Bernoulli **10**(4), 2004, 565–582

Nonparametric independent component analysis

ALEXANDER SAMAROV[1] and ALEXANDRE TSYBAKOV[2]

[1]*Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139-4307, USA. E-mail: samarov@mit.edu*
[2]*Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VI, 4 place Jussieu, Tour 56, 75252 Paris Cedex 5, France. E-mail: tsybakov@ccr.jussieu.fr*

We consider the problem of nonparametric estimation of a d-dimensional probability density and its 'principal directions' in the independent component analysis model. A new method of estimation based on diagonalization of nonparametric estimates of certain matrix functionals of the density is suggested. We show that the proposed estimators of principal directions are $\sqrt{n}$-consistent and that the corresponding density estimators converge at the optimal rate.

Keywords: estimation of functionals; independent component analysis; nonparametric density estimation; projection pursuit

1. Introduction

Let $X_1, \ldots, X_n$ be independent and identically distributed random vectors with common probability density p on $\mathbb{R}^d$, $d \geq 2$. We consider the problem of nonparametric estimation of the density p, assuming that it has the form

$$p(x) = |\det(B)| \prod_{j=1}^{d} p_j(x^{\mathrm{T}}\beta_j), \qquad x \in \mathbb{R}^d, \tag{1}$$

where $\beta_1, \ldots, \beta_d$ are unknown, linearly independent, unit-length column vectors in $\mathbb{R}^d$, $\det(B)$ is the determinant of the matrix

$$B = (\beta_1, \ldots, \beta_d),$$

and $p_j(\cdot)$, $j = 1, \ldots, d$, are unknown probability densities on $\mathbb{R}^1$. We assume that the p_j are smooth densities belonging to Hölder classes $\Sigma(s_j, L_j)$ with some $s_j > 2$, $L_j > 0$, respectively. Here the Hölder class $\Sigma(s_j, L_j)$ is defined as the class of all probability densities on $\mathbb{R}^1$ satisfying

$$|p_j^{(l_j)}(z) - p_j^{(l_j)}(z')| \leq L_j |z - z'|^{s_j - l_j}, \quad \forall z, z' \in \mathbb{R}^1,$$

where $l_j = \lfloor s_j \rfloor$ and $0 < s_j, L_j < \infty$.

Model (1) has recently become popular in the engineering literature in the context of independent components analysis (ICA); see Hyvärinen *et al*. (2001) and Roberts and

1350–7265

Everson (2001). ICA is a statistical and computational technique for identifying hidden factors that underlie sets of random variables, measurements, or signals. ICA defines a model for the observed multivariate data, in which the data variables are assumed to be linear mixtures of some unknown latent variables, called the independent components, and the mixing system is also unknown. The goal of ICA is to estimate these independent components, also called sources or factors. The data analysed by ICA include digital images, biomedical, financial and telecommunications data, as well as economic indicators and psychometric measurements. Typically, ICA is applied in blind source separation problems where the measurements are given as a set of parallel signals, for example, mixtures of simultaneous speech signals that have been picked up by several microphones, brain waves recorded by multiple sensors, interfering radio signals arriving at a mobile phone, or parallel time series obtained from some industrial process.

In the engineering literature model (1) is usually stated in an equivalent form,

$$U = B^{\mathrm{T}} X,$$

where X is a random vector in $\mathbb{R}^d$, U is a random vector in $\mathbb{R}^d$ with independent components, and B is an unknown non-degenerate matrix. The goal of ICA is to estimate the matrix B based on a sample $X_1, \ldots, X_n$ from X. Most known methods of solving the ICA problem involve specification of the parametric form of the latent component densities p_j and estimation of B together with parameters of p_j using maximum likelihood or minimization of the empirical versions of various divergence criteria between densities; see Hyvärinen *et al*. (2001), Roberts and Everson (2001) and references therein. Statistical properties of such methods are usually not analysed in this literature, but can be derived from the theory of minimum contrast parametric estimation. In practical applications, the distributions p_j of latent independent components are generally unknown, and it is preferable to consider ICA as a semi-parametric model in which these distributions are left unspecified.

Note that in the ICA literature it is usually assumed that the right-hand side of (1) represents not the true density p but only its best independent component approximation. One is then supposed to find the best approximation of p in a class of product densities, with respect to a given loss criterion. Accordingly, the problem of estimation of p is not considered. In a recent paper in this direction, Bach and Jordan (2002) propose an ICA algorithm which does not rely on a specific parametric form of component distribution. Among earlier methods of this kind we mention those of Pham (1996) and Cardoso (1999) because we believe that under appropriate conditions these methods provide consistent estimators of the directions β_j. However, the proof of this conjecture is not available.

Model (1) is a special case of the projection pursuit density estimation (PPDE) model. As in ICA, in PPDE one uses the approximation of the joint density by a product of univariate densities of rotated components, rather than the exact representation assumed in (1); see, for example, Huber (1985). The difference from ICA is that the number of univariate densities used to approximate p in PPDE is not necessarily equal to the dimension d; it can be larger or smaller. Using minimization of the Kullback–Liebler divergence method, Hall (1988) has shown that the direction vector of the one-component projection pursuit approximation can be estimated at $n^{-1/2}$ rate; extension of this result to multiple-component approximation as in the right-hand side of (1) is not obvious.

The aim of this paper is twofold. First, we propose a method of simultaneous estimation of the directions $\beta_1, \ldots, \beta_d$ which does not rely on parametric distributional assumptions on $p_j(\cdot)$. Our method is based on nonparametric estimation of the average outer product of the density gradient and on simultaneous diagonalization of this estimated matrix and the sample covariance matrix of the data. We show that our estimates converge to the true directions $\beta_1, \ldots, \beta_d$ with the parametric $n^{-1/2}$ rate. Second, unlike in the ICA framework, we also consider the estimation of p. In fact, we show that (1) is a particular form of multivariate density for which the 'curse of dimensionality' can be avoided in the asymptotic sense. We show that the component densities can be estimated at the usual one-dimensional nonparametric rate and that the resulting product estimator of the joint density $p(x)$ has the one-dimensional (optimal) nonparametric rate corresponding to the independent component density with the worst smoothness.

In Section 2 we outline our approach and define the estimators of independent component directions and of the density (1). Root-n consistency of the direction estimators is proved in Section 3, while the rate of convergence of the joint density estimators is established in Section 4. The Appendix contains proofs of our technical lemmas.

2. The method of estimation

We first outline the idea of our $\sqrt{n}$-consistent estimators of the directions β_j, $j = 1, \ldots, d$. Denote by X a random vector in $\mathbb{R}^d$ distributed with the density p. Under model (1) the components of the random vector $U = B^{\mathrm{T}}X$ are independent, so that the covariance matrix $\mathrm{var}(U) = D$ is diagonal, that is,

$$B^{\mathrm{T}} \mathrm{var}(X) B = D. \tag{2}$$

Consider also a matrix-valued functional

$$T(p) = \mathrm{E}[\nabla p(X) \nabla^{\mathrm{T}} p(X)], \tag{3}$$

where ∇p is the gradient of p and E denotes the expectation with respect to p. For densities satisfying (1) the functional $T(p)$ takes the form

$$T(p) = \sum_{j=1}^{d} \sum_{k=1}^{d} c_{jk} \beta_j \beta_k^{\mathrm{T}}, \tag{4}$$

where $c_{jk} = (\det(B))^2 \mathrm{E}[\prod_{i \neq j} p_i(X^{\mathrm{T}}\beta_i) \prod_{m \neq k} p_m(X^{\mathrm{T}}\beta_m) p_j'(X^{\mathrm{T}}\beta_j) p_k'(X^{\mathrm{T}}\beta_k)]$ (we assume that the c_{jk} are finite). Making the change of variables $u_l = X^{\mathrm{T}}\beta_l$, $l = 1, \ldots, d$, and integrating out the u_l with $l \neq j, k$, we obtain, under mild boundary assumptions on the marginal densities p_j,

$$c_{jk} = C_{jk} \int p_j'(u_j) p_k'(u_k) p_j^2(u_j) p_k^2(u_k) \mathrm{d}u_j \, \mathrm{d}u_k = 0, \tag{5}$$

for $j \neq k$ and some constants C_{jk}. Hence, (4) can be written as

$$T(p) = \sum_{j=1}^{d} c_{jj}\beta_j\beta_j^{\mathrm{T}},$$

or, in matrix form,

$$T = T(p) = BCB^{\mathrm{T}}, \tag{6}$$

where $C = \mathrm{diag}(c_{jj})$. It is easy to see that $c_{jj} > 0$ for all j. Thus T is positive definite and (6) implies that

$$B^{\mathrm{T}}T^{-1}B = C^{-1}, \tag{7}$$

Denote $\Sigma = \mathrm{var}(X)$, $P = BC^{1/2}$ and $\Lambda = C^{1/2}DC^{1/2}$. Then equations (2) and (7) imply that

$$P^{\mathrm{T}}\Sigma P = \Lambda, \tag{8}$$

$$P^{\mathrm{T}}T^{-1}P = I, \tag{9}$$

where I is the identity matrix. It is known from matrix algebra (see Lewis 1991, Section 6.7), that for any two positive semidefinite symmetric matrices Σ and T^{-1}, such that at least one of these matrices is positive definite, there exists a non-singular matrix P and a diagonal matrix Λ such that (8) and (9) hold, where the elements λ_j of Λ are eigenvalues of the matrix $T\Sigma$ and the columns of P are the corresponding eigenvectors of $T\Sigma$. So, if all the λ_j are different, the columns of P can be uniquely identified as vectors $\mathbf{p}_j$ solving

$$T\Sigma\mathbf{p}_j = \lambda_j\mathbf{p}_j, \qquad j = 1, \ldots, d.$$

This last equation can be rewritten as

$$T^{1/2}\Sigma T^{1/2}\mathbf{q}_j = \lambda_j\mathbf{q}_j, \qquad j = 1, \ldots, d, \tag{10}$$

where $T^{1/2}$ is the symmetric square root of T and $\mathbf{q}_j = T^{-1/2}\mathbf{p}_j$. Since the matrix $W = T^{1/2}\Sigma T^{1/2}$ is symmetric, the vectors $\mathbf{q}_j$, $\mathbf{q}_k$ are orthogonal for $j \neq k$.

Now our plan for estimating the β_j is as follows: we first construct $\sqrt{n}$-consistent estimators of matrices Σ and T and use them to $\sqrt{n}$-consistently estimate the matrix $W = T^{1/2}\Sigma T^{1/2}$, then find the principal components of this estimator and use them to estimate $\mathbf{q}_j$, $j = 1, \ldots, d$. These last estimates are then used to construct $\sqrt{n}$-consistent estimates of $\mathbf{p}_j = T^{1/2}\mathbf{q}_j$, and finally, since

$$\beta_j = c_{jj}^{-1/2}\mathbf{p}_j = \mathbf{p}_j/\|\mathbf{p}_j\|, \tag{11}$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^d$, the estimators of β_j are obtained from estimators of $\mathbf{p}_j$ by normalizing them to have unit length.

We now define our estimators of β_j more precisely. Denote

$$S = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^{\mathrm{T}},$$

the sample covariance matrix. Consider next an estimator of $T(p)$ defined by

$$\hat{T} = \frac{1}{n}\sum_{i=1}^{n} \nabla \hat{p}_{-i}(X_i)\nabla^{\mathrm{T}} \hat{p}_{-i}(X_i),$$

where $\nabla \hat{p}_{-i}(X_i)$ is the column vector with components

$$\frac{\partial \hat{p}_{-i}}{\partial x_l}(X_i) = \frac{1}{(n-1)h^{d+1}} \sum_{j=1, j\neq i}^{n} Q_l\left(\frac{X_j - X_i}{h}\right), \qquad l = 1, \ldots, d. \tag{12}$$

Here

$$Q_l\left(\frac{X_j - X_i}{h}\right) = K_1\left(\frac{X_{jl} - X_{il}}{h}\right) \prod_{k=1, k\neq l}^{d} K\left(\frac{X_{jk} - X_{ik}}{h}\right),$$

X_{ik} is the kth component of X_i, $h > 0$ is a bandwidth and $K : \mathbb{R}^1 \to \mathbb{R}^1$, $K_1 : \mathbb{R}^1 \to \mathbb{R}^1$ are kernels satisfying the conditions that will be stated below. If K is differentiable, one can take $K_1 = K'$; in this case (12) can be viewed as a partial derivative of the leave-one-out kernel density estimator

$$\hat{p}_{-i}(X_i) = \frac{1}{(n-1)h^d} \sum_{j=1, j\neq i}^{n} \prod_{k=1}^{d} K\left(\frac{X_{jk} - X_{ik}}{h}\right).$$

The symmetric non-negative definite matrix $\hat{T}$ admits a spectral decomposition $\hat{T} = \hat{V}\hat{M}\hat{V}^{\mathrm{T}}$ with an orthogonal matrix $\hat{V}$ and a diagonal matrix $\hat{M} = \mathrm{diag}(\hat{m}_1, \ldots, \hat{m}_d)$, where $\hat{m}_j \geqslant 0$, $j = 1, \ldots, d$. The square root matrix estimator $\hat{T}^{1/2}$ is then defined as $\hat{T}^{1/2} = \hat{V}\hat{M}^{1/2}\hat{V}^{\mathrm{T}}$.

We next compute the orthogonal eigenvectors $\hat{\mathbf{q}}_j$, $j = 1, \ldots, d$, of the symmetric matrix $\hat{W} = \hat{T}^{1/2} S \hat{T}^{1/2}$, and finally obtain our direction estimators $\hat{\beta}_j$ by normalizing vectors $\hat{T}^{1/2}\hat{\mathbf{q}}_j$ to have unit length.

Our final step is to estimate $p(x)$ by

$$\hat{p}(x) = |\det(\hat{B})| \prod_{j=1}^{d} \frac{1}{n\tilde{h}_j} \sum_{i=1}^{n} \tilde{K}\left(\frac{X_i^{\mathrm{T}}\hat{\beta}_j - x^{\mathrm{T}}\hat{\beta}_j}{\tilde{h}_j}\right), \tag{13}$$

where the matrix $\hat{B} = (\hat{\beta}_1, \ldots, \hat{\beta}_d)$, with vectors $\hat{\beta}_j$, $j = 1, \ldots, d$, constructed by the procedure described above, and the kernel $\tilde{K}(\cdot)$ and bandwidths $\tilde{h}_j$ satisfy the conditions that will be stated below.

Remark 1. The matrix-valued functional (4) can be written in the form

$$T(p) = \mathrm{E}[p^2(X)\nabla \log p(X) \nabla^{\mathrm{T}} \log p(X)].$$

Dropping the scalar weight function $p^2(X)$ under the expectation, we obtain the Fisher information matrix of the density p, $I(p) = \mathrm{E}[\nabla \log p(X) \nabla^{\mathrm{T}} \log p(X)]$. Under suitable conditions $I(p)$ satisfies the analogue of (6): $I(p) = B\tilde{C}B^{\mathrm{T}}$, where $\tilde{C}$ is a diagonal matrix. More generally, one can replace $T(p)$ by a functional of the form

$$\Phi(p) = \mathrm{E}[w(X)\nabla p(X) \nabla^{\mathrm{T}} p(X)], \tag{14}$$

where $w(X)$ is a scalar weight function. For example, one can take $w(x) = p^\gamma(x)$ for some power $\gamma > 0$, and our argument again carries through. Therefore, the above method of estimation of the directions β_j works with $I(p)$ or $\Phi(p)$ in place of $T(p)$, with suitably defined estimators of those functionals. We prefer to consider the functional $T(p)$ since it makes proofs less technical and allows us to obtain results under milder conditions.

Finally, the matrix Σ in (8) can be also replaced by a matrix of the form (14) since under mild conditions the latter is also transformed to a diagonal one by the action of B. Therefore, instead of simultaneously diagonalizing Σ and T^{-1} we can, in general, simultaneously diagonalize Φ_1 and Φ_2^{-1}, where Φ_1 and Φ_2 are matrix functionals of the form (14) with two distinct weight functions $w = w_1$ and $w = w_2$. An advantage of such a procedure is robustness to outliers: in general, no moment assumptions on X are needed to ensure the existence of Φ_1 and Φ_2.

Remark 2. In this paper we assume that the smoothness indices s_j are known: they enter in the definition of bandwidths of density estimators (see Theorem 2). In practice the bandwidths should be selected in a data-driven manner. Classical bandwidth selection methods cannot be applied here since a possibility of anisotropic smoothness s_j for different directions β_j should be taken into account. Using the estimators of the present paper as a building block, one can construct a density estimator that adapts to anisotropic smoothness using the ideas of aggregation of estimators; see, for example, Nemirovski (2000). This topic will be discussed in our forthcoming work.

Remark 3. The idea of using matrix average derivative functionals of the underlying functions to identify multivariate nonparametric models appeared earlier in regression context (Samarov 1993; see also Härdle and Tsybakov 1991). Samarov (1993) shows that, under certain regularity conditions, kernel plug-in estimators of a class of integral functionals are $\sqrt{n}$-consistent and asymptotically normal. While these results are applicable in our context, we give a separate proof of $\sqrt{n}$-consistency of $\hat{T}$ under weaker regularity conditions.

Remark 4. Our method can be applied to a generalization of ICA where the independent components are multivariate:

$$p(x) = |\det(B)| \prod_{j=1}^{k} p_j(B_j x), \qquad x \in \mathbb{R}^d,$$

with some unknown matrices B_j of dimension $d \times n_j$, and $B = (B_1, \ldots, B_k)$, $n_1 + \cdots + n_k = d$, $B_j^{\mathrm{T}} B_i = 0$ for $i \neq j$, such that B is of full rank. Here the functional $T(p)$ has a block-diagonal form, and we can estimate the subspaces corresponding to the matrices B_j, up to an arbitrary non-singular transformation.

3. Root-n consistency of the estimators $\hat{\beta}_j$

Assume that the bandwidth h and the kernels K and K_1 satisfy the following conditions for some $b > d + 3$.

Condition 1. *$nh^{2d+4} \to \infty$, $nh^{2b-2} \to 0$, as $n \to \infty$.*

Condition 2. *The kernels K and K_1 are bounded functions supported on $[-1, 1]$ and such that*

$$\int K(u)\mathrm{d}u = 1, \qquad \int u^l K(u)\mathrm{d}u = 0, \quad l = 1, \ldots, \lfloor b \rfloor,$$

$$\int uK_1(u)\mathrm{d}u = 1, \qquad \int u^l K_1(u)\mathrm{d}u = 0, \quad l = 0, 2, \ldots, \lfloor b \rfloor.$$

Note that there exist several ways of constructing the kernels satisfying Condition 2. One of them is to take $K(u) = \sum_{j=0}^{\lfloor b \rfloor} \phi_j(u)\phi_j(0)\mathbb{1}(|u| \leqslant 1)$, $K_1(u) = \sum_{j=0}^{\lfloor b \rfloor} \phi_j'(u)\phi_j(0)\mathbb{1}(|u| \leqslant 1)$, where $\{\phi_j\}_{j=0}^{\lfloor b \rfloor}$ are the first orthonormal Legendre polynomials on $[-1, 1]$. Here $\mathbb{1}(\cdot)$ denotes the indicator function.

Next, we introduce assumptions on the density p. The first is a usual smoothness assumption:

Condition 3. *The density p satisfies* (1)*, where the p_j are probability densities on $\mathbb{R}^1$ belonging to Hölder classes $\Sigma(s_j, L_j)$ with $b \leqslant s_j < \infty$ and $0 < L_j < \infty$, for $j = 1, \ldots, d$.*

It is easy to see (in view of standard embedding theorems) that Condition 3 implies uniform boundedness and continuity of all the p_j, as well as of all their derivatives up to order $\lfloor b \rfloor$, for $j = 1, \ldots, d$. This, in particular, implies that the diagonal elements c_{jj} of the matrix C are finite. These elements are always positive since the p_j are probability densities, thus T is positive definite provided that (5) holds.

The next assumption guarantees that (5) holds.

Condition 4.

$$\int p_j'(u)p_j^2(u)\mathrm{d}u = 0, \qquad j = 1, \ldots, d. \tag{15}$$

Note that Condition 4 is very mild: it is satisfied, for example, for any density p_j supported on $\mathbb{R}^1$ (or any symmetric density supported on a bounded subset of $\mathbb{R}^1$) such that the integral in (15) is well defined.

Condition 5. *The matrices T and $T^{1/2}\Sigma T^{1/2}$ do not have multiple eigenvalues.*

This condition, in particular, rules out the case where the densities p_j are normal: in fact,

it is easy to see that the matrix T in this case is proportional to Σ^{-1}. Note that the condition of non-normality (of all independent components except possibly one) is always imposed in the context of ICA, see, for example, Hyvärinen *et al*. (2001). It is important to emphasize that the non-normality is only a necessary condition for identifying independent components. Additional conditions are required to obtain root-n consistency of the estimators. For our method Condition 5 plays exactly this role. More precisely, it is needed in Lemma 2 below. If Condition 5 is not satisfied, any vectors in the subspaces spanned by the eigenvectors with equal eigenvalues can be chosen by our method. So, one cannot prove consistency of estimation of the β_j corresponding to those subspaces; one may only consider consistency of subspace estimation in the spirit of Li (1991), but this is beyond the scope of this paper.

Theorem 1. *Assume that Conditions 1–5 are satisfied and* $\mathrm{E}\|X\|^4 < \infty$. *Then*

$$\|\hat{\beta}_j - \beta_j\| = O_p(n^{-1/2}), \qquad j = 1, \ldots, d. \tag{16}$$

The proof of Theorem 1 is based on the following two lemmas. For a $d \times d$ matrix A, define

$$\|A\|_2 \stackrel{\text{def}}{=} \sup_{\|\mathbf{v}\|=1} \sqrt{\mathbf{v}^{\mathrm{T}} A^{\mathrm{T}} A \mathbf{v}}.$$

Lemma 1. *Assume that Conditions 1–3 are satisfied and* $\mathrm{E}\|X\|^4 < \infty$. *Then*

$$S - \Sigma = \frac{1}{\sqrt{n}} \Delta_\Sigma, \tag{17}$$

and

$$\hat{T} - T = \frac{1}{\sqrt{n}} \Delta_T, \tag{18}$$

where the matrices Δ_Σ *and* Δ_T *are such that* $\|\Delta_\Sigma\|_2 = O_p(1)$ *and* $\|\Delta_T\|_2 = O_p(1)$, *as* $n \to \infty$.

The proof of this lemma is given in the Appendix.

Lemma 2. *Let* A *and* Δ_A *be symmetric* $d \times d$ *matrices such that* A *has distinct eigenvalues* $\lambda_j(A)$ *and the corresponding eigenvectors* $\mathbf{e}_j(A)$, $j = 1, \ldots, d$, *and let* $\hat{A} = A + \Delta_A$ *be a perturbed matrix with eigenvalues and eigenvectors* $(\lambda_j(\hat{A}), \mathbf{e}_j(\hat{A}))$, $j = 1, \ldots, d$. *Then, there exists a constant* $0 < c_1 < \infty$, *depending only on* A, *such that, for* $j = 1, \ldots, d$,

$$|\lambda_j(\hat{A}) - \lambda_j(A) - \mathbf{e}_j^{\mathrm{T}}(A) \Delta_A \mathbf{e}_j(A)| \leqslant c_1 \|\Delta_A\|_2^2$$

and

$$\left\| \mathbf{e}_j(\hat{A}) - \mathbf{e}_j(A) - \sum_{s=1, s \neq j}^{d} \frac{\mathbf{e}_s^{\mathrm{T}}(A) \Delta_A \mathbf{e}_j(A)}{\lambda_s(A) - \lambda_j(A)} \mathbf{e}_s(A) \right\| \leqslant c_1 \|\Delta_A\|_2^2.$$

This lemma can be viewed as a mild refinement of standard results in matrix perturbation theory (see Kato 1995; or Stewart and Sun 1990). It follows from Lemma A of Kneip and Utikal (2001), so we omit the proof.

Proof of Theorem 1. Let $T = VMV^{\mathrm{T}}$ be the spectral decomposition of T with an orthogonal matrix V and a diagonal matrix $M = \mathrm{diag}(m_1, \ldots, m_d)$ where $m_j > 0$, $j = 1, \ldots, d$. The columns $\mathbf{v}_j$ of V are normalized eigenvectors of T with eigenvalues $m_j : T\mathbf{v}_j = m_j\mathbf{v}_j$, $j = 1, \ldots, d$. Now, we set $A = T$, $\hat{A} = \hat{T}$, $\mathbf{e}_j(A) = \mathbf{v}_j$, $\mathbf{e}_j(\hat{A}) = \hat{\mathbf{v}}_j$, $\lambda_j(A) = m_j$, $\lambda_j(\hat{A}) = \hat{m}_j$, and obtain, using Lemmas 1 and 2,

$$m_j - \hat{m}_j = O_p(n^{-1/2}), \qquad j = 1, \ldots, d \tag{19}$$

$$\|\mathbf{v}_j - \hat{\mathbf{v}}_j\| = O_p(n^{-1/2}), \qquad j = 1, \ldots, d, \tag{20}$$

as $n \to \infty$. Now,

$$R \stackrel{\mathrm{def}}{=} \hat{T}^{1/2} - T^{1/2} = (\hat{V} - V)\hat{M}^{1/2}\hat{V}^{\mathrm{T}} + V(\hat{M}^{1/2} - M^{1/2})\hat{V}^{\mathrm{T}} + VM^{1/2}(\hat{V} - V)^{\mathrm{T}}.$$

Using (19), (20) and the fact that V and $\hat{V}$ are orthogonal matrices we find

$$\|R\|_2 = O_p(n^{-1/2}). \tag{21}$$

Next,

$$\begin{aligned}\hat{W} - W &= (T^{1/2} + R)S(T^{1/2} + R) - T^{1/2}\Sigma T^{1/2} \\ &= n^{-1/2}T^{1/2}\Delta_\Sigma T^{1/2} + T^{1/2}SR + RST^{1/2} + RSR,\end{aligned}$$

and from Lemma 1 and (21) we deduce that $\hat{W} = W + \Delta_W$, where $\|\Delta_W\|_2 = O_p(n^{-1/2})$, as $n \to \infty$. Applying Lemma 2 again to the eigenvectors $\hat{\mathbf{q}}_j$ of $\hat{W}$, we obtain

$$\|\hat{\mathbf{q}}_j - \mathbf{q}_j\| = O_p(n^{-1/2}), \tag{22}$$

as $n \to \infty$. Also,

$$\hat{T}^{1/2}\hat{\mathbf{q}}_j - T^{1/2}\mathbf{q}_j = (T^{1/2} + R)(\hat{\mathbf{q}}_j - \mathbf{q}_j) + R\mathbf{q}_j.$$

This, together with (21) and (22), entails that $\hat{T}^{1/2}\hat{\mathbf{q}}_j - T^{1/2}\mathbf{q}_j = O_p(n^{-1/2})$, as $n \to \infty$. Now, since the estimators $\hat{\boldsymbol{\beta}}_j$ of $\boldsymbol{\beta}_j$ are obtained by normalizing the vectors $\hat{T}^{1/2}\hat{\mathbf{q}}_j$, the theorem is proved. □

4. Asymptotics for estimators of the density p

We now prove that the estimator $\hat{p}$ defined in (13) has optimal rate of convergence when Condition 3 holds. We will need the following assumption on the kernel $\tilde{K}$ used in (13).

Condition 6. *The kernel* $\tilde{K} : \mathbb{R}^1 \to \mathbb{R}^1$ *is a function supported on* $[-1, 1]$ *that satisfies the*

Lipschitz condition, that is, $|\tilde{K}(u) - \tilde{K}(u')| \leqslant L_K|u - u'|$, *for all* $u, u' \in \mathbb{R}^1$, *for some* $L_K < \infty$, *and*

$$\int \tilde{K}(u)\mathrm{d}u = 1, \qquad \int u^l \tilde{K}(u)\mathrm{d}u = 0, \quad l = 1, \ldots, \lfloor s \rfloor,$$

where $s = \min_{1 \leqslant j \leqslant d} s_j$.

Theorem 2. *Let* $\hat{\beta}_j$ *be estimators of* β_j *based on* $X_1, \ldots, X_n$ *such that* (16) *holds. Assume that Conditions 3 and 6 are met and* $\mathrm{E}\|X\|^a < \infty$ *for some* $a > d(2b+1)/(2b-1)$. *Then, for every* $x \in \mathbb{R}^d$, *the estimator* $\hat{p}$ *defined in* (13) *with* $\tilde{h}_j \asymp n^{-1/(2s_j+1)}$, $j = 1, \ldots, d$, *satisfies*

$$\hat{p}(x) - p(x) = O_p\big(n^{-s/(2s+1)}\big),$$

as $n \to \infty$.

The rate given in Theorem 2 (characterized by $s = \min_{1 \leqslant j \leqslant d} s_j$) is optimal in a minimax sense under our smoothness assumptions on the densities p_j. In fact, at the expense of a routine technical effort, the result of Theorem 2 can be turned into a uniform one over the class of densities satisfying (1) such that $p_j \in \Sigma(s_j, L_j)$, $j = 1, \ldots, d$. This gives an upper bound on the minimax risk for estimation of p at a fixed point x. The lower bound with the same rate $n^{-s/(2s+1)}$ is a trivial consequence of the one-dimensional lower bound (Stone 1980) applied successively to d subclasses of densities such that only one component density in the right-hand side of (1) is allowed to vary, all others being fixed.

Proof of Theorem 2. Let $\tilde{p}_j(u)$ denote a marginal kernel estimator based on the sample $X_1^{\mathrm{T}}\hat{\beta}_j, \ldots, X_n^{\mathrm{T}}\hat{\beta}_j$:

$$\tilde{p}_j(u) = \frac{1}{n\tilde{h}_j} \sum_{i=1}^{n} \tilde{K}\left(\frac{X_i^{\mathrm{T}}\hat{\beta}_j - u}{\tilde{h}_j}\right).$$

We have

$$|\hat{p}(x) - p(x)| \leqslant |\det(B) - \det(\hat{B})| \prod_{j=1}^{d} p_j(x^{\mathrm{T}}\beta_j) + |\det(\hat{B})| \left| \prod_{j=1}^{d} \tilde{p}_j(x^{\mathrm{T}}\hat{\beta}_j) - \prod_{j=1}^{d} p_j(x^{\mathrm{T}}\beta_j) \right|.$$

Define the random event

$$\Omega_1 = \{\|\hat{\beta}_j - \beta_j\| \leqslant n^{-1/2} \log n, \ \text{ for all } j = 1, \ldots, d\}.$$

In view of (16) we have, for the complement $\hat{\Omega}_1$, $\mathrm{P}(\hat{\Omega}_1) = o(1)$, as $n \to \infty$, so it suffices in what follows to work on the event Ω_1. Since the p_j are uniformly bounded, on Ω_1 we have

$$|\hat{p}(x) - p(x)| \leqslant C\left(n^{-1/2}\log n + \left|\prod_{j=1}^{d}\tilde{p}_j(x^{\mathrm{T}}\hat{\beta}_j) - \prod_{j=1}^{d}p_j(x^{\mathrm{T}}\beta_j)\right|\right)$$

$$\leqslant C\left(n^{-1/2}\log n + \max_{1\leqslant j\leqslant d}|\tilde{p}_j(x^{\mathrm{T}}\hat{\beta}_j) - p_j(x^{\mathrm{T}}\beta_j)|\right).$$

(Here and later we denote by C finite positive constants, possibly different on different occasions.) Thus, to prove the theorem it remains to show that

$$\max_{1\leqslant j\leqslant d}|\tilde{p}_j(x^{\mathrm{T}}\hat{\beta}_j) - p_j(x^{\mathrm{T}}\beta_j)| = O_p\big(n^{-s/(2s+1)}\big), \tag{23}$$

as $n \to \infty$. Introduce the functions

$$g_j(v) = \int \tilde{K}(u)p_j(v + u\tilde{h}_j)\mathrm{d}u = \frac{1}{\tilde{h}_j}\int \tilde{K}\left(\frac{w - v}{\tilde{h}_j}\right)p_j(w)\mathrm{d}w, \qquad j = 1, \ldots, d.$$

We have

$$|\tilde{p}_j(x^{\mathrm{T}}\hat{\beta}_j) - p_j(x^{\mathrm{T}}\beta_j)| \leqslant J_1 + J_2 + J_3, \tag{24}$$

where

$$J_1 = |\tilde{p}_j(x^{\mathrm{T}}\beta_j) - p_j(x^{\mathrm{T}}\beta_j)|, \qquad J_2 = |g_j(x^{\mathrm{T}}\hat{\beta}_j) - g_j(x^{\mathrm{T}}\beta_j)|,$$

$$J_3 = \left|[\tilde{p}_j(x^{\mathrm{T}}\hat{\beta}_j) - \tilde{p}_j(x^{\mathrm{T}}\beta_j)] - [g_j(x^{\mathrm{T}}\hat{\beta}_j) - g_j(x^{\mathrm{T}}\beta_j)]\right|.$$

It follows from standard bias-variance evaluations for one-dimensional kernel estimates (see Ibragimov and Has'minskii 1981, Chapter 4) that

$$J_1 = O_p\big(n^{-s/(2s+1)}\big) \tag{25}$$

as $n \to \infty$. Next, using the fact that p_j is Lipschitz (cf. Condition 3 and the remarks after it), we obtain

$$J_2 \leqslant \int |\tilde{K}(u)|\,|p_j(x^{\mathrm{T}}\hat{\beta}_j + u\tilde{h}_j) - p_j(x^{\mathrm{T}}\beta_j + u\tilde{h}_j)|\mathrm{d}u$$

$$\leqslant C\|x\|\,\|\hat{\beta}_j - \beta_j\|\int |\tilde{K}(u)|\mathrm{d}u \leqslant Cn^{-1/2}\log n, \tag{26}$$

provided the event Ω_1 holds. We now prove that

$$\lim_{n\to\infty}\mathrm{P}\big(J_3 \geqslant n^{-s/(2s+1)}\big) = 0. \tag{27}$$

We have

$$\mathrm{P}\big(J_3 \geqslant n^{-s/(2s+1)}\big) \leqslant \mathrm{P}\big(J_3 \geqslant n^{-s_j/(2s_j+1)}\big) \leqslant \mathrm{P}(\overline{\Omega}_1) + \mathrm{P}(\Omega_2),$$

where Ω_2 is the random event defined by

$$\Omega_2 = \left\{ \sup_{\beta:\|\beta_j-\beta\|\leqslant n^{-1/2}\log n} |[\tilde{p}_j(x^{\mathrm{T}}\beta) - \tilde{p}_j(x^{\mathrm{T}}\beta_j)] - [g_j(x^{\mathrm{T}}\beta) - g_j(x^{\mathrm{T}}\beta_j)]| \geqslant n^{-s_j/(2s_j+1)} \right\}.$$

Hence, to prove (27), it remains to show that $\mathrm{P}(\Omega_2) = o(1)$, as $n \to \infty$. We will do this using the following lemma (Ibragimov and Has'minskii 1981, Appendix 1):

Lemma 3. *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*

$$\mathrm{E}|\eta(t+\Delta) - \eta(t)|^a \leqslant H\|\Delta\|^a, \qquad \text{for all } t, \Delta \in \mathbb{R}^d,$$

$$\mathrm{E}|\eta(t)|^a \leqslant H, \qquad \text{for all } t \in \mathbb{R}^d.$$

Then for every $\delta > 0$ and $t_0 \in \mathbb{R}^d$ such that $\|t_0\| \leqslant C_0$,

$$\mathrm{E}\left[\sup_{t:\|t-t_0\|\leqslant\delta} |\eta(t) - \eta(t_0)|\right] \leqslant B_0(C_0+\delta)^d H^{1/a}\delta^{1-d/a}, \tag{28}$$

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

Consider now the process

$$\eta(\beta) = \tilde{p}_j(x^{\mathrm{T}}\beta) - g_j(x^{\mathrm{T}}\beta) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\tilde{h}_j}\tilde{K}\left(\frac{(X_i - x)^{\mathrm{T}}\beta}{\tilde{h}_j}\right) - g_j(x^{\mathrm{T}}\beta), \qquad \beta \in \mathbb{R}^d, \tag{29}$$

and choose a satisfying the assumptions of the theorem. Since $\eta(\beta)$ defined in (29) for any fixed β is an average of independent and identically distributed zero-mean bounded random variables

$$\xi_i = \frac{1}{\tilde{h}_j}\tilde{K}\left(\frac{(X_i - x)^{\mathrm{T}}\beta}{\tilde{h}_j}\right) - g_j(x^{\mathrm{T}}\beta),$$

we have, by Rosenthal's inequality (see Petrov 1995),

$$\mathrm{E}|\eta(\beta)|^a \leqslant Cn^{-a/2}\mathrm{E}|\xi_1|^a \leqslant Cn^{-a/2}\tilde{h}_j^{1-a}, \tag{30}$$

where the last inequality follows from the fact that p and $\tilde{K}$ are bounded and $\tilde{K}$ is compactly supported. Again, by Rosenthal's inequality,

$$\begin{aligned}
\mathrm{E}|\eta(\beta+\Delta)-\eta(\beta)|^a &\leqslant Cn^{-a/2}\tilde{h}_j^{-a}\mathrm{E}\left|\tilde{K}\left(\frac{(X_i-x)^{\mathrm{T}}(\beta+\Delta)}{\tilde{h}_j}\right)-\tilde{K}\left(\frac{(X_i-x)^{\mathrm{T}}\beta}{\tilde{h}_j}\right)\right|^a \\
&\leqslant Cn^{-a/2}\tilde{h}_j^{-a}\int\left|\tilde{K}\left(\frac{(z-x)^{\mathrm{T}}(\beta+\Delta)}{\tilde{h}_j}\right)-\tilde{K}\left(\frac{(z-x)^{\mathrm{T}}\beta}{\tilde{h}_j}\right)\right|^a p(z)\mathrm{d}z \\
&\leqslant Cn^{-a/2}\tilde{h}_j^{-a}\left(L_K\tilde{h}_j^{-a}\int\|z-x\|^a\|\Delta\|^a p(z)\mathrm{d}z\right) \\
&\leqslant Cn^{-a/2}\tilde{h}_j^{-2a}\|\Delta\|^a, \qquad (31)
\end{aligned}$$

in view of Condition 6 and of the assumption that $\mathrm{E}\|X\|^a<\infty$. It follows from (30) and (31) that the assumptions of Lemma 3 are satisfied for the process η defined in (29) with $H=Cn^{-a/2}\tilde{h}_j^{-2a}$, and thus (28) yields

$$\mathrm{E}\left[\sup_{\beta:\|\beta-\beta_j\|\leqslant\delta}|\eta(\beta)-\eta(\beta_j)|\right]\leqslant CH^{1/a}\delta^{1-d/a}\leqslant Cn^{-1/2}\tilde{h}_j^{-2}\delta^{1-d/a}.$$

Using this inequality for $\delta=n^{-1/2}\log n$ and the Markov inequality, we may bound $\mathrm{P}(\Omega_2)$ as follows:

$$\mathrm{P}(\Omega_2)\leqslant Cn^{s_j/(2s_j+1)-1/2}\tilde{h}_j^{-2}\delta^{1-d/a}=O\big((\log n)^{1-d/a}n^{(s_j+2)/(2s_j+1)-1+d/(2a)}\big)=o(1),$$

as $n\to\infty$, whenever $a>d(2s_j+1)/(2s_j-1)$ (which is implied by the assumption on a in the theorem). This proves (27). Now, (24)–(27) entail (23), and hence the theorem. □

Appendix. Proof of Lemma 1

In view of the assumption $\mathrm{E}\|X\|^4<\infty$, relation (17) easily follows from application of Chebyshev's inequality to each component of the matrix S. We therefore prove only (18). It suffices to prove (18) componentwise, that is, to show that

$$\hat{t}_{lk}-t_{lk}=O_p(n^{-1/2}),\qquad n\to\infty,$$

where $\hat{t}_{lk}$ and t_{lk} are the (l,k)th entries of the matrices $\hat{T}$ and T, respectively. Due to the bias-variance decomposition, the last relation is proved if we show that

$$\mathrm{bias}(\hat{t}_{lk})=\mathrm{E}(\hat{t}_{lk})-t_{lk}=O(n^{-1/2}),\qquad n\to\infty, \qquad (32)$$

$$\mathrm{var}(\hat{t}_{lk})=\mathrm{E}[(\hat{t}_{lk}-\mathrm{E}(\hat{t}_{lk}))^2]=O(n^{-1}),\qquad n\to\infty. \qquad (33)$$

Since $b\leqslant\min_{1\leqslant i\leqslant d}s_i$, Condition 3 entails that all the partial derivatives up to order $\lfloor b\rfloor$ of the density p are bounded and continuous and satisfy the Hölder condition with Hölder exponent $b-\lfloor b\rfloor$. Using this fact and Condition 2 we obtain, for any $l=1,\ldots,d$,

$$\mathrm{E}\left(Q_l\left(\frac{X_j - X_i}{h}\right)\middle| X_i\right) = \int K_1\left(\frac{z_l - X_{il}}{h}\right) \prod_{r=1, r\neq l}^{d} K\left(\frac{z_r - X_{ir}}{h}\right) p(z_1, \ldots, z_d)\mathrm{d}z_1 \ldots \mathrm{d}z_d$$

$$= h^d \int K_1(u_l) \prod_{r=1, r\neq l}^{d} K(u_r) p(X_{i1} + hu_1, \ldots, X_{id} + hu_d)\mathrm{d}u_1 \ldots \mathrm{d}u_d$$

$$= h^d \left(\frac{\partial p}{\partial x_l}(X_i)h + O(h^b)\right) \tag{34}$$

as $h \to 0$. Consequently, for any $l = 1, \ldots, d$, uniformly in X_i, we get that (almost surely)

$$\mathrm{E}\left(Q_l\left(\frac{X_j - X_i}{h}\right)\middle| X_i\right) = O(h^{d+1}) \tag{35}$$

as $h \to 0$.

Without loss of generality, we prove (33) for $l = 1$, $k = 2$. We have

$$\hat{t}_{12} = \frac{1}{n(n-1)^2 h^{2(d+1)}} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \sum_{m=1, m\neq i}^{n} U_{ijm},$$

where

$$U_{ijm} = Q_1\left(\frac{X_j - X_i}{h}\right) Q_2\left(\frac{X_m - X_i}{h}\right).$$

Clearly,

$$\mathrm{var}(\hat{t}_{12}) \leqslant \frac{C}{n^6 h^{4(d+1)}} (A_1 + A_2), \tag{36}$$

where

$$A_1 = \mathrm{var}\left(\sum\sum\sum{}^{*} U_{ijm}\right), \qquad A_2 = \mathrm{var}\left(\sum\sum{}^{*} \mathrm{V}_{ij}\right), \qquad V_{ij} = U_{ijj},$$

and $\sum\sum^{*}$, $\sum\sum\sum^{*}$ are the sums over i, j and i, j, m varying from 1 to n such that all the indices of every summand are distinct. To evaluate the right-hand side of (36) we will use the following lemma.

Lemma 4. *Let $U(x, y, z)$ and $V(x, y)$ be real-valued functions of $x, y, z \in \mathbb{R}^d$, such that $\mathrm{E}(U_{123}^2)$ and $\mathrm{E}(V_{12}^2)$ are bounded, where $U_{ijm} = U(X_i, X_j, X_m)$ and $V_{ij} = V(X_i, X_j)$. Then there exists an absolute constant $C_* < \infty$ such that*

$$\mathrm{var}\left(\sum\sum{}^{*} V_{ij}\right) \leqslant C_*\left(n^2 \mathrm{E}(V_{12}^2) + n^3[\mathrm{E}(\mathrm{V}_{1\cdot}^2) + \mathrm{E}(V_{\cdot 2}^2)]\right) \tag{37}$$

and

$$\operatorname{var}\left(\sum\sum\sum{}^{*} U_{ijm}\right) \leqslant C_{*}(n^3 \mathrm{E}(U_{123}^2) + n^4[\mathrm{E}(U_{12\cdot}^2) + \mathrm{E}(U_{1\cdot 3}^2) + \mathrm{E}(U_{\cdot 23}^2)] + n^5[\mathrm{E}(U_{1\cdot\cdot}^2) + \mathrm{E}(U_{\cdot 2\cdot}^2) + \mathrm{E}(U_{\cdot\cdot 3}^2)]), \tag{38}$$

where the dot in the index means that the random variable in the corresponding position is integrated out: for example, $V_{i\cdot} = \mathrm{E}(V_{ij}|X_i)$, $V_{\cdot j} = \mathrm{E}(V_{ij}|X_j)$, $U_{ij\cdot} = \mathrm{E}(U_{ijm}|X_i, X_j)$, $U_{i\cdot\cdot} = \mathrm{E}(U_{ijm}|X_i)$.

The proof of Lemma 4 is analogous to the well-known Hoeffding projection argument for U-statistics, the only difference being in the lack of symmetry. Similar calculations can be found in Hall (1988, Lemma 5.5): to prove (37) and (38) it suffices to write $V_{ij} - \mathrm{E}(V_{ij}) = \zeta_{ij} + V_{i\cdot} + V_{\cdot j}$ and

$$U_{ijm} - \mathrm{E}(U_{ijm}) = W_{ijm} + (U_{ij\cdot} + U_{i\cdot m} + U_{\cdot ij}) - (U_{i\cdot\cdot} + U_{\cdot j\cdot} + U_{\cdot\cdot k}),$$

and to note that the random variables ζ_{ij} and W_{ijm} satisfy the assumptions of Lemma 5.5 in Hall (1988), namely all their conditional expectations are 0. Application of that lemma finishes the proof of Lemma 4.

We first evaluate A_2 using Lemma 4. Since the kernels Q_1, Q_2 are bounded and compactly supported and p is bounded, we obtain

$$\mathrm{E}(V_{12}^2) \leqslant Ch^d, \qquad \mathrm{E}(V_{1\cdot}^2) \leqslant Ch^{2d}, \qquad \mathrm{E}(V_{\cdot 2}^2) \leqslant Ch^{2d}, \tag{39}$$

which yields, together with (37), that

$$A_2 \leqslant C(n^2 h^d + n^3 h^{2d}). \tag{40}$$

We now control the term A_1. For the same reason as in (39), we obtain

$$\mathrm{E}(U_{123}^2) \leqslant Ch^{2d}. \tag{41}$$

Note also that, in view of (35),

$$|U_{12\cdot}| = \left|Q_1\left(\frac{X_2 - X_1}{h}\right)\mathrm{E}\left(Q_2\left(\frac{X_3 - X_1}{h}\right)\middle| X_1\right)\right| \leqslant Ch^{d+1}\left|Q_1\left(\frac{X_2 - X_1}{h}\right)\right|,$$

which implies, together with the fact that the kernel Q_1 is bounded and compactly supported and that p is bounded, that

$$\mathrm{E}(U_{12\cdot}^2) \leqslant Ch^{3d+2}. \tag{42}$$

Quite similarly,

$$\mathrm{E}(U_{1\cdot 3}^2) \leqslant Ch^{3d+2}. \tag{43}$$

Next, since p is bounded,

$$U_{\cdot 23} = \mathrm{E}_{X_1}\left(Q_1\left(\frac{X_2 - X_1}{h}\right)Q_2\left(\frac{X_3 - X_1}{h}\right)\right) \leqslant Ch^d Q_*\left(\frac{X_2 - X_3}{h}\right),$$

where E_{X_1} denotes the expectation with respect to X_1 and $Q_* = |Q_1| * |Q_2|$ is the convolution kernel. This, and the fact that Q_* is bounded and compactly supported and p is bounded, yields

$$\mathrm{E}(U_{\cdot 23}^2) \leqslant Ch^{3d}. \tag{44}$$

Now, using (35),

$$|U_{1\cdot\cdot}| = \left|\mathrm{E}\left(Q_1\left(\frac{X_2 - X_1}{h}\right)\middle| X_1\right)\mathrm{E}\left(Q_2\left(\frac{X_3 - X_1}{h}\right)\middle| X_1\right)\right| \leqslant Ch^{2(d+1)},$$

and therefore

$$\mathrm{E}(U_{1\cdot\cdot}^2) \leqslant Ch^{4(d+1)}. \tag{45}$$

To evaluate $\mathrm{E}(U_{\cdot 2\cdot}^2)$ and $\mathrm{E}(U_{\cdot\cdot 3}^2)$ we use the following relation obtained in the same way as (34), (35):

$$\mathrm{E}_{X_1}\left[Q_1\left(\frac{X_2 - X_1}{h}\right)\frac{\partial p}{\partial x_2}(X_1)\right] = O(h^{d+1})$$

as $h \to 0$. This relation and (34) yield

$$\begin{aligned} U_{\cdot 2\cdot} &= \mathrm{E}_{X_1}\left[Q_1\left(\frac{X_2 - X_1}{h}\right)\mathrm{E}\left(Q_2\left(\frac{X_3 - X_1}{h}\right)\middle| X_1\right)\right] \\ &= \mathrm{E}_{X_1}\left[Q_1\left(\frac{X_2 - X_1}{h}\right)h^d\left(\frac{\partial p}{\partial x_2}(X_1)h + O(h^b)\right)\right] \\ &= O(h^{2(d+1)}). \end{aligned}$$

A similar calculation is valid for $U_{\cdot\cdot 3}$. Thus,

$$\mathrm{E}(U_{\cdot 2\cdot}^2) \leqslant Ch^{4(d+1)}, \qquad \mathrm{E}(U_{\cdot\cdot 3}^2) \leqslant Ch^{4(d+1)}. \tag{46}$$

Combining (38), (41)–(45) and (46), we obtain

$$A_1 \leqslant C(n^3h^{2d} + n^4h^{3d} + n^5h^{4(d+1)}).$$

This inequality, together with (36) and (40), gives

$$\mathrm{var}(\hat{t}_{12}) = O\left(\frac{1}{n^4h^{3d+4}} + \frac{1}{n^3h^{2d+4}} + \frac{1}{n^2h^{d+4}} + \frac{1}{n}\right) = O\left(\frac{1}{n}\right)$$

as $n \to \infty$, where the last equality holds because $nh^{2d+4} \to \infty$ (see Condition 1). This finishes the proof of (33).

We now prove (32). Again set $l = 1$, $k = 2$. The bias of $\hat{t}_{12}$ is

$$\operatorname{bias}(\hat{t}_{12}) = \frac{1}{n(n-1)^2 h^{2(d+1)}} \left[\left(\sum\sum\sum {}^{*}\mathrm{E}(U_{ijm}) - t_{12} \right) + \sum\sum {}^{*}\mathrm{E}(V_{ij}) \right]$$

$$= \left(\frac{(n-2)\mathrm{E}(U_{123})}{(n-1)h^{2(d+1)}} - t_{12} \right) + \frac{\mathrm{E}(V_{12})}{(n-1)h^{2(d+1)}}. \tag{47}$$

Now, using (34), we find

$$\mathrm{E}(U_{123}) = \mathrm{E}_{X_1} \left[\mathrm{E}\left(Q_1\left(\frac{X_2 - X_1}{h} \right) \middle| X_1 \right) \mathrm{E}\left(Q_2\left(\frac{X_3 - X_1}{h} \right) \middle| X_1 \right) \right]$$

$$= h^{2(d+1)} \mathrm{E}\left[\frac{\partial p}{\partial x_1}(X_1) \frac{\partial p}{\partial x_2}(X_1) \right] + O(h^{2d+b+1})$$

$$= h^{2(d+1)} t_{12} + O(h^{2d+b+1}), \tag{48}$$

as $h \to 0$. Also, as in (39),

$$\mathrm{E}(V_{12}) = O(h^d),$$

as $h \to 0$. Substitution of this relation and of (48) into (47) and application of Condition 1 yields

$$\operatorname{bias}(\hat{t}_{12}) = O\left(h^{b-1} + \frac{1}{nh^{d+2}} \right) = O(n^{-1/2}),$$

as $n \to \infty$. This proves (32) and hence the lemma.

Acknowledgement

The authors would like to thank the referee who pointed out the paper by Eriksson *et al*. (2001) which suggests an approach to ICA different from ours and makes use of empirical characteristic functions. The referee also indicated the preprint by Chen and Bickel (2003) which appeared after our paper was submitted and which proves the root-n consistency of the method proposed in Eriksson *et al*. (2001).

References

Bach, F. and Jordan, M. (2002) Kernel independent component analysis. *J. Machine Learning Res.*, **3**, 1–48.

Cardoso, J.-F. (1999) High-order contrasts for independent component analysis. *Neural Computation*, **11**, 157–192.

Chen, A. and Bickel, P. (2003) Efficient independent component analysis. Technical Report no. 634, Department of Statistics, University of California, Berkeley.

Eriksson, J., Kankainen, A., and Koivunen, V. (2001) Novel characteristic function based criteria for

ICA. In T.-W. Lee, A. Jung, S. Makeig and T. Sejnowski (eds), *Proceedings of the 3rd International Conference on Independent Component Analysis and Signal Separation*. San Diego: UCSD Institute for Neural Computation.

Hall, P. (1988) Estimating the direction in which a data set is most interesting. *Probab. Theory Related Fields*, **80**, 51–77.

Härdle, W. and Tsybakov, A.B. (1991) Discussion of 'Sliced inverse regression'. *J. Amer. Statist. Assoc.*, **86**, 333–335.

Huber, P. (1985) Projection pursuit. *Ann. Statist.*, **13**, 435–475.

Hyvärinen, A., Karhunen, J. and Oja, E. (2001) *Independent Component Analysis*. New York: Wiley.

Ibragimov, I.A. and Has'minskii, R.Z. (1981) *Statistical Estimation: Asymptotic Theory*. New York: Spinger-Verlag.

Kato, T. (1995) *Perturbation Theory for Linear Operators*. New York: Springer-Verlag.

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

Lewis, D.W. (1991) *Matrix Theory*. Teaneck NJ: World Scientific.

Li, K.-C. (1991) 'Sliced inverse regression for dimension reduction'. *J. Amer. Statist. Assoc.*, **86**, 316–342.

Nemirovski, A. (2000) Topics in non-parametric statistics. In P. Bernard (ed.), *Lectures on Probability and Statistics: Ecole d'Été de Probabilités de Saint-Flour XXVIII – 1998*. Lecture Notes in Math. 1738, pp. 85–277. Berlin: Springer-Verlag.

Petrov, V.V. (1995) *Limit Theorems of Probability Theory*. Oxford: Clarendon Press.

Pham, D.T. (1996) Blind separation of instantaneous mixture of sources via an independent component analysis. *IEEE Trans. Signal Process.*, **44**, 2768–2779.

Roberts, S. and Everson, R. (2001) *Independent Component Analysis: Principles and Practice*. Cambridge University Press.

Samarov, A.M. (1993) Exploring regression structure using nonparametric functional estimation. *J. Amer. Statist. Assoc.*, **88**, 836–847.

Stewart, G.W. and Sun, J. (1990) *Matrix Perturbation Theory*. London: Academic Press.

Stone, C.J. (1980) Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, **8**, 1348–1360.

Received November 2002 and revised February 2004