O. (2004). *Statistical Learning Theory and Stochastic Optimization. Ecole d'Eté de Probabilités de Saint-Flour XXXI - 2001*. Lecture Notes in Mathematics, vol.1851, Springer, New York.
- [4] Cox, D.R. (1969). Some Sampling Problems in Technology. In: Johnson, N. and Smith, H. (eds.), *New Developments in Survey Sampling*, Wiley-Interscience, New York, pp.506-529.
- [5] Devroye, L. and Lugosi, G. (2000). *Combinatorial Methods in Density Estimation*. Springer, New-York.
- [6] Hristache, M., Juditsky, A., Polzehl J., and Spokoiny, V. (2001). Structure Adaptive Approach for Dimension Reduction. *Ann. Statist.*, **29**, 1537-1566.
- [7] Huber, P. (1985). Projection Pursuit. *Ann. Statist.*, **13**, 435-475.
- [8] Juditsky, A.B., Nazin, A.V., Tsybakov, A.B., and Vayatis, N. (2005a) Generalization error bounds for aggregation by mirror descent. *Proceedings of NIPS-2005* (to appear).
- [9] Juditsky, A.B., Nazin, A.V., Tsybakov, A.B., and Vayatis, N. (2005b) Recursive aggregation of estimators by a mirror descent method with averaging. *Problems of Information Transmission*, **41**, n.4 (to appear).
- [10] Li, K-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, **86**, 316-342.
- [11] Nemirovski, A. (2000). Topics in Non-parametric Statistics. In: *Ecole d'Eté de Probabilités de Saint-Flour XXVIII - 1998*, Lecture Notes in Mathematics, vol. 1738, Springer, New York.
- [12] Patil, G. and Rao, C.R. (1977). The Weighted Distributions: the Survey of Their Applications. In: P.R. Krishnaiah (ed.), *Applications of Statistics*, Amsterdam, North Holland, pp. 383-405.
- [13] Rigollet, Ph. and Tsybakov, A.B. (2004) Linear and convex aggregation of density estimators. Submitted.

- [14] Samarov, A. and Tsybakov, A. (2004) Nonparametric Independent Component Analysis. *Bernoulli*, **10**, 565-582.
- [15] Serfling, R. (1980) *Approximation Theorems of Mathematical Statistics*, J. Wiley, New York.
- [16] Tsybakov, A. (2003). Optimal rates of aggregation. In: *Computational Learning Theory and Kernel Machines*, (B.Schölkopf and M.Warmuth, eds.), Lecture Notes in Artificial Intelligence, v.2777. Springer, Heidelberg, 303-313.
- [17] Tsybakov, A. (2004). *Introduction à l'estimation non-paramétrique*. Springer, Berlin-Heidelberg.
- [18] Wegkamp, M.H. (1999). Quasi-universal bandwidth selection for kernel density estimators. *Canad. J. Statist.*, **27**, 409-420.
- [19] Wegkamp, M.H. (2003). Model selection in nonparametric regression. *Ann. Statist.*, **31**, 252 – 273.
- [20] Wellner, J. and van der Vaart, A. (1996). *Weak convergence and empirical processes*. Springer, New York.
- [21] Yang, Y. (2000). Mixing strategies for density estimation. *Ann. Statist.*, **28**, 75-87.