Electronic Journal of Statistics
Vol. 4 (2010) 707–736
ISSN: 1935-7524
DOI: 10.1214/09-EJS498

Noisy independent factor analysis model for density estimation and classification

Umberto Amato[‖]

Istituto per le Applicazioni del Calcolo 'Mauro Picone' CNR – Sede di Napoli, Italy
e-mail: u.amato@iac.cnr.it
url: www.na.iac.cnr.it/amato

Anestis Antoniadis[†¶]

Laboratoire Jean Kuntzmann, Université Joseph Fourier, Grenoble 38041, France
e-mail: anestis.antoniadis@imag.fr
url: http://ljk.imag.fr/membres/Anestis.Antoniadis/

Alexander Samarov[*‡¶]

Department of Mathematical Sciences, University of Massachusetts Lowell and Sloan School of Management, MIT, MA 02139, USA
e-mail: samarov@mit.edu
url: http://www.uml.edu/college/arts_sciences/Math/faculty/samarov.html

Alexandre B. Tsybakov[§¶]

Laboratoire de Statistique, CREST, Malakoff, 92240 France
e-mail: alexandre.tsybakov@ensae.fr
url: http://www.proba.jussieu.fr/pageperso/tsybakov/tsybakov.html

Abstract: We consider the problem of multivariate density estimation when the unknown density is assumed to follow a particular form of dimensionality reduction, a noisy independent factor analysis (IFA) model. In this model the data are generated by a number of latent independent components having unknown distributions and are observed in Gaussian noise. We do not assume that either the number of components or the matrix mixing the components are known. We show that the densities of this form can be estimated with a fast rate. Using the mirror averaging aggregation algorithm, we construct a density estimator which achieves a nearly parametric rate $(\log^{1/4} n)/\sqrt{n}$, independent of the dimensionality of the data, as the sample size n tends to infinity. This estimator is adaptive

[*] Corresponding author.

[†] Financial support from the IAP research network of the Belgian government (Belgian Federal Science Policy) is gratefully acknowledged.

[‡] Research partially supported by NSF grant DMS-0505561 and by a grant from Singapore-MIT Alliance (CSB).

[§] Research partially supported by the grant ANR-06-BLAN-0194 and by the PASCAL Network of Excellence.

[¶] A. Antoniadis, A. Samarov, and A.B. Tsybakov would like to thank U. Amato for his hospitality while visiting the Istituto per le Applicazioni del Calcolo 'M. Picone' CNR, Napoli to carry out this work.

[‖] U. Amato would like to thank A. Antoniadis for his excellent hospitality while visiting the Laboratoire J. Kuntzmann, Grenoble.

to the number of components, their distributions and the mixing matrix. We then apply this density estimator to construct nonparametric plug-in classifiers and show that they achieve the best obtainable rate of the excess Bayes risk, to within a logarithmic factor independent of the dimension of the data. Applications of this classifier to simulated data sets and to real data from a remote sensing experiment show promising results.

AMS 2000 subject classifications: Primary 62H25; secondary 62G07, 62H30.
Keywords and phrases: Nonparametric density estimation, independent factor analysis, aggregation, plug-in classifier, remote sensing.

Received October 2009.

Contents

1 Introduction 708
2 Independent factor analysis model for density estimation 710
3 Application to nonparametric classification 716
4 The algorithm 718
5 Simulations and examples 719
5.1 Density estimation 719
5.2 Classification: A real data example 723
6 Conclusions 727
A Proofs 727
References 734

1. Introduction

Complex data sets lying in multidimensional spaces are a commonplace occurrence in many areas of science and engineering. There are various sources of this kind of data, including biology (genetic networks, gene expression microarrays, molecular imaging data), communications (internet data, cell phone networks), risk management, and many others. One of the important challenges of the analysis of such data is to reduce its dimensionality in order to identify and visualize its structure.

It is well known that common nonparametric density estimators are quite unreliable even for moderately high-dimensional data. This motivates the use of dimensionality reduction models. The literature on dimensionality reduction is very extensive, and we mention here only some recent publications that are connected to our context and contain further references [9, 10, 26, 28, 31].

In this paper we consider the independent factor analysis (IFA) model, which generalizes the ordinary factor analysis (FA), principal component analysis (PCA), and independent component analysis (ICA). The IFA model was introduced by Attias [6] as a method for recovering independent hidden sources from their observed mixtures. In the ordinary FA and PCA, the hidden sources are assumed to be uncorrelated and the analysis is based on the covariance matrices, while IFA assumes that the hidden sources (factors) are independent and have unknown distributions. The ICA, in its standard form, assumes that the number of sources is equal to the number of observed variables and that the mixtures are observed without noise. Mixing of sources in realistic situations, however, generally involves noise and different numbers of sources (factors) and observed variables, and the IFA model allows for both of these extensions of ICA.

Most of the existing ICA algorithms concentrate on recovering the mixing matrix and either assume the known distribution of sources or allow for their limited, parametric flexibility [16]. Attias [6] and more recent IFA papers [1, 4, 23] either use mixture of Gaussian distributions as source models or assume that the number of independent sources is known, or both. In the present paper the IFA serves as a dimensionality reduction model for multivariate nonparametric density estimation; we suppose that the distribution of the sources (factors) and their number are unknown.

Samarov and Tsybakov [27] have shown that densities which have the standard, noiseless ICA representation can be estimated at an optimal one-dimensional nonparametric rate, without knowing the mixing matrix of the independent sources. Here our goal is to estimate a multivariate density in the noisy IFA model with unknown number of latent independent components observed in Gaussian noise. It turns out that the density generated by this model can be estimated with a very fast rate. In Section 2 we show that, using recently developed methods of aggregation [18, 19] we can estimate the density of this form at a parametric root-n rate, up to a logarithmic factor independent of the dimension d.

One of the main applications of multivariate density estimators is in the supervised learning. They can be used to construct plug-in classifiers by estimating the densities of each labeled class. Recently, Audibert and Tsybakov [7] have shown that plug-in classifiers can achieve fast rates of the excess Bayes risk and under certain conditions perform better than classifiers based on the (penalized) empirical risk minimization. A difficulty with such density-based plug-in classifiers is that, even when the dimension d is moderately large, most density estimators have poor accuracy in the tails, i.e., in the region which is important for classification purposes. Amato, Antoniadis and Grégoire [2] have suggested to overcome this problem using the ICA model for multivariate data. The resulting method appears to outperform linear, quadratic and flexible discriminant analysis [14] in the training set, but its performance is rather poor in the testing set. Earlier, Polzehl [25] suggested a discrimination-oriented version of projection pursuit density estimation, which appears to produce quite good results but at a high computational cost. His procedure depends on some tuning steps, such as bandwidth selection, which are left open and appear to be crucial for the implementation. More recently, Montanari et al. [23] constructed plug-in classifiers based on the IFA model, with the sources assumed to be distributed according to a mixture of Gaussian distributions, and reported promising numerical results.

In Section 3 we give a bound to the excess risk of nonparametric plug-in classifiers in terms of the MISE of the density estimators of each class. Combining this bound with the results of Section 2, we show that if the data in each class are generated by a noisy IFA model, the corresponding plug-in classifiers achieve, within a logarithmic factor independent of the dimensionality d, the best obtainable rate of the excess Bayes risk. In Section 4 we describe the algorithm implementing our classifier. Section 5 reports results of the application of the algorithm to simulated and real data.

2. Independent factor analysis model for density estimation

We consider the noisy IFA model:

$$\mathbf{X} = A\mathbf{S} + \boldsymbol{\epsilon}, \tag{2.1}$$

where A is a $d \times m$ unknown deterministic matrix of factor loadings with unknown $m < d$, $\mathbf{S}$ is an unobserved m-dimensional random vector with independent zero-mean components (called factors) having unknown distributions with finite variance, and $\boldsymbol{\epsilon}$ is a random vector of noise, independent of $\mathbf{S}$, which we will assume to have d-dimensional normal distribution with zero mean and covariance matrix $\sigma^2 \mathbf{I}_d$, $\sigma^2 > 0$. Here $\mathbf{I}_d$ denotes the $d \times d$ identity matrix.

Assume that we have independent observations $\mathbf{X}_1, \dots, \mathbf{X}_n$, where each $\mathbf{X}_i$ has the same distribution as $\mathbf{X}$. As mentioned in the Introduction, this model is an extension of the ICA model, which is widely used in signal processing for blind source separation. In the signal processing literature the components of $\mathbf{S}$ are called sources rather than factors. The basic ICA model assumes $\boldsymbol{\epsilon} = 0$ and $m = d$ (cf., e.g., [16]). Unlike in the signal processing literature, our goal here is to estimate the target density $p_{\mathbf{X}}(\cdot)$ of $\mathbf{X}$, and model (2.1) serves as a particular form of dimensionality reduction for density estimation.

Somewhat different versions of this model where the signal $\mathbf{S}$ has not necessarily independent components and needs to be non-Gaussian were considered recently in [9, 28]. Blanchard et al. [9] and the follow-up paper by Kawanabe et al. [20] use projection pursuit type techniques to identify the non-Gaussian subspace spanned by the columns of A with known number of columns m, while Samarov and Tsybakov [28] propose aggregation methods to estimate the density of $\mathbf{X}$ when neither the non-Gaussian subspace, nor its dimension are known.

By independence between the noise and the vector of factors $\mathbf{S}$, the target density $p_{\mathbf{X}}$ can be written as a convolution:

$$p_{\mathbf{X}}(\mathbf{x}) = \int_{\mathbb{R}^m} \phi_{d,\sigma^2}(\mathbf{x} - A\mathbf{s}) F_{\mathbf{S}}(d\mathbf{s}), \tag{2.2}$$

where ϕ_{d,σ^2} denotes the density of a d-dimensional Gaussian distribution $N_d(0, \sigma^2 \mathbf{I}_d)$ and $F_{\mathbf{S}}$ is the distribution of $\mathbf{S}$.

Note that (2.2) can be viewed as a variation of the Gaussian mixture model which is widely used in classification, image analysis, mathematical finance and other areas, cf., e.g., [22, 32]. In Gaussian mixture models, the matrix A is the identity matrix, $F_{\mathbf{S}}$ is typically a discrete distribution with finite support, and variances of the Gaussian terms are usually different.

Since in (2.2) we have a convolution with a Gaussian distribution, the density $p_{\mathbf{X}}$ has very strong smoothness properties, no matter how irregular the distribution $F_{\mathbf{S}}$ of the factors is, whether or not the factors are independent, and whether or not the mixing matrix A is known. In the Appendix, we construct a kernel estimator $\hat{p}_n^*$ of $p_{\mathbf{X}}$ such that

$$\mathbb{E}||\hat{p}_n^* - p_{\mathbf{X}}||_2^2 \leq C\frac{(\log n)^{d/2}}{n}, \tag{2.3}$$

where C is a constant and $||\cdot||_2$ is the $L_2(\mathbb{R}^d)$ norm. As in [5, 8] it is not hard to show that the rate given in (2.3) is optimal for the class of densities $p_{\mathbf{X}}$ defined by (2.2) with arbitrary probability distribution $F_{\mathbf{S}}$.

Though this rate appears to be very fast asymptotically, it does not guarantee good accuracy for most practical values of n, even if d is moderately large. For example, if $d = 10$, we have $(\log n)^{d/2} > n$ for all $n \leq 10^5$.

We will make further assumptions on the model (2.1) which will allow us to eliminate this dependence of the rate on the dimension d. It is well known that the standard, covariance-based factor analysis model is not fully identifiable without extra assumptions (see, e.g., [3]). Indeed, the factors are defined only up to an arbitrary rotation. The independence of factors assumed in (2.1) excludes this indeterminacy provided that at most one factor is allowed to have a Gaussian distribution. This last assumption, needed for the identifiability of A, is standard in the ICA literature, see, e.g., [16]. However, we will not need it in this paper because it turns out that for estimation of the density under the assumptions listed below it suffices to identify A only up to a permutation of its columns.

We will collect here the assumptions used in the paper.

Assumption 1. The columns of the matrix A are orthonormal.

Assumption 2. The number of factors m does not exceed an upper bound M, $M < d$.

Assumption 3. The M largest eigenvalues of the covariance matrix $\Sigma_{\mathbf{X}}$ of the random vector $\mathbf{X}$ are distinct and the 4th moments of the components of $\mathbf{X}$ are finite.

Assumption 1 is rather restrictive but, as we show below, together with the assumed independence of the factors, it is crucial to obtain a weak dependency of the rate in (2.3) on the dimension d.

Assumption 2 means that model (2.1) indeed provides the dimensionality reduction. The assumption $M < d$ is only needed to estimate the variance σ^2 of the noise; if σ^2 is known we can allow $M = d$.

Assumption 3 is needed to establish, in the proof of Theorem 2.1, root-n consistency of the eigenvectors of the sample covariance matrix of $\mathbf{X}$.

In order to construct our estimator, we first consider the estimation of $p_{\mathbf{X}}$ when the dimension m, the mixing matrix A, and the level of noise σ^2 are specified; the fact that none of these quantities is known is addressed later in this section.

Since the columns of A are orthonormal, we have $A^T\mathbf{X} = \mathbf{S} + A^T\boldsymbol{\epsilon}$ and

$$\begin{aligned}\phi_{d,\sigma^2}(\mathbf{x} - A\mathbf{s}) &= \left(\frac{1}{2\pi\sigma^2}\right)^{d/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{x} - A\mathbf{s})^T(\mathbf{x} - A\mathbf{s})\right\} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{d/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{s} - A^T\mathbf{x})^T(\mathbf{s} - A^T\mathbf{x})\right\} \\ &\quad \exp\left\{-\frac{1}{2\sigma^2}\mathbf{x}^T(\mathbf{I}_d - AA^T)\mathbf{x}\right\}.\end{aligned}$$

Substitution of the above expression in (2.2) gives:

$$p_{\mathbf{X}}(\mathbf{x}) = \left(\frac{1}{2\pi\sigma^2}\right)^{(d-m)/2} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{x}^T(\mathbf{I}_d - AA^T)\mathbf{x}\right\}\int_{\mathbb{R}^m} \phi_{m,\sigma^2}(\mathbf{s}-A^T\mathbf{x})F_{\mathbf{S}}(d\mathbf{s}).$$

Now, by independence of the factors, we get:

$$p_{\mathbf{X}}(\mathbf{x}) \equiv p_{m,A}(\mathbf{x}) = \left(\frac{1}{2\pi\sigma^2}\right)^{(d-m)/2} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{x}^T(\mathbf{I}_d - AA^T)\mathbf{x}\right\}\prod_{k=1}^m g_k(\mathbf{a}_k^T\mathbf{x}) \tag{2.4}$$

where $\mathbf{a}_k$ denotes the kth column of A and

$$g_k(u) = \int_{\mathbb{R}} \phi_{1,\sigma^2}(u - s)F_{S_k}(ds),$$

where F_{S_k} denotes the distribution of the kth component S_k of $\mathbf{S}$. We see that to estimate the target density $p_{\mathbf{X}}$ it suffices to estimate nonparametrically each one-dimensional density g_k using the projections of an observed sample $\mathbf{X}_1, \dots, \mathbf{X}_n$ generated by the model 2.1) onto the kth direction $\mathbf{a}_k$.

Note that, similarly to (2.2), the density g_k is obtained from convolution with a one-dimensional Gaussian density, and therefore has very strong smoothness properties. To estimate g_k we will use the kernel estimators

$$\hat{g}_k(x) = \frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{x - \mathbf{a}_k^T\mathbf{X}_i}{h_n}\right), \quad k = 1, \dots, m, \tag{2.5}$$

with a bandwidth $h_n \asymp (\log n)^{-1/2}$ and the sinc function kernel $K(u) = \sin u/\pi u$. We could also use here any other kernel K whose Fourier transform is bounded and compactly supported, for example, the de la Vallée-Poussin kernel $K(u) = (\cos(u)-\cos(2u))/(\pi u^2)$, which is absolutely integrable and therefore well suited for studying the L_1-error.

A potential problem of negative values of $\hat{g}_k$ in the regions where the data are sparse can be corrected using several methods (see, for example, [12, 13]). For our practical implementation we will follow the method suggested in [13], and our estimators will be obtained by truncating the estimator $\hat{g}_k(x)$ outside the “central” range where it is nonnegative, and then renormalizing.

Once each “projection” density g_k is estimated by the corresponding kernel estimator (2.5), the full target density $p_{\mathbf{X}}$ is then estimated using (2.4):

$$\hat{p}_{n,m,A}(\mathbf{x}) = \left(\frac{1}{2\pi\sigma^2}\right)^{(d-m)/2} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{x}^T(\mathbf{I}_d - AA^T)\mathbf{x}\right\} \prod_{k=1}^{m} \hat{g}_k(\mathbf{a}_k^T\mathbf{x}). \quad (2.6)$$

The following proposition proved in the Appendix summarizes the discussion for the case when A and σ^2 are known.

Proposition 2.1. *Consider a random sample of size n from the density $p_{\mathbf{X}}$ given by (2.4) with known A and σ^2 and let Assumption 1 hold. Then the estimator (2.6) with $\hat{g}_k$ given in (2.5) has the mean integrated square error of the order $(\log n)^{1/2}/n$:*

$$\mathbb{E}\|\hat{p}_{n,m,A} - p_{\mathbf{X}}\|_2^2 = \mathcal{O}\left(\frac{(\log n)^{1/2}}{n}\right).$$

Note that neither m nor d affect the rate. Note also that Proposition 2.1 is valid with no assumption on the distribution of the factors. The identifiability assumption (that at most one factor is allowed to have a Gaussian distribution) is not used in the proof, since we do not estimate the matrix A. Also in Proposition 2.1 we do not use the assumption that the factors have finite variances. This assumption is needed to make possible the estimation of the variance σ^2, while in Proposition 2.1 the variance is a given value.

So far in this section we have assumed that A and σ^2 are known. When σ^2 is an unknown parameter, it is still possible to obtain the same rates based on the approach outlined above, provided that the dimensionality reduction holds in the strict sense, i.e., $m < d$. Indeed, assume that we know an upper bound M for the number of factors m and that $M < d$, i.e. that Assumption 2 holds.

The assumed independence and finite variance of the factors imply that their covariance matrix, which we will denote by W, is diagonal. The covariance matrix $\Sigma_{\mathbf{X}}$ of $\mathbf{X}$ is given by:

$$\Sigma_{\mathbf{X}} = AWA^T + \sigma^2\mathbf{I}_d.$$

If $\lambda_1(\Sigma_{\mathbf{X}}) \geq \cdots \geq \lambda_d(\Sigma_{\mathbf{X}})$ denote the eigenvalues of $\Sigma_{\mathbf{X}}$ sorted in decreasing order, then $\lambda_i(\Sigma_{\mathbf{X}}) = w_i + \sigma^2$, for $i = 1, \ldots, m$, and $\lambda_i(\Sigma_{\mathbf{X}}) = \sigma^2$ for $i > m$, where w_i denote the diagonal elements of W. We estimate σ^2 with

$$\hat{\sigma}^2 = \frac{1}{d-M}\sum_{i=M+1}^{d} \hat{\lambda}_i,$$

where $\hat{\lambda}_i$, $i = 1, \ldots, d$, are the eigenvalues of the sample covariance matrix $\hat{\Sigma}_{\mathbf{X}}$ arranged in decreasing order. Note that $\hat{\sigma}^2$ is a root-n consistent estimator.

Indeed, the root-n consistency of each $\hat{\lambda}_i$ is a consequence of elementwise root-n consistency of $\hat{\Sigma}_{\mathbf{X}}$ and a classical inequality from matrix perturbation theory:

$$|\lambda_i(C+D) - \lambda_i(C)| \leq |||D|||, \quad i = 1, 2, \ldots, d,$$

where C and D are any symmetric matrices and $|||D|||$ is the spectral norm of D, see, e.g. Corollary 4.10, p.203, in [30].

Using the root-n consistency of $\hat{\sigma}^2$, it is not hard to show that the estimation of σ^2 does not affect a slower density estimator rate, and so in what follows we will assume that σ^2 is known.

Consider now the case where the matrix A, and hence its rank m, are unknown. We will use a model selection type aggregation procedure similar to the one developed recently in [28] and, more specifically, the mirror averaging algorithm of [19]. We aggregate estimators of the type (2.6) corresponding to candidate pairs $(k, \hat{B}_k)$, $k = 1, \ldots, M$. Here $\hat{B}_k$ is a $d \times k$ matrix whose columns are the first k (in the decreasing order of eigenvalues) orthonormal eigenvectors of $\hat{\Sigma}_{\mathbf{X}} - \hat{\sigma}^2 \mathbf{I}_d$ (and thus of $\hat{\Sigma}_{\mathbf{X}}$). Note also that the columns of A are orthonormal eigenvectors of $\Sigma_{\mathbf{X}}$ corresponding to its first m largest eigenvalues. In view of this, when the true number of factors is m, it follows from Lemma A.2 in the Appendix that, provided Assumption 3 holds, $\hat{B}_m$ is a $\sqrt{n}$-consistent estimator of A.

We can now define the aggregate estimator, applying the results of [19] in our framework. We split the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ in two parts, $\mathcal{D}_1$ and $\mathcal{D}_2$ with $n_1 = \mathrm{Card}(\mathcal{D}_1)$, $n_2 = \mathrm{Card}(\mathcal{D}_2)$, $n = n_1 + n_2$. From the first subsample $\mathcal{D}_1$ we construct the estimators

$$\hat{p}_k(\mathbf{x}) \equiv \hat{p}_{n_1,k,\hat{B}_k}(\mathbf{x}) = \left(\frac{1}{2\pi\sigma^2}\right)^{(d-k)/2} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{x}^T(\mathbf{I}_d - \hat{B}_k\hat{B}_k^T)\mathbf{x}\right\} \prod_{j=1}^{k} \hat{g}_j(\mathbf{b}_j^T\mathbf{x}) \tag{2.7}$$

for $k = 1, \ldots, M$, where $\mathbf{b}_j$ denotes the jth column of $\hat{B}_k$, the estimators $\hat{g}_j(\cdot)$ are defined in (2.5), and both $\hat{B}_k$ and $\hat{g}_j(\cdot)$ are based only on the first subsample $\mathcal{D}_1$.

The collection $\mathcal{C}$ of density estimators $\left\{\hat{p}_{n_1,k,\hat{B}_k},\ k = 1, \ldots, M\right\}$ of the form (2.7) constructed from the subsample $\mathcal{D}_1$ can be considered as a collection of fixed functions when referring to the second subsample $\mathcal{D}_2$. The cardinality of this collection is M.

To proceed further, we need some more notation. Let Θ be the simplex

$$\Theta = \left\{\boldsymbol{\theta} \in \mathbb{R}^M\ : \sum_{k=1}^{M} \theta_k = 1,\ \theta_k \geq 0,\ k = 1, \ldots, M\right\},$$

and

$$\mathbf{u}(\mathbf{X}) = (u_1(\mathbf{X}), \ldots, u_M(\mathbf{X}))^T,$$

where

$$u_k(\mathbf{x}) = \int \hat{p}_k^2(\mathbf{x})d\mathbf{x} - 2\hat{p}_k(\mathbf{x}). \tag{2.8}$$

Introduce the vector function

$$\mathbf{H}(\mathbf{x}) = (\hat{p}_1(\mathbf{x}), \dots, \hat{p}_M(\mathbf{x}))^T .$$

As in [19], the goal of aggregation is to construct a new density estimator $\tilde{p}_n(\mathbf{x})$ of the form

$$\tilde{p}_n(\mathbf{x}) = \tilde{\boldsymbol{\theta}}^T \mathbf{H}(\mathbf{x}) \tag{2.9}$$

which is nearly as good in terms of the L_2-risk as the best one in the collection $\mathcal{C}$. Using the mirror averaging algorithm, the aggregate weights $\tilde{\boldsymbol{\theta}}$ are computed by a simple procedure which is recursive over the data. Starting with an arbitrary value $\tilde{\boldsymbol{\theta}}^{(0)} \in \Theta$, these weights are defined in the form:

$$\tilde{\boldsymbol{\theta}} = \frac{1}{n_2} \sum_{\ell=1}^{n_2} \tilde{\boldsymbol{\theta}}^{(\ell-1)}, \tag{2.10}$$

where the components of $\tilde{\boldsymbol{\theta}}^{(\ell)}$ are given by

$$\tilde{\theta}_k^{(\ell)} = \frac{\exp\left(-\beta^{-1} \sum_{r=1}^{\ell} u_k(\mathbf{X}_r)\right)}{\sum_{t=1}^{M} \exp\left(-\beta^{-1} \sum_{r=1}^{\ell} u_t(\mathbf{X}_r)\right)}, \quad k = 1, \dots, M, \tag{2.11}$$

with $\mathbf{X}_r$, $r = 1, \dots, n_2$, denoting the elements of the second subsample $\mathcal{D}_2$. Here $\beta > 0$ is a random variable measurable w.r.t. the first subsample $\mathcal{D}_1$.

Our main result about the convergence of the aggregated density estimator is given in Theorem 2.1 below. We will consider the norms restricted to a Euclidean ball $B \subset \mathbb{R}^d$: $\|f\|_{2,B}^2 = \int_B f^2(\mathbf{x})d\mathbf{x}$, $\|f\|_{\infty,B} = \sup_{t\in B} |f(t)|$ for $f : \mathbb{R}^d \to \mathbb{R}$. Accordingly, in Theorem 2.1 we will restrict our estimators to B and define $\tilde{p}_n$ by the above aggregation procedure where $\hat{p}_k(x)$ are replaced by $\hat{p}_k(x)I\{x \in B\}$. Here $I\{\cdot\}$ denotes the indicator function.

Clearly, all densities $p_{\mathbf{X}}$ of the form (2.4) are bounded: $\|p_{\mathbf{X}}\|_{\infty,B} \le L_0 := (2\pi\sigma^2)^{-d/2}$ for all m and A (as mentioned earlier, we assume that σ^2 is known because it can be estimated at the root-n rate). We set $\hat{L}_1 = \max_{k=1,\dots,M} \|\hat{p}_k\|_{\infty,B}$ and $\hat{L} = \max(L_0, \hat{L}_1)$. In the Appendix we prove that

$$\mathbb{E}\|\hat{p}_k\|_{\infty,B} \le L', \quad \forall k = 1, \dots, M, \tag{2.12}$$

where L' is a constant.

Theorem 2.1. *Let $p_{\mathbf{X}}$ be the density of $\mathbf{X}$ in model (2.1) and Assumptions 1 through 3 hold. Let $n_2 = \lceil cn/\sqrt{\log n}\rceil$ for some constant $c > 0$ such that $1 \le n_2 < n$. Then for $\beta = 12\hat{L}$, the aggregate estimator $\tilde{p}_n$ with $\tilde{\boldsymbol{\theta}}$ obtained by the mirror averaging algorithm restricted to a Euclidean ball B satisfies*

$$\mathbb{E}\|\tilde{p}_n - p_{\mathbf{X}}\|_{2,B}^2 = \mathcal{O}\left(\frac{(\log n)^{1/2}}{n}\right), \tag{2.13}$$

as $n \to +\infty$.

The theorem implies that the estimator $\tilde{p}_n$ adapts to the unknown m and A, i.e., has the same rate, independent of m and d, as in the case when the dimension m and the matrix A are known. The proof is given in the Appendix.

Remarks.

1. Note that Theorem 2.1 holds with mild assumptions on distributions of the factors. In particular, we do not need the factors to have discrete distributions as in standard mixture models or to have densities with respect to the Lebesgue measure.

2. We state Theorem 2.1 with a restricted L_2-norm $\|\cdot\|_{2,B}$. Under mild assumptions on the densities of the factors we can extend it to the L_2-norm on $\mathbb{R}^d$. Indeed, inspection of the proof shows that Theorem 2.1 remains valid for balls B of radius r_n which tends to infinity slowly enough as $n \to \infty$. If $p_{\mathbf{X}}$ behaves itself far from the origin roughly as a Gaussian density (which is true under mild assumptions on factor densities), then the integral of $p_{\mathbf{X}}^2$ outside of the ball reduces to a value smaller than the right hand side of (2.13).

3. Application to nonparametric classification

One of the main applications of multivariate density estimators is in the supervised learning, where they can be used to construct plug-in classifiers by estimating the densities of each labeled class. The difficulty with such density-based plug-in classifiers is that, even for moderately large dimensions d, standard density estimators have poor accuracy in the tails, i.e., in the region which is important for classification purposes. In this section we consider the nonparametric classification problem and bound the excess misclassification error of a plug-in classifier in terms of the MISE of class-conditional density estimators. This bound implies that, for the class-conditional densities obeying the noisy IFA model (2.2), the resulting plug-in classifier has nearly optimal excess error.

Assume that we have J independent training samples $\{X_{j1}, \dots, X_{jN_j}\}$ of sizes N_j, $j = 1, \dots, J$, from J populations with densities $f_1, \dots, f_J$ on $\mathbb{R}^d$. We will denote by $\mathcal{D}$ the union of training samples. Assume that we also have an observation $\mathbf{X} \in \mathbb{R}^d$ independent of these samples and distributed according to one of the f_j. The classification problem consists in predicting the corresponding value of the class label $j \in \{1, \dots, J\}$. We define a classifier or prediction rule as a measurable function $T(\cdot)$ which assigns a class membership based on the explanatory variable, i.e., $T : \mathbb{R}^d \to \{1, \dots, J\}$. The misclassification error associated with a classifier T is usually defined as

$$R(T) = \sum_{j=1}^{J} \pi_j \mathbb{P}_j(T(\mathbf{X}) \neq j) = \sum_{j=1}^{J} \pi_j \int_{\mathbb{R}^d} I(T(\mathbf{x}) \neq j) f_j(\mathbf{x}) d\mathbf{x},$$

where $\mathbb{P}_j$ denotes the class-conditional population probability distribution with density f_j, and π_j is the prior probability of class j. We will consider a slightly more general definition:

$$R_B(T) = \sum_{j=1}^{J} \pi_j \int_B I(T(\mathbf{x}) \neq j) f_j(\mathbf{x}) d\mathbf{x},$$

where B is a Borel subset of $\mathbb{R}^d$. The Bayes classifier T^* is the one with the smallest misclassification error:

$$R_B(T^*) = \min_T R_B(T).$$

In general, the Bayes classifier is not unique. It is easy to see that there exists a Bayes classifier T^* which does not depend on B and which is defined by

$$\pi_{T^*(\mathbf{x})} f_{T^*(\mathbf{x})}(\mathbf{x}) = \min_{1 \leq j \leq J} \pi_j f_j(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^d.$$

A classifier trained on the sample $\mathcal{D}$ will be denoted by $T_{\mathcal{D}}(\mathbf{x})$. A key characteristic of such a classifier is the misclassification error $R_B(T_{\mathcal{D}})$. One of the main goals in statistical learning is to construct a classifier with the smallest possible excess risk

$$\mathcal{E}(T_{\mathcal{D}}) = \mathbb{E} R_B(T_{\mathcal{D}}) - R_B(T^*).$$

We consider plug-in classifiers $\hat{T}(\mathbf{x})$ defined by:

$$\pi_{\hat{T}(\mathbf{x})} \hat{f}_{\hat{T}(\mathbf{x})}(\mathbf{x}) = \min_{1 \leq j \leq J} \pi_j \hat{f}_j(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^d$$

where $\hat{f}_j$ is an estimator of density f_j based on the training sample $\{X_{j1}, \dots, X_{jN_j}\}$.

The following proposition relates the excess risk $\mathcal{E}(\hat{T})$ of plug-in classifiers to the rate of convergence of the estimators $\hat{f}_j$.

Proposition 3.1.

$$\mathcal{E}(\hat{T}) \leq \sum_{j=1}^{J} \pi_j \, \mathbb{E} \int_B |\hat{f}_j(\mathbf{x}) - f_j(\mathbf{x})| d\mathbf{x}$$

Proof of the proposition is given in the Appendix.

Assume now that the class-conditional densities follow the noisy IFA model (2.2) with different unknown mixing matrices and that $N_j \asymp n$ for all j. Let B be a Euclidean ball in $\mathbb{R}^d$ and define each of the estimators $\hat{f}_j$ using the mirror averaging procedure as in the previous section. Then, using Theorem 2.1, we have

$$\mathbb{E} \int_B |\hat{f}_j(\mathbf{x}) - f_j(\mathbf{x})| d\mathbf{x} \leq \sqrt{|B|} \, \mathbb{E} \|\hat{f}_j - f_j\|_{2,B} = \mathcal{O}\left(\frac{(\log n)^{1/4}}{\sqrt{n}} \right)$$

as $n \to \infty$, where $|B|$ denotes the volume of the ball B. Thus, the excess risk $\mathcal{E}(\hat{T})$ converges to 0 at the rate $(\log n)^{1/4}/\sqrt{n}$ independently of the dimension d. Following the argument in [11] or [35], it is easy to show that this is the best obtainable rate for the excess risk, up to the $\log^{1/4} n$ factor.

4. The algorithm

In this section we discuss numerical aspects of the proposed density estimator.

Clearly, one-dimensional kernel density estimators $\hat{g}_k$ with given bandwidth, say $h_n \propto (\log n)^{-1/2}$, can be computed in a fast way. Similarly, estimating the variance of the noise component in the noisy IFA model amounts to implementing a single singular value decomposition (SVD) of the $d \times n$ data matrix $D = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$. Let $D = V \Lambda U^T$ be the SVD of D, where Λ is the diagonal matrix and U, V are matrices with orthonormal columns. We assume w.l.o.g. that $\mathbf{X}_i$ are centered. Then an estimate of the variance $\hat{\sigma}_k^2$ with rank k approximation, $k \le M$, is given by

$$\hat{\sigma}_k^2 = \frac{1}{d-k} \sum_{i=k+1}^{d} s_i^2, \quad k = 1, \ldots, M \tag{4.1}$$

where s_i are the diagonal elements of $\Lambda/\sqrt{n}$ sorted in the decreasing order. When the index matrix A is unknown, the rank k approximation $\hat{B}_k$ of A used in the density estimator $\hat{p}_k$, cf. (2.7), can be easily obtained from the SVD of D. Indeed, we can take $\hat{B}_k = V_k$, where V_k is formed by the first k columns of V. So, accurate computation of the density estimators (2.7) is feasible, reasonably fast and does not require a huge amount of memory even for very large n and d.

Therefore, the complexity of the procedure is controlled by the numerical implementation of the mirror averaging algorithm which, in particular, requires the computation of the score functions $u_k(\mathbf{x})$, involving integration of $\hat{p}_k^2$, see (2.8). The numerical implementation of the integral of the square of density estimates $\hat{p}_k$ in $\mathbb{R}^d$ can be realized by means of cubature formulas. Recall that for the calculation of $\int \hat{p}_k(\mathbf{x})^2 d\mathbf{x}$, say, a cubature has the form $\sum_{i=1}^N w_i \hat{p}_k^2(\mathbf{x}_i)$ where $\mathbf{x}_i$ are the nodes and w_i are the associated weights. In our setting, M integrals involving the $\hat{B}_k$-projections need to be calculated for each θ_k, so formulas with fixed nodes will be actually more economical. On multidimensional domains, product quadratures quickly become prohibitive (they grow exponentially in d for the same accuracy), and therefore this approach is not realistic.

An alternative is to use Monte-Carlo integration methods which require much more evaluations but do not depend on the dimension d, or a more clever implementation through Gibbs sampling by generating samples from some suitable distribution for the Monte-Carlo estimates. Several Gibbs sampling strategies were considered in the present work. The fastest one was to generate samples directly from $\hat{p}_k$, so that

$$\int \hat{p}_k^2(\mathbf{x}) d\mathbf{x} \simeq \frac{1}{Q} \sum_{i=1}^{Q} \hat{p}_k(\mathbf{x}_i),$$

where Q is the number of generated i.i.d. random realizations $\mathbf{x}_i$ from the density $\hat{p}_k$.

The overall algorithm implementing our approach is the following:

Algorithm 4.1.

```
- Compute the singular value decomposition of the data array $D$:
      $D = V\Lambda U^T,$
  with matrices $U$, $V$, and $\Lambda$ having dimensions $n \times d$, $d \times d$ and $d \times d$, respectively;
- for $k$=1,...,$M$
    Take $\hat{B}_k$ as the matrix built from the first $k$ columns of $V$;
    Compute $\hat{\sigma}_k^2$ from (4.1);
    Compute the density estimator $\hat{p}_k(\mathbf{x})$ from (2.7) based on the subsample $\mathcal{D}_1$;
    Compute $u_k(\mathbf{x})$ from (2.8).
- end for
- Estimate the weights through (2.10)–(2.11) and output the final density estimator (2.9).
```

To speed up computations, one-dimensional kernel density estimators $\hat{g}_j$, $j = 1, \dots, M$, in (2.7) are obtained through a Fast Fourier Transform algorithm, cf. [29].

The algorithm for estimating $\int \hat{p}_k^2(\mathbf{x}) d\mathbf{x}$ in (2.8) goes through the following steps.

Algorithm 4.2.

- Generate Q independent random numbers, $y_k^{(i)}$, $i = 1, \dots, Q$, from each $\hat{g}_k$, $k = 1, \dots, M$, and compute the corresponding density $\hat{g}_k(y_k^{(i)})$ by kernel density estimation;
- Generate the corresponding d-dimensional $\mathbf{x}^{(i)}$ as $\mathbf{x}^{(i)} = \hat{B}_k \mathbf{y}^{(i)} + (I_d - \hat{B}_k \hat{B}_k^T)\boldsymbol{\epsilon}^{(i)}$, $\mathbf{y}^{(i)} \equiv (y_1^{(i)}, \dots, y_k^{(i)})$, with $\boldsymbol{\epsilon}^{(i)}$ being random numbers extracted from a d-variate Gaussian density function having 0 mean and diagonal covariance $\hat{\sigma}_k^2 I_d$;
- Compute $\hat{p}_k(\mathbf{x}^{(i)})$ through (2.7);
- Output the estimate $\frac{1}{Q}\sum_{i=1}^{Q} \hat{p}_k(\mathbf{x}^{(i)})$ of the integral $\int \hat{p}_k^2(\mathbf{x}) d\mathbf{x}$.

Here Q is chosen so that generating more random numbers does not change the estimated value of the integral within a predefined tolerance. Random numbers generated from the density estimator $\hat{g}_k$ are based on the corresponding cumulative functions and pre-computed on a high resolution grid with linear interpolation.

5. Simulations and examples

5.1. Density estimation

To study the performance of density estimates based on our noisy IFA model we have conducted an extensive set of simulations. As samples drawn form factors $\mathbf{s}$, we used data generated from a variety of source distributions, including

Table 1

List of basic functions considered for the numerical experiments. $\mathcal{G}(q,r)$ stands for Gaussian distribution with mean q and standard deviation r; $\chi^2(r)$ indicates chi-square density function with r degrees of freedom; $\gamma(r)$ is Gamma distribution of parameters r and 1; $t(r)$ is Student distribution with r degrees of freedom

Index	Test function
1	$\mathcal{G}(0,1)$
2	$\chi^2(1)$
3	$0.5\mathcal{G}(-3,1)+0.5\mathcal{G}(2,1)$
4	$0.4\gamma(5,1)+0.6\gamma(13,1)$
5	$\chi^2(8)$
6	$t(5)$
7	Double exponential : $\exp(-\lvert x\rvert)$

subgaussian and supergaussian distributions, as well as distributions that are nearly Gaussian. We studied unimodal, multimodal, symmetric, and nonsymmetric distributions. Table 1 lists the basic (one-dimensional) test densities from which multidimensional density functions are built.

Experiments were run up to dimension $d = 10$ with a number of independent factors equal to 1 and 2. Random i.i.d. noise was generated and added to the simulated signals so that the Signal to Noise Ratio (SNR) was equal to 3, 5 or 7, where from (2.1) and by Assumption 1 the SNR is computed as

$$\mathrm{SNR} = \frac{\sum_{k=1}^{M} \mathrm{var}(\mathbf{s}_\mathrm{k})}{D\sigma^2}$$

The kernels K for density estimators $\hat{g}_j$ in (2.7) were the Gaussian, the sinc and de la Vallée-Poussin kernels; the bandwidth h was chosen as $h = \sigma/\log^{1/2} n$. To obtain legitimate (i.e., nonnegative) density functions they were post-processed by the procedure of [13]. The size of the sample was chosen as n=200, 300, 500, 700, 1000, 2000 and 4000. In order to apply the aggregate estimation involved in Noisy IFA, the sample has been randomly split in two parts $\mathcal{D}_1$ and $\mathcal{D}_2$ of equal size $n_1 = n_2 = n/2$. The following criterion was used for evaluating the performance of density estimators:

$$I_1 := 100\left(1 - \frac{\int \left(p_{\mathrm{estimated}}(\mathbf{x}) - p_{\mathbf{X}}(\mathbf{x})\right)^2 d\mathbf{x}}{\int p^2_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}}\right). \tag{5.1}$$

The performance of IFA density estimation was compared with kernel smoothing (KS) [34] as implemented in the `KS` package available in `R`. IFA density estimation has been implemented in the `MATLAB` environment and the scripts are available upon request. We note that KS can be effectively computed only up to $d = 6$ if the FFT algorithm is used. In contrast with this, our method has no practical restrictions on the dimension. This is due to the use of a proper Gibbs sampling for estimating integrals (2.8); in addition the density estimate can be computed on any set in $\mathbb{R}^d$, not necessarily on a lattice imposed by the FFT.

We conducted numerical experiments by generating random samples of size n from the independent components of Table 1, random mixing matrices, and

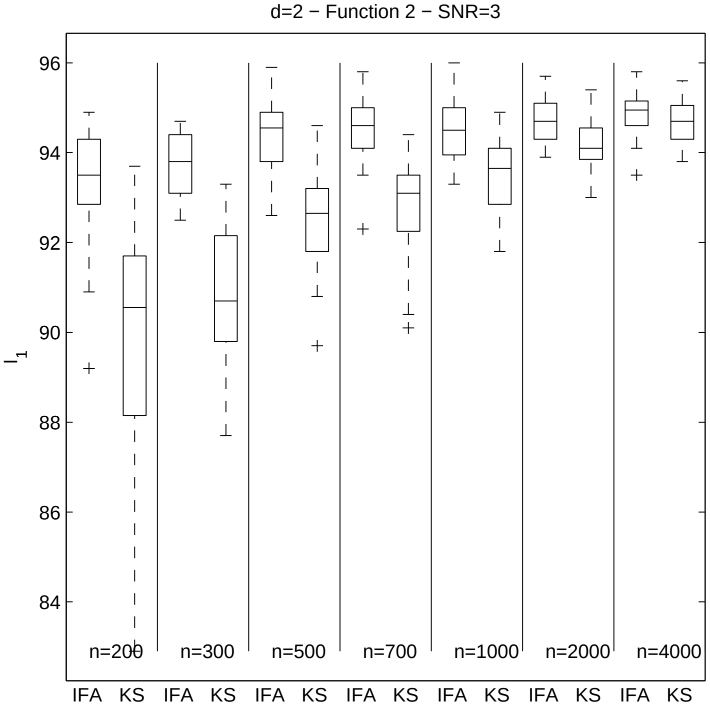

FIG 1. *Boxplot of the error criterion I_1 (Eq. (5.1)) in the case $d = 2$, Signal to Noise Ratio 3 and test function 2 for several sample sizes.*

different realizations of Gaussian noise. In particular, the elements of the mixing matrix A were generated as i.i.d. standard Gaussian random variables and then the matrix was orthonormalized by a Gram-Schmidt procedure. We perform 50 Monte-Carlo replications for each case and output the corresponding values I_1. Results over all experiments show a very good performance of Noisy IFA. For brevity we only show some representative figures in the form of boxplots. We display different test functions to demonstrate good performances over all of them. Moreover, we present only the case of SNR=3 because it seems to be more interesting for applications and because improvement of performance for both methods flattens the differences. Figure 1 shows the case of $d = 2$, SNR=3 and test function 2 (chi-square function), where the superiority of the aggregated Noisy IFA with respect to KS is clear. Figure 2 shows analogous boxplots in the case $d = 3$ and test function 3 (mixture of Gaussians), again when SNR=3. This case is interesting because the dimension d is larger, whereas the number of independent factors is kept constant with respect to the previous experiment. Figure 2 clearly shows that difference of performance between Noisy IFA and KS increases in favor of the former. Figure 3 shows boxplots in the case $d = 5$ and test functions 5 and 6 (chi-square and Student, respectively), again for SNR=3.

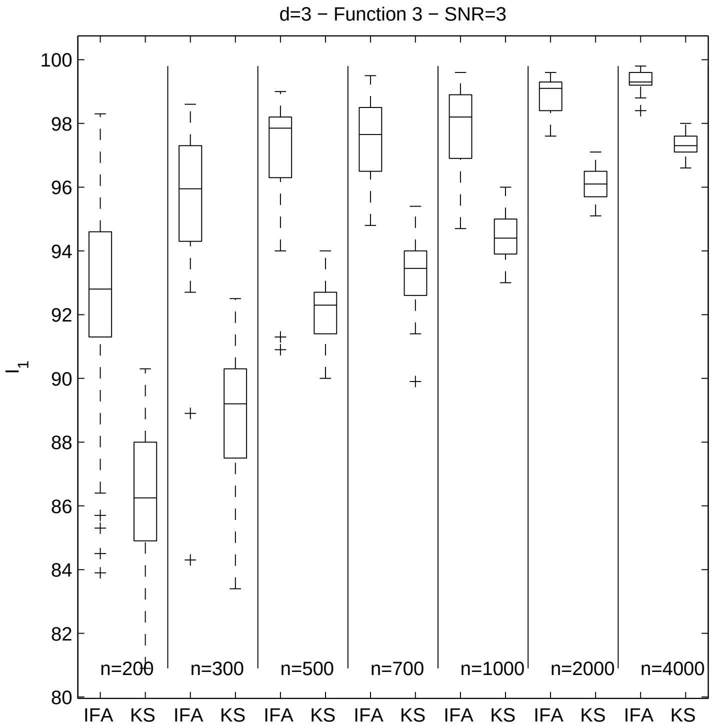

Fig 2. *Boxplot of the error criterion I_1 (Eq. (5.1)) in the case $d = 3$, Signal to Noise Ratio 3 and test function 3 for several sample sizes.*

Better performance of Noisy IFA with respect to KS is confirmed, especially when d increases. Finally, Figure 4 refers to the case $d = 10$ and test function 4 (mixture of Gammas) again for SNR=3. KS boxplots are not shown because the software is available only for $d \leq 6$. The figure shows that accuracy of NoisyIFA is very good also for higher dimensions.

Finally, Table 2 shows typical computational times of aggregated IFA and KS density estimators. Executions were run on a single core 64-bit Opteron 248 processor with MATLAB version R2008a, R 2.9.0 and Linux Operating System. We see that the aggregated IFA is more than one order of magnitude faster than KS.

Table 2

Computational time (sec) of aggregated IFA and KS for some test configurations

Experiment	Aggregated IFA	KS
$d = 2$, $n = 500$	0.3	3
$d = 3$, $n = 500$	0.9	15
$d = 5$, $n = 500$	4	120

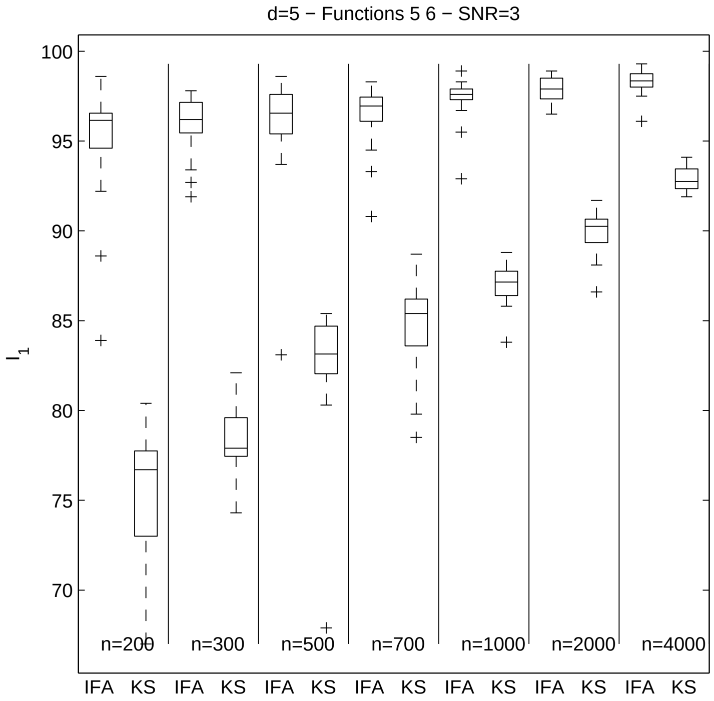

FIG 3. *Boxplot of the error criterion I_1 (Eq. (5.1)) in the case $d = 5$, Signal to Noise Ratio 3 and test functions 5 and 6 for several sample sizes.*

5.2. Classification: A real data example

In this subsection we apply the nonparametric classification method suggested in Section 3 to real data. We consider only a two-class problem and we assume that the class-conditional distributions follow the noisy IFA model. To evaluate the performance of our approach in comparison with other classification methods that are often used in this context, we have also applied to these data three other classification procedures, one parametric and two nonparametric:

LDA (Linear Discriminant Analysis). It relies on the estimate of the class-conditional density functions (supposed to be Gaussian with a common covariance matrix among classes), and on the consequent separation of the two classes by a hyperplane in d-dimensional space.

NPDA (Nonparametric Discriminant Analysis [2]). In this procedure class-conditional density functions are estimated nonparametrically by the kernel method, assuming that the density obeys an ICA model. The kernel functions mentioned above in this section were considered in the experiments. The smoothing procedure uses an asymptotic estimate of the bandwidth and a correction for getting non-negative density estimators.

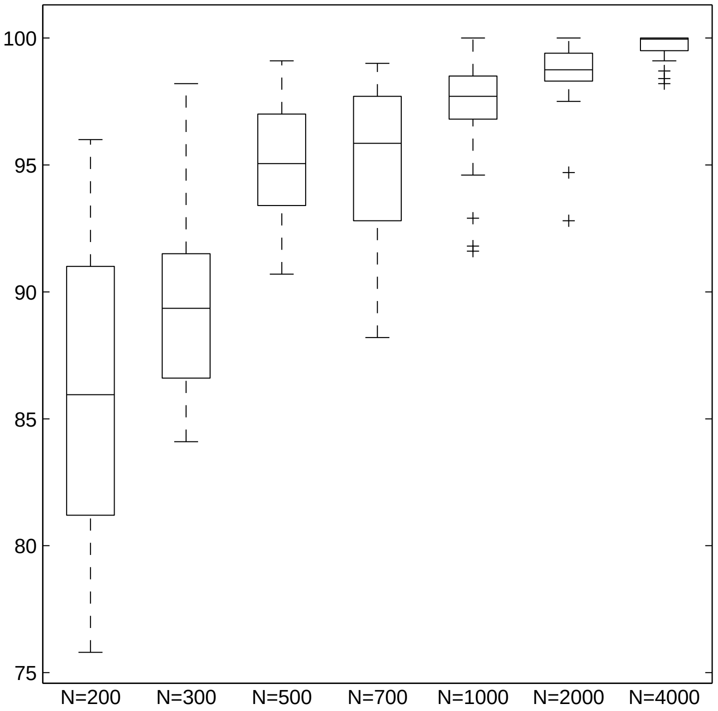

Fig 4. *Boxplot of the error criterion I_1 (Eq. (5.1)) in the case $d = 10$, Signal to Noise Ratio 3 and test function 4 for several sample sizes.*

FDA (Flexible Discriminant Analysis [14]). This method is also nonparametric, but classification is performed through an equivalent regression problem where the regression function is estimated by the spline method.

We have compared the performance of the classification methods on a data set from a remote sensing experiment. MSG (METEOSAT Second Generation) is a series of geostationary satellites launched by EUMETSAT (EUropean organization for the exploitation of METeorological SATellites) mainly aimed at providing data useful for the weather forecast. A primary instrument onboard MSG is SEVIRI, a radiometer measuring radiance emitted by Earth at $d = 11$ spectral channels having a resolution of 3 Km^2 at sub-satellite point. Essentially, SEVIRI produces 11 images of the whole Earth hemisphere centered at 0^o degrees latitude every 15 minutes. Recognizing whether each pixel of the images is clear or affected by clouds (cloud detection) is a mandatory preliminary task for any processing of satellite data. In this respect multispectral radiance data are prone to improve the detectability of clouds, thanks to the peculiar behavior of clouds in selected spectral bands. Figure 5 shows an RGB image of the Earth taken by SEVIRI on June 30th 2006 UTC time 11:12 composed by 3 selected spectral channels. This problem is faced by classification where starting from a variate with dimension $d = 11$ (i.e., radiance measured over pixels

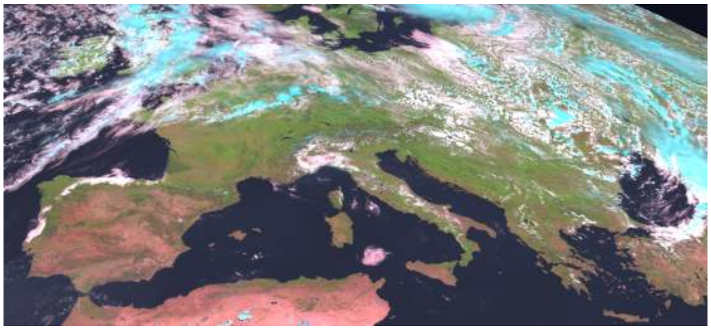

Fig 5. *RGB image obtained from the SEVIRI sensor onboard MSG on June 30th 2006 UTC Time 11:12.*

of the 11 images), a 2-class label (clear or cloudy) is assigned to each pixel. In order to accomplish this task by supervised classification a training set has to be defined. Here we take the training set from a cloud mask produced by sensor MODIS onboard NOAA EOS series satellites. MODIS sensor is endowed with a product (MOD35) aimed to produce a reliable cloud mask in many pixels (confident classification in the terminology of MOD35). The algorithm underlying MOD35 is based on physical arguments, with a series of simple threshold tests mostly based on couples of spectral bands (see [24] for details of the algorithm). Troubles in dealing with the increasing number of spectral bands of current and next generation instrumentation from the physical point of view is fostering investigation of statistical methods for detecting clouds. Due to the very different spectral characteristics of water and land pixels, two separate independent classifications are performed for the two cases. Over land the SEVIRI data set is composed of 36415 cloudy pixels and 61361 clear ones; for water pixels we have 47048 cloudy pixels and 53610 clear ones. We assume that labels assigned by MOD35 are the truth.

In order to evaluate the methods, for each case (land and water) we divide the data set randomly into two parts; a training set of about 2/3 of the pixels used for estimation and learning (training set) and a test set of about 1/3 of the pixels used for evaluating the prediction capability of the estimated discrimination. The split was done 50 times in such a way that the proportion of clear and cloudy pixels of the whole original data set was respected. As in the density estimation experiments, the training dataset has been split in two parts $\mathcal{D}_1$ and $\mathcal{D}_2$ of equal size to apply the aggregate estimator involved in noisy IFA. The results are summarized in the Figure 6 showing the boxplots of misclassification errors for the various classification methods over 50 random splits for land (left) and water (right). For the land pixels, apart the NPDA method which has a poor behavior, none of the other three methods clearly stands out and they all perform essentially well. For the water panels (cf. the right panel of Figure 6) we get different conclusions. Here the boxplots clearly indicate that our noisy IFA classification method has the smallest error. Finally, Figure 7 shows the cloud mask overimposed to the analyzed area.

Table 3

Computational time taken by LDA, NPDA, FDA, NoisyIFA methodologies to classify the image of the cloud experiment (land pixels)

Method	CPU time (s)
LDA	0.4
NPDA	2.9
FDA	265
Noisy IFA - recursive	314

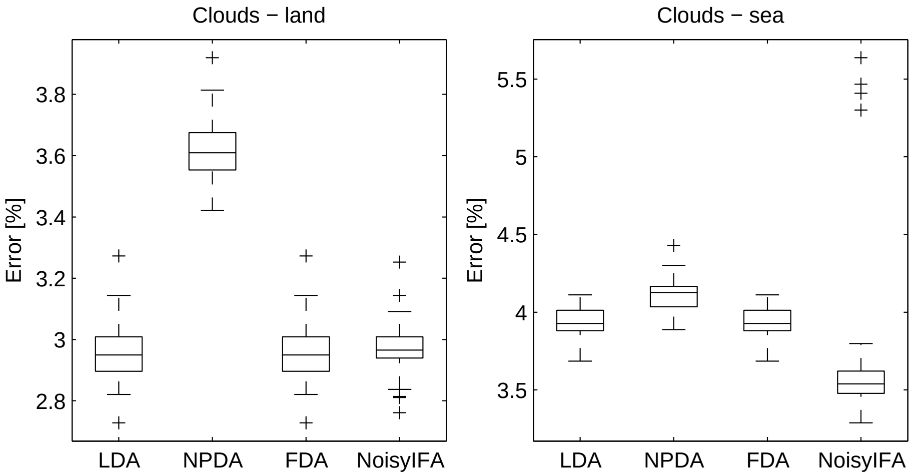

Fig 6. *Boxplot of the misclassifications for the considered classifiers. Results refer to land (left) and water (right) pixels of the remote sensing data.*

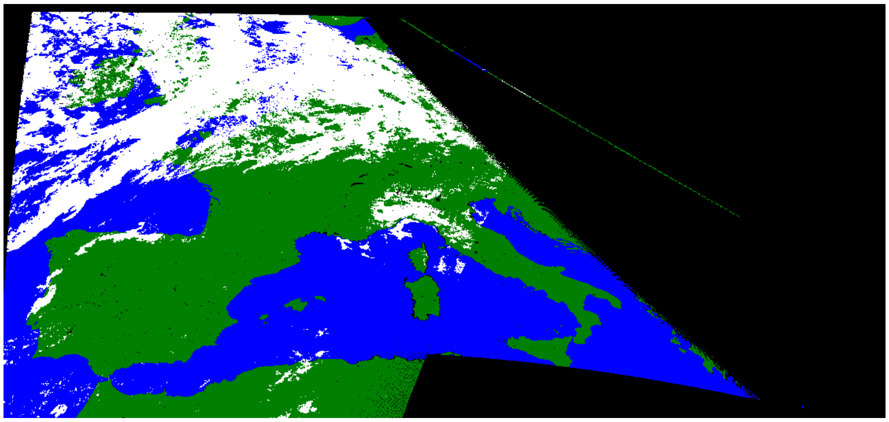

Fig 7. *Cloud mask estimated over a part of the region in Fig. 5 by Noisy IFA. Black: area not subject to classification; blue: pixels over water classified as clear; green: pixels over land classified as clear; white: pixels over land or water classified as cloudy.*

6. Conclusions

We have considered multivariate density estimation with dimensionality reduction expressed in terms of noisy independent factor analysis (IFA) model. In this model the data are generated by a (small) number of latent independent components having unknown distributions and observed in Gaussian noise.

Without assuming that either the number of components or the mixing matrix are known, we have shown that the densities of this form can be estimated with a fast rate. Using the mirror averaging aggregation algorithm, we constructed a density estimator which achieves a nearly parametric rate $\log^{1/4} n/\sqrt{n}$, independent of the dimension of the data.

We then applied these density estimates to construct nonparametric plug-in classifiers and have shown that they achieve, within a logarithmic factor independent of d, the best obtainable rate of the excess Bayes risk.

These theoretical results were supported by numerical simulations and by an application to a complex data set from a remote sensing experiment in which our IFA classifier outperformed several commonly used classification methods. Implementation of the IFA-based density estimator and of the related classifier is computationally intensive; therefore an efficient computational algorithm has been developed that makes mirror averaging aggregation feasible from computational point of view.

Appendix A: Proofs

Proof of (2.3). Note that (2.2) implies that the Fourier transform $\varphi_{\mathbf{X}}(\mathbf{u}) = \int_{\mathbb{R}^d} p_{\mathbf{X}}(\mathbf{x}) e^{i\mathbf{x}^T\mathbf{u}} d\mathbf{x}$ of the density $p_{\mathbf{X}}$ satisfies the inequality

$$|\varphi_{\mathbf{X}}(\mathbf{u})| \leq e^{-\sigma^2\|\mathbf{u}\|^2/2} \tag{A.1}$$

for all $\mathbf{u} \in \mathbb{R}^d$, where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^d$. Define the kernel estimator

$$\hat{p}_n^*(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right)$$

with the kernel $K : \mathbb{R}^d \to \mathbb{R}$, such that $K(\mathbf{x}) = \prod_{k=1}^d K_0(x_k)$, $\mathbf{x}^T = (x_1, x_2, \ldots, x_d)$, where K_0 is the sinc kernel: $K_0(x) = \frac{\sin x}{\pi x}$, for $x \neq 0$, and $K(0) = 1/\pi$, with the Fourier transform $\Phi^{K_0}(t) = I(|t| \leq 1)$.

Using Plancherel theorem and acting as in Theorem 1.4 on p. 21 of [33], we have

$$\begin{aligned} \mathbb{E}\|\hat{p}_n^* - p_{\mathbf{X}}\|_2^2 &= \frac{1}{(2\pi)^d} E \int |\varphi_n(\mathbf{u})\Phi^K(h\mathbf{u}) - \varphi_{\mathbf{X}}(\mathbf{u})|^2 d\mathbf{u} \\ &\leq \frac{1}{(2\pi)^d} \left[\int |1 - \Phi^K(h\mathbf{u})|^2 |\varphi_{\mathbf{X}}(\mathbf{u})|^2 d\mathbf{u} + \frac{1}{n} \int |\Phi^K(h\mathbf{u})|^2 d\mathbf{u} \right], \end{aligned}$$

where $\varphi_n(\mathbf{u}) = n^{-1} \sum_{j=1}^n e^{i\mathbf{X}_j^T\mathbf{u}}$ is the empirical characteristic function and $\Phi^K(\mathbf{v})$ is the Fourier transform of K. Note that $\Phi^K(\mathbf{v}) = \prod_{j=1}^d I\{|v_j| \leq 1\}$

where v_j are the components of $\mathbf{v} \in \mathbb{R}^d$. Now, for the bias term we have, using (A.1),

$$\begin{aligned}\int |1-\Phi^K(h\mathbf{u})|^2 |\varphi_{\mathbf{X}}(\mathbf{u})|^2 d\mathbf{u} &= \int I\left\{\exists j : |u_j| > \frac{1}{h}\right\} |\varphi_{\mathbf{X}}(\mathbf{u})|^2 d\mathbf{u} \\ &\le \int I\left\{\exists j : |u_j| > \frac{1}{h}\right\} e^{-\sigma^2\mathbf{u}^2/4} e^{-\sigma^2\mathbf{u}^2/4} d\mathbf{u} \\ &\le e^{-\sigma^2/4h^2} \int e^{-\sigma^2\mathbf{u}^2/4} d\mathbf{u} = e^{-\sigma^2/4h^2} \left(\frac{4\pi}{\sigma^2}\right)^{d/2}.\end{aligned}$$

Next, the variance term

$$\frac{1}{n}\int |\Phi^K(h\mathbf{u})|^2 d\mathbf{u} = \frac{1}{n}\prod_{j=1}^d \int I\left\{|u_j| \le \frac{1}{h}\right\} du_j = \frac{2^d}{nh^d}.$$

Combining the last two expressions, we get

$$\mathbb{E}\|\hat{p}_n^* - p_{\mathbf{X}}\|_2^2 \le C\left(e^{-\sigma^2/4h^2} + \frac{1}{nh^d}\right)$$

with some constant $C > 0$. Taking here $h = \sigma(4\log n)^{-1/2}$, we get (2.3). □

Proof of Proposition 2.1. W.l.o.g. we will suppose here that $\mathbf{a}_k$ are the canonical basis vectors in $\mathbb{R}^d$. Note first that the proof of (2.3) with $d = 1$ implies that the estimators (2.5) achieve the convergence rate of $(\log n)^{1/2}/n$ for the quadratic risk:

$$\mathbb{E}\|\hat{g}_k - g_k\|_2^2 = \mathcal{O}((\log n)^{1/2}/n) \quad \forall k = 1,\dots,m. \tag{A.2}$$

Denoting $C > 0$ a constant, not always the same, we have for the estimator (2.6)

$$\begin{aligned}\mathbb{E}\|\hat{p}_{n,m,A} - p_{\mathbf{X}}\|_2^2 &\le C\mathbb{E}\left\|\prod_{j=1}^m \hat{g}_j - \prod_{j=1}^m g_j\right\|_2^2 \\ &= C\mathbb{E}\left[\left\|\sum_{k=1}^m \prod_{j=1}^{k-1} g_j(\hat{g}_k - g_k) \prod_{j=k+1}^m \hat{g}_j\right\|_2^2\right] \\ &\le C\sum_{k=1}^m \mathbb{E}\left[\left\|\prod_{j=1}^{k-1} g_j\right\|_2^2 \|\hat{g}_k - g_k\|_2^2 \left\|\prod_{j=k+1}^m \hat{g}_j\right\|_2^2\right] \\ &\le C\sum_{k=1}^m \prod_{j=1}^{k-1} \|g_j\|_2^2 \mathbb{E}\left[\|\hat{g}_k - g_k\|_2^2 \prod_{j=k+1}^m \|\hat{g}_j\|_2^2\right] \\ &\le C\max_{k=1}^m \mathbb{E}\big[\|\hat{g}_k - g_k\|_2^2 \prod_{j=k+1}^m \|\hat{g}_j\|_2^2\big],\end{aligned}$$

where $\prod_{i=l}^{u} a_i = 1$ when $l > u$ and we have used that the L_2-norms of g_j are bounded for all $j = 1, \dots, m$. The latter is due to the fact that $\|g_j\|_2 \le \|\phi_{1,\sigma^2}\|_2$. The equality in the first line of the last display is just a telescopic sum while the inequality in its second line uses that the arguments of g_j and $\hat{g}_j$ are $\mathbf{a}_k^T \mathbf{x}$, and the integral factorizes because of the orthonormality of $\mathbf{a}_k$'s.

We now evaluate the L_2-norms of $\hat{g}_j$. By separating the diagonal and off-diagonal terms,

$$\|\hat{g}_j\|_2^2 = \frac{1}{nh} \int K_0^2 + \frac{1}{n^2} \sum_{i \neq m} \frac{1}{h} K^* \left(\frac{Y_i - Y_m}{h} \right), \tag{A.3}$$

with the convolution kernel $K^* = K_0 * K_0$ and we write for brevity $Y_i = \mathbf{a}_j^T X_i$. The second term in (A.3) is a U-statistic that we will further denote by U_n. Since all the summands $\frac{1}{h} K^* \left(\frac{Y_i - Y_m}{h} \right)$ in U_n are uniformly $\le C/h$, by Hoeffding inequality for U-statistics [15] we get

$$P(|U_n - E(U_n)| > t) \le 2 \exp(-cnh^2 t^2) \tag{A.4}$$

for some constant $c > 0$ independent of n. On the other hand, it is straightforward to see that there exists a constant C_0 such that $|E(U_n)| \le C_0$. This and (A.4) imply:

$$P(|U_n| > 2C_0) \le 2 \exp(-c' nh^2) \tag{A.5}$$

for some constant $c' > 0$ independent of n. From (A.3) and (A.5) we get

$$P(\mathcal{A}) \le 2d \exp(-c' nh^2), \tag{A.6}$$

for the random event $\mathcal{A} = \{\exists j : \|\hat{g}_j\|_2^2 \ge C_1\}$, where $C_1 = 2C_0 + \int K_0^2/(nh)$.

Using (A.6), (A.2) and the fact that $\|g_j\|_2^2$ and $\|\hat{g}_j\|_2^2$ are uniformly $\le C/h$ we find

$$\begin{aligned}
\mathbb{E}\left[\|\hat{g}_k - g_k\|_2^2 \prod_{j=k+1}^{m} \|\hat{g}_j\|_2^2 \right] &\le \mathbb{E}\left[\|\hat{g}_k - g_k\|_2^2 \prod_{j=k+1}^{m} \|\hat{g}_j\|_2^2 I\{\mathcal{A}\} \right] \\
&\quad + (C_1)^{m-k} \mathbb{E}\left[\|\hat{g}_k - g_k\|_2^2 I\{\mathcal{A}^c\} \right] \\
&\le (C/h)^{m-k+1} P\{\mathcal{A}\} + C(\log n)^{1/2}/n \\
&\le Ch^{-(m-k+1)} \exp(-c' nh^2) + C(\log n)^{1/2}/n \\
&\le C(\log n)^{1/2}/n.
\end{aligned}$$

Thus, the proposition follows. □

Proof of (2.12). We will show first that for some constant $C > 0$ and for all $j = 1, \dots, M$

$$\mathbb{P}(\|\hat{g}_j\|_{\infty,[-1,1]} > C) \le \frac{1}{n^{1/2} h^{3/2}}, \tag{A.7}$$

where $\|f\|_{\infty,[-1,1]} = \sup_{t \in [-1,1]} |f(t)|$ for $f : \mathbb{R} \to \mathbb{R}$. Note that the sinc kernel K_0 satisfies the inequality $|K_0(u)| \le 1/\pi$ for all $u \in \mathbb{R}$. Now because

$$\|\hat{g}_j\|_{\infty,[-1,1]} \le \mathbb{E}\|\hat{g}_j\|_{\infty,[-1,1]} + \|\hat{g}_j - \mathbb{E}\hat{g}_j\|_{\infty,[-1,1]}$$

and

$$|\mathbb{E}\hat{g}_j(t)| = \left|\int K_0(u)g_j(t-uh)du\right| \leq \frac{1}{\pi}, \quad \forall t \in \mathbb{R},$$

we have

$$\mathbb{P}(\|\hat{g}_j\|_{\infty,[-1,1]} > C) \leq \mathbb{P}\left(\|\hat{g}_j - \mathbb{E}\hat{g}_j\|_{\infty,[-1,1]} > C - \frac{1}{\pi}\right). \tag{A.8}$$

Now for $\eta(t) := \hat{g}_j(t) - \mathbb{E}\hat{g}_j(t)$ we have

$$\begin{aligned}\mathbb{E}(\eta(t+\Delta)-\eta(t))^2 =& \frac{1}{nh^2}\mathrm{Var}\left(K_0\left(\frac{t+\Delta-Z}{h}\right) - K_0\left(\frac{t-Z}{h}\right)\right)\\ \leq& \frac{1}{nh^2}\int\left(K_0\left(\frac{t+\Delta-z}{h}\right) - K_0\left(\frac{t-z}{h}\right)\right)^2 g_k(z)dz\\ \leq& \frac{C_0^2}{nh^3}\Delta^2\end{aligned} \tag{A.9}$$

for $t, \Delta \in [-1,1]$, where we used that $|K_0'(u)| \leq C_0$ with some constant C_0 for all $u \in \mathbb{R}$. Also, the standard bound for the variance of kernel estimator $\hat{g}_j$ gives

$$\mathbb{E}\eta^2(t) \leq \frac{C_2}{nh}, \quad \forall t \in [-1,1] \tag{A.10}$$

with $C_2 = \int K_0^2(u)du$. Now (A.9) and (A.10) verify conditions of the following lemma.

Lemma A.1. *[17, Appendix 1] Let $\eta(t)$ be a continuous real-valued random function defined on $\mathbb{R}^d$ such that, for some $0 < H < \infty$ and $d < a < \infty$ we have*

$$\begin{aligned}&\mathbb{E}|\eta(t+\Delta)-\eta(t)|^a \leq H\|\Delta\|^a, \qquad \forall\ t, \Delta \in \mathbb{R}^d,\\ &\mathbb{E}|\eta(t)|^a \leq H, \qquad \forall\ t \in \mathbb{R}^d.\end{aligned}$$

Then for every $\delta > 0$ and $t_0 \in \mathbb{R}^d$ such that $\|t_0\| \leq D$,

$$\mathbb{E}\left[\sup_{t:\|t-t_0\|\leq\delta}|\eta(t)-\eta(t_0)|\right] \leq B_0(D+\delta)^d H^{1/a}\delta^{1-d/a}$$

where B_0 is a finite constant depending only on a and d.

Applying this lemma with $d = 1$, $a = 2$, $H = \frac{C_0^2}{nh^3}$, $t_0 = 0$, and $\delta = 1$, we get

$$\mathbb{E}\sup_{t\in[-1,1]}|\eta(t)| \leq \mathbb{E}\sup_{t\in[-1,1]}|\eta(t)-\eta(0)| + \mathbb{E}|\eta(0)| \leq \frac{C_3}{n^{1/2}h^{3/2}} + \frac{C_2^{1/2}}{(nh)^{1/2}} \leq \frac{C_4}{n^{1/2}h^{3/2}}.$$

Applying now in (A.8) Markov inequality and choosing $C = C_4 + 1/\pi$, we obtain (A.7).

Next, assume w.l.o.g. that B is the unit ball in $\mathbb{R}^d$. We note that (A.7) implies

$$\begin{aligned}
\mathbb{P}\left(\left\|\prod_{j=1}^{k}\hat{g}_j\right\|_{\infty,B} > C^k\right) &\leq \mathbb{P}\left(\prod_{j=1}^{k}\|\hat{g}_j\|_{\infty,[-1,1]} > C^k\right)\\
&\leq \mathbb{P}(\cup_{j=1}^{k}\{\|\hat{g}_j\|_{\infty,[-1,1]} > C\}) \leq \frac{k}{n^{1/2}h^{3/2}}.
\end{aligned}$$

Using this and definition (2.7) of $\hat{p}_k$ we have that

$$\begin{aligned}
\mathbb{E}\|\hat{p}_k\|_{\infty,B} &\leq (2\pi\sigma^2)^{(d-k)/2}\mathbb{E}\left\|\prod_{j=1}^{k}\hat{g}_j\right\|_{\infty,B}\\
&\leq (2\pi\sigma^2)^{(d-k)/2}\left[C^k + \mathbb{E}\left\|\prod_{j=1}^{k}\hat{g}_j\right\|_{\infty,B} I\left\{\left\|\prod_{j=1}^{k}\hat{g}_j\right\|_{\infty,B} > C^k\right\}\right]\\
&\leq (2\pi\sigma^2)^{(d-k)/2}\left[C^k + \frac{1}{(\pi h)^k}\frac{k}{n^{1/2}h^{3/2}}\right],
\end{aligned}$$

where we also used the fact that $\|\hat{g}_j\|_{\infty,[-1,1]} \leq (\pi h)^{-1}$ for all $j = 1, \ldots, k$. Since $h \asymp (\log n)^{-1/2}$, we get that, for some constant L_k,

$$\mathbb{E}\|\hat{p}_k\|_{\infty,B} \leq L_k, \quad \forall k = 1, \ldots, M,$$

and (2.12) follows with $L' = \max(L_1, L_2, \ldots, L_M)$. □

In the proof of the theorem below, we make use of the following lemma.

Lemma A.2. *[21, Lemma A.1] Let Σ and D be two symmetric $d \times d$ matrices. For an arbitrary symmetric $d \times d$ matrix C, denote by $\lambda_1(C) \geq \lambda_2(C) \geq \cdots \geq \lambda_d(C)$ its d eigenvalues and by $\mathbf{a}_1(C), \mathbf{a}_2(C), \ldots, \mathbf{a}_d(C)$ the corresponding orthonormal eigenvectors. If $\lambda_r(\Sigma)$ is not a multiple eigenvalue of Σ (i.e. $\lambda_{r-1} > \lambda_r > \lambda_{r+1}$), then*

$$\mathbf{e}_r(\Sigma + D) - \mathbf{e}_r(\Sigma) = -S_r(\Sigma)D\mathbf{e}_r(\Sigma) + R,$$

where $S_r(\Sigma) := \sum_{s\neq r}\frac{1}{\lambda_s-\lambda_r}P_s(\Sigma)$, $P_s(\Sigma)$ is the projector on the eigenspace corresponding to the eigenvalue $\lambda_s = \lambda_s(\Sigma)$, and

$$\|R\| \leq \frac{6\|D\|_F}{\min_{s,s\neq r}|\lambda_s - \lambda_r|^2}$$

where $\|R\|$ is the Euclidean norm of R.

Proof of Theorem 2.1. To prove the theorem we use Corollary 5.7 in [19], which implies that for $\beta = 12\hat{L}$ the corresponding aggregate estimator $\tilde{p}_n$ satisfies:

$$\mathbb{E}_{\mathcal{D}_2}\|\tilde{p}_n - p_{\mathbf{X}}\|_2^2 \leq \min_{k=1,\ldots,M}\|\hat{p}_{n_1,k,\hat{B}_k} - p_{\mathbf{X}}\|_2^2 + \frac{\beta\log M}{n_2}, \tag{A.11}$$

where $\mathbb{E}_{\mathcal{D}_2}$ denotes the expectation over the second, aggregating subsample. Here $\hat{p}_{n_1,k,\hat{B}_k}$ are the estimators constructed from the first, training subsample $\mathcal{D}_1$, which is supposed to be frozen when applying the result of [19] and the inequality holds for any fixed training subsample. Taking expectation in inequality (A.11) with respect to the training subsample, using that, by construction, $\tilde{p}_n$ and $\hat{p}_{n_1,k,\hat{B}_k}$ vanish outside B, and interchanging the expectation and the minimum on the right hand side we get

$$\mathbb{E}\|\tilde{p}_n - p_{\mathbf{X}}\|_{2,B}^2 \le \min_{k=1,\dots,M} \mathbb{E}\|\hat{p}_{n_1,k,\hat{B}_k} - p_{\mathbf{X}}\|_{2,B}^2 + \frac{\log M}{n_2}\mathbb{E}\beta,$$

where now $\mathbb{E}$ is the expectation over the entire sample.

Recalling now that $M < d$, $n_2 = [cn/\sqrt{\log n}]$, and that $\mathbb{E}\beta \le C$ by (2.12), we obtain

$$\mathbb{E}\|\tilde{p}_n - p_{\mathbf{X}}\|_{2,B}^2 \le \min_{k=1,\dots,M} \mathbb{E}\|\hat{p}_{n_1,k,\hat{B}_k} - p_{\mathbf{X}}\|_{2,B}^2 + \frac{C(\log n)^{1/2}}{n}. \tag{A.12}$$

Now,

$$\min_{k=1,\dots,M} \mathbb{E}\|\hat{p}_{n_1,k,\hat{B}_k} - p_{\mathbf{X}}\|_{2,B}^2 \le \mathbb{E}\|\hat{p}_{m,\hat{A}} - p_{\mathbf{X}}\|_{2,B}^2, \tag{A.13}$$

where we set $\hat{A} = \hat{B}_m$ with m being the true rank of A. We set for brevity $\hat{p}_{m,D} \equiv \hat{p}_{n_1,m,D}$ for any $d \times m$ matrix D. In view of Lemma A.2 and Assumption 3, each of the columns of $\hat{A}$ estimates $\sqrt{n}$-consistently some column of A (cf. remarks in Section 2). Note also that $\hat{p}_{m,D}$ is invariant under permutation of columns of D. Thus, w.l.o.g. we will consider in the sequel that the columns $\hat{\mathbf{a}}_j$ of $\hat{A}$ in the expression for $\hat{p}_{m,\hat{A}}$ are numbered in the same order as the corresponding columns $\mathbf{a}_j$ of A.

Since $p_{\mathbf{X}} = p_{m,A}$, we have

$$\|\hat{p}_{m,\hat{A}} - p_{\mathbf{X}}\|_{2,B}^2 \le 2(\|\hat{p}_{m,\hat{A}} - \hat{p}_{m,A}\|_{2,B}^2 + \|\hat{p}_{m,A} - p_{m,A}\|_{2,B}^2). \tag{A.14}$$

Since $n_1 = n(1+o(1))$, by Proposition 2.1 we get

$$\mathbb{E}\|\hat{p}_{m,A} - p_{m,A}\|_{2,B}^2 = \mathcal{O}((\log n)^{1/2}/n). \tag{A.15}$$

It remains to prove that

$$\mathbb{E}\|\hat{p}_{m,\hat{A}} - \hat{p}_{m,A}\|_{2,B}^2 = \mathcal{O}((\log n)^{1/2}/n). \tag{A.16}$$

Setting for brevity $G_{\mathbf{x}}(A) = \left(\frac{1}{2\pi\sigma^2}\right)^{(d-m)/2} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{x}^T(\mathbf{I}_d - AA^T)\mathbf{x}\right\}$ we can write (see (2.6) and (2.7)),

$$\|\hat{p}_{m,\hat{A}} - \hat{p}_{m,A}\|_{2,B} = \|G_{\mathbf{x}}(\hat{A})\prod_{j=1}^m \hat{g}_j(\hat{\mathbf{a}}_j^T\mathbf{x}) - G_{\mathbf{x}}(A)\prod_{j=1}^m \hat{g}_j(\mathbf{a}_j^T\mathbf{x})\|_{2,B}$$

$$\le C\|\prod_{j=1}^m \hat{g}_j(\hat{\mathbf{a}}_j^T\mathbf{x}) - \prod_{j=1}^m g_j(\hat{\mathbf{a}}_j^T\mathbf{x})\|_{2,B} + C\|\prod_{j=1}^m \hat{g}_j(\mathbf{a}_j^T\mathbf{x}) - \prod_{j=1}^m g_j(\mathbf{a}_j^T\mathbf{x})\|_{2,B} +$$

$$\|G_{\mathbf{x}}(\hat{A})\prod_{j=1}^m g_j(\hat{\mathbf{a}}_j^T\mathbf{x}) - G_{\mathbf{x}}(A)\prod_{j=1}^m g_j(\mathbf{a}_j^T\mathbf{x})\|_{2,B} =: I_1 + I_2 + I_3.$$

As in the proof of Proposition 2.1 we get $\mathbb{E}I_i^2 = \mathcal{O}((\log n)^{1/2}/n)$, $i = 1, 2$. Next, we show that $\mathbb{E}I_3^2 = \mathcal{O}(1/n)$. We write $I_3 \leq I_{3,1} + I_{3,2}$ where

$$I_{3,1} = \|G_{\mathbf{x}}(\hat{A}) - G_{\mathbf{x}}(A)\|_{2,B} \| \prod_{j=1}^{m} g_j(\mathbf{a}_j^T\mathbf{x})\|_{2,B},$$

$$I_{3,2} = C\| \prod_{j=1}^{m} g_j(\hat{\mathbf{a}}_j^T\mathbf{x}) - \prod_{j=1}^{m} g_j(\mathbf{a}_j^T\mathbf{x})\|_{2,B}.$$

To bound these terms we will use the fact that $\| \prod_{j=k}^{l} g_j(\mathbf{a}_j^T\mathbf{x})\|_{2,B} \leq C$ for all $1 \leq k \leq l \leq m$ (and the same with $\hat{\mathbf{a}}_j$ instead of $\mathbf{a}_j$). This fact, the definition of $G_{\mathbf{x}}(\cdot)$ and the boundedness of the Frobenius norms of A and $\hat{A}$ imply that $I_{3,1} \leq C\|A - \hat{A}\|_F$, where $\|M\|_F$ denotes the Frobenius norm of matrix M. Now, from Lemma A.2 and Assumption 3 we get $\mathbb{E}\|\hat{A} - A\|_F^2 = \mathcal{O}(1/n)$. Thus, $\mathbb{E}I_{3,1}^2 = \mathcal{O}(1/n)$. We also get $\mathbb{E}I_{3,2}^2 = \mathcal{O}(1/n)$. This follows from the Lipschitz continuity of $g_j(\cdot)$ and from the fact that (cf. proof of Proposition 2.1):

$$\mathbb{E}I_{3,2}^2 \leq C\sum_{k=1}^{m} \mathbb{E}\left[\left\| \prod_{j=1}^{k-1} g_j(\mathbf{a}_j^T\mathbf{x})\right\|_{2,B}^2 \|g_k(\mathbf{a}_k^T\mathbf{x}) - g_k(\hat{\mathbf{a}}_k^T\mathbf{x})\|_{2,B}^2 \left\| \prod_{j=k+1}^{m} g_j(\hat{\mathbf{a}}_j^T\mathbf{x})\right\|_{2,B}^2\right]$$

So, we have $\mathbb{E}I_3^2 = \mathcal{O}(1/n)$. This finishes the proof of (A.16).

Inequalities (A.14), (A.15), and (A.16) give

$$\mathbb{E}\|\hat{p}_{m,\hat{A}} - p_{\mathbf{x}}\|_{2,B}^2 \leq \mathcal{O}((\log n)^{1/2}/n),$$

which together with (A.12) and (A.13) completes the proof. □

Proof of Proposition 3.1. For any classifier T we have

$$\begin{aligned} R_B(T) - R_B(T^*) &= \sum_{j=1}^{J} \pi_j \int_B (I(T(\mathbf{x}) \neq j) - I(T^*(\mathbf{x}) \neq j)) f_j(\mathbf{x}) d\mathbf{x} \\ &= \sum_{j=1}^{J} \pi_j \int_B (I(T^*(\mathbf{x}) = j) - I(T(\mathbf{x}) = j)) f_j(\mathbf{x}) d\mathbf{x} \\ &= \int_B (\pi_{T^*(\mathbf{x})} f_{T^*(\mathbf{x})}(\mathbf{x}) - \pi_{T(\mathbf{x})} f_{T(\mathbf{x})}(\mathbf{x})) d\mathbf{x}. \end{aligned}$$

Therefore, the excess risk of the plug-in classifier $\hat{T}$ can be written in the form

$$\begin{aligned} \mathcal{E}(\hat{T}) &\equiv \mathbb{E}(R_B(\hat{T})) - R_B(T^*) \\ &= \mathbb{E}\int_B (\pi_{T^*} f_{T^*}(\mathbf{x}) - \pi_{\hat{T}} \hat{f}_{\hat{T}}(\mathbf{x}) + \pi_{\hat{T}} \hat{f}_{\hat{T}}(\mathbf{x}) - \pi_{\hat{T}} f_{\hat{T}}(\mathbf{x})) d\mathbf{x}, \quad \text{(A.17)} \end{aligned}$$

where we omit for brevity the argument $\mathbf{x}$ of T^* and $\hat{T}$. Note that, by the definition of $\hat{T}$, for all $\mathbf{x} \in \mathbb{R}^d$ we have:

$$\begin{aligned}
&\pi_{T^*} f_{T^*}(\mathbf{x}) - \pi_{\hat{T}} \hat{f}_{\hat{T}}(\mathbf{x}) + \pi_{\hat{T}} \hat{f}_{\hat{T}}(\mathbf{x}) - \pi_{\hat{T}} f_{\hat{T}}(\mathbf{x}) \\
&\quad \le \pi_{T^*} f_{T^*}(\mathbf{x}) - \pi_{T^*} \hat{f}_{T^*}(\mathbf{x}) + \pi_{\hat{T}} |\hat{f}_{\hat{T}}(\mathbf{x}) - f_{\hat{T}}(\mathbf{x})| \\
&\quad \le \sum_{j=1}^{J} \pi_j |\hat{f}_j(\mathbf{x}) - f_j(\mathbf{x})|.
\end{aligned}$$

Combining the last display with (A.17) proves the proposition. □

References

[1] Aladjem, M. (2005). Projection Pursuit Mixture Density Estimation. *IEEE Trans. Signal Process.* **53** 4376–4383. MR2242178

[2] Amato, U., Antoniadis, A., and Grégoire, G. (2003). Independent Component Discriminant Analysis. *Int. J. Math.* **3** 735–753. MR1975044

[3] Anderson, T. W., and Rubin, H. (1956). Statistical inference in factor analysis. *Proc. Third Berkeley Symposium on Mathematical Statistics and Probability* (Vol. V), ed. J. Neyman. Berkeley and Los Angeles, University of California Press, 111–150. MR0084943

[4] An, Y., Hu, X., and Xu, L. (2006). A comparative investigation on model selection in independent factor analysis. *J. Math. Modeling Algorithms* **5** 447–473. MR2244277

[5] Artiles, L. M. (2001). *Adaptive minimax estimation in classes of smooth functions.* University of Utrecht, Ph.D. thesis.

[6] Attias, H. (1999). Independent Factor Analysis. *Neural Computation* **11** 803–851.

[7] Audibert, J. U., and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *Annals Statist.* **35** 608–633. MR2336861

[8] Belitser, E., and Levit, B. (2001). Asymptotically local minimax estimation of infinitely smooth density with censored data. *Annals Inst. Statist. Math.* **53** 289–306. MR1841137

[9] Blanchard, B., Kawanabe, G. M., Sugiyama, M., Spokoiny, V., and Müller, K. R. (2006). In search of non-gaussian components of a high-dimensional distribution. *J. of Mach. Learn. Research* **7** 247–282. MR2274368

[10] Cook, R. D., and Li, B. (2002). Dimension reduction for conditional mean in regression. *Annals Statist.* **32** 455–474. MR1902895

[11] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* New York, Springer. MR1383093

[12] Glad, I. K., Hjort, N. L., and Ushakov, N.G. (2003). Correction of density estimators that are not densities. *Scand. J. Statist.* **30** 415–427. MR1983134

[13] Hall, P., and Murison, R. D. (1993). Correcting the negativity of high-order kernel density estimators. *J. Multivar. Analysis* **47** 103–122. MR1239108

[14] Hastie, T., Tibshirani, R., and Buja, A. (1994). Flexible Discriminant Analysis by Optimal Scoring. *J. Am. Statist. Assoc.* **89** 1255–1270. MR1310220

[15] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Am. Statist. Assoc.* **58** 13–30. MR0144363

[16] Hyvarinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis.* New York, Wiley and Sons.

[17] Ibragimov, I. A., and Khasminskiĭ, R. Z. (1982). An estimate of the density of a distribution belonging to a class of entire functions (Russian). *Teoriya Veroyatnostei i ee Primeneniya* **27** 514–524. MR0673923

[18] Juditsky, A. B., Nazin, A. V, Tsybakov, A. B., and Vayatis, N. (2005). Recursive Aggregation of Estimators by the Mirror Descent Algorithm with Averaging. *Problems Informat. Transmiss.* **41** 368–384.

[19] Juditsky, A., Rigollet, P., and Tsybakov, A. B. (2008). Learning by mirror averaging. *Annals Statist.* **36** 2183–2206. MR2458184

[20] Kawanabe, M., Sugiyama, M., Blanchard, G., and Müller, K. R. (2007). A new algorithm of non-Gaussian component analysis with radial kernel functions. *Annals Inst. Statist. Math.* **59** 57–75. MR2405287

[21] Kneip, A., and Utikal, K. (2001). Inference for density families using functional principal components analysis (with discussion). *J. Am. Statist. Assoc.* **96** 519–542. MR1946423

[22] McLachlan, G.J., and Peel, D. (2000). *Finite Mixture Models.* New York, Wiley. MR1789474

[23] Montanari, A., Calò, D., and Viroli, C. (2008). Independent factor discriminant analysis. *Comput. Statist. Data Anal.* **52** 3246–3254. MR2424789

[24] Platnick, S., King, M. D., Ackerman, S. A., Menzel, W. P, Baum, P. A., Ridi, J. C, and Frey, R. A. (2003). The MODIS cloud products: Algorithms and examples from Terra. *IEEE Trans. Geosc. Remote Sens.* **41** 459–473.

[25] Polzehl, J. (1995). Projection pursuit discriminant analysis. *Comput. Statist. Data Anal.* **20** 141–157. MR1353784

[26] Roweis, S., and Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290** 2323–2326.

[27] Samarov, A., and Tsybakov, A. B. (2004). Nonparametric independent component analysis. *Bernoulli* **10** 565–582. MR2076063

[28] Samarov, A., and Tsybakov, A. B. (2007). Aggregation of density estimators and dimension reduction. In *Advances in Statistical Modeling and Inference, Essays in Honor of K. Doksum*, Series in Biostatistics (Vol. 3), V. Nair (ed.). London, World Scientific 233–251. MR2416118

[29] Silverman, B. W. (1982). Kernel density estimation using the fast Fourier transform. *Appl. Statist.* **31** 93–99.

[30] Stewart, G. W., Sun, J. (1990), *Matrix Perturbation Theory*. New York, Academic Press.

[31] Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319–2323.

[32] Titterington, D., A. Smith, and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York, Wiley. MR0838090

[33] Tsybakov, A. B. (2009), *Introduction to Nonparametric Estimation*. New York, Springer.

[34] Wand, M. P., and Jones, M. C. (1995). *Kernel Smoothing*. London, Chapman & Hall/CRC. MR1319818

[35] Yang, Y. (1999). Minimax nonparametric classification. I. Rates of convergence. II. Model selection for adaptation. *IEEE Trans. Inform. Theory* **45** 2271–2292. MR1725115,1725116