Statistics

A Journal of Theoretical and Applied Statistics

ISSN: 0233-1888 (Print) 1029-4910 (Online) Journal homepage: https://www.tandfonline.com/loi/gsta20

Plug-in L_2-upper error bounds in deconvolution, for a mixing density estimate in R^d and for its derivatives, via the L_1-error for the mixture

Yannis G. Yatracos

To cite this article: Yannis G. Yatracos (2019) Plug-in L_2-upper error bounds in deconvolution, for a mixing density estimate in R^d and for its derivatives, via the L_1-error for the mixture, Statistics, 53:6, 1251-1268, DOI: 10.1080/02331888.2019.1632313

To link to this article: https://doi.org/10.1080/02331888.2019.1632313

Published online: 30 Jul 2019.

STATISTICS
2019, VOL. 53, NO. 6, 1251–1268
https://doi.org/10.1080/02331888.2019.1632313

Taylor & Francis
Taylor & Francis Group

Plug-in L_2-upper error bounds in deconvolution, for a mixing density estimate in R^d and for its derivatives, via the L_1-error for the mixture

Yannis G. Yatracos[a,b]

[a]Yau Mathematical Sciences Center, Tsinghua University, Beijing, People's Republic of China; [b]School of Communication and Media Studies, Cyprus University of Technology, Lemesos, Cyprus

ABSTRACT

In deconvolution in R^d, $d \geq 1$, with mixing density $p(\in \mathcal{P})$ and kernel h, the mixture density $f_p(\in \mathcal{F}_p)$ is estimated with MDE $f_{\hat{p}_n}$, having upper L_1-error rate, a_n, in probability or in risk; $\hat{p}_n \in \mathcal{P}$. In one application, $\mathcal{P}$ consists of L_1-separable densities in R with differences changing sign at most J times and $h(x-y)$ Totally Positive. When h is known and p is $\tilde{q}$-smooth, vanishing outside a compact in R^d, plug-in upper bounds are provided for the L_2-error rate of $\hat{p}_n$ and its $[s]$-th mixed partial derivative $\hat{p}_n^{(s)}$, via $\|f_{\hat{p}_n} - f_p\|_1$, with rates $(\log a_n^{-1})^{-N_1}$ and $a_n^{N_2}$, respectively, for h super-smooth and smooth; $\tilde{q} \in R^+$, $[s] \leq \tilde{q}$, $d \geq 1$, $N_1 > 0$, $N_2 > 0$. For $a_n \sim (\log n)^{\zeta} \cdot n^{-\delta}$, the former rate is optimal for *any* $\delta > 0$ and the latter misses the optimal by the factor $(\log n)^{\xi}$ when $\delta = .5$; $\zeta > 0, \xi > 0$. N_1 and N_2 appear in optimal rates and lower error and risk bounds in the deconvolution literature.

ARTICLE HISTORY
Received 16 December 2017
Accepted 11 June 2019

KEYWORDS
Deconvolution; minimum distance estimation; plug-in upper error/risk bounds; totally positive kernels; Vapnik–Chervonenkis classes

AMS SUBJECT CLASSIFICATIONS
62G07; 62G20

1. Introduction

In the deconvolution problem, random vectors Y and X in R^d, d ≥ 1, have densities, respectively, p and f_p and satisfy the equation

$$X = Y + \epsilon; \tag{1}$$

Y is independent of the error ϵ that has density h,

$$f_p(x) = h * p(x) = \int_{R^d} h(x-y)p(y)\mathrm{d}y,\ p \in \mathcal{P}, \tag{2}$$

$$\mathcal{F}_{\mathcal{P}} = \mathcal{F}_{\mathcal{P},d} = \{f_p,\ p \in \mathcal{P}\}; \tag{3}$$

$\mathcal{P}$ is any class of densities of interest, '$*$' denotes convolution. Independent copies $X_1, \ldots, X_n$ of X are observed and the goal is to estimate p, its derivative(s) and calculate the estimation errors. Usually, h is assumed known, with non-vanishing Fourrier transform $\tilde{h}$. The classic approach is to estimate p via a kernel estimate of f_p.

CONTACT Yannis G. Yatracos yannis.yatracos@yahoo.com Yau Mathematical Sciences Center, Tsinghua University, Jingzhai 106, Haidian District, Beijing 100084, People's Republic of China

Until recently, research has been devoted mainly to the one-dimensional deconvolution problem. However, X-observations in R^d can be used to estimate f_p, e.g. with a kernel estimate $\hat{f}_n$; $d > 1$. A Minimum Distance Estimate (*MDE*) $f_{\hat{p}_n}$ with $\hat{p}_n \in \mathcal{P}$ can then be obtained, either via $\hat{f}_n$ or directly as described in Section 3, with calculation of upper L_1-error rate for $||f_{\hat{p}_n} - f_p||_1$ when h is either known or unknown. In applications, $\mathcal{P}$ consists either of products of d $\tilde{q}$-smooth densities defined on a compact in R, or L_1-separable densities in R with their differences changing sign at most J times and $h(x-y)$ Totally Positive; J is either known or unknown. The problem that has not been tackled so far in the literature is to derive 'plug-in' upper error and risk bounds for $\hat{p}_n$ and the s-th order mixed partial derivative, $\hat{p}_n^{(s)}$, from the rate of convergence of $f_{\hat{p}_n}$ to f_p.

This problem is addressed herein when $\mathcal{P}$ is a sup-norm compact family of $\tilde{q}$-smooth densities vanishing outside a compact $\mathcal{Y}$ in R^d (see Definition 2.3); $d \geq 1$. Upper bounds in probability for the L_2-errors of $\hat{p}_n$ and of $\hat{p}_n^{(s)}$ and for their risks are provided that depend on the L_1-error $||f_{\hat{p}_n} - f_p||_1$, non-vanishing $\tilde{h}$ and the smoothing parameter b_n of a trapezoidal kernel K [1] that is used as approximation tool.

If $f_{\hat{p}_n}$ is L_1-optimal with respect to some criterion, e.g. minimax, the difference of $\hat{p}_n$'s L_2-error rate from the optimal is not expected to be substantial. For example, if h is super-smooth (see (28)) and $f_{\hat{p}_n}$ converges to f_p in L_1-distance with the typical rate $n^{-\delta} \cdot (\log n)^{\zeta}$ in probability or in risk, it follows from (32) that $\hat{p}_n$ has upper L_2-error rate the optimal, $(\log n)^{-\tilde{q}/k}$, for *any* $\delta,\ \zeta$; k determines the rate of the exponential decay of $\tilde{h}, \delta > 0,\ \zeta \in R$. If h is smooth (see (29)), for $f_{\hat{p}_n}$'s typical rate and from (33), $\hat{p}_n$ has upper L_2-error rate $[(\log n)^{2\zeta}/n^{2\delta})]^{\frac{\tilde{q}}{2\tilde{q}+2\sum_{i=1}^{d}\beta_i+d}}$, which misses the optimal by the factor $(\log n)^{\xi}$ when $\delta = .5;\ \xi > 0,\ \beta_1, \ldots, \beta_d$ are the exponents determining $\tilde{h}$'s algebraic decay.

The exponents in the rates' bounds coincide with those of the lower or optimal rates for the isotropic Hölder and Sobolev classes and for the isotropic and bounded Nikolskii class pointwise, for the mean integrated square error, and in L_u-distance and risk, $2 \leq u \leq \infty$ [2–4]. The same holds in univariate deconvolution (see, e.g. [5–8], [9, Chapter 2], [10]).

Some readers may, as the referees did, acknowledge the generality of the assumptions concerning the noise density h but express concerns about the isotropy condition imposed on p and the non-adaptivity of the obtained estimates. Rates can be obtained under anisotropic Lipschitz, Sobolev and Nikolskii conditions by the interested reader following the same approach and appropriate Taylor expansions; it is expected that either the least favourable anisotropic component or their average will determine the convergence rate. Remarks 4.1 and 4.3 indicate how the rates are affected by anisotropic conditions. In Minimum Distance Estimation the existence of unknown parameters is addressed by enlarging the minimization grid, including additional grids from the corresponding parameter spaces; see, e.g. Proposition 3.2. In the present context, the unknown parameters, i.e. the number q_i of existing partial derivatives and the constants L_i, γ_i for the Lipschitz condition in each coordinate for the q_i-th partial derivative, are not expected to affect the convergence rate by more than the factor $(\ln n)^{\alpha}$ since their parameter spaces are not as 'rich' as those of $p, h, f_p;\ 0 < \alpha < 1,\ i = 1, \ldots, d$. When the modulus of continuity, w_{q_i} of the ultimate partial derivative is element of a 'richer' space, the rate of convergence may be affected by more than $(\ln n)^{\alpha}$; see, e.g. Proposition 3.2.

In multidimensional deconvolution, estimates have been obtained also by [11,12]. For the deconvolution in R, consistent estimates have been provided and, when p is $\tilde{q}$-smooth, optimality of the error rates has been established for smooth and super-smooth h, pointwise and in weighted L_u-distance, among others by [13–20] $1 \leq u \leq \infty$. Devroye [14] showed in particular that one can construct a consistent kernel estimate of p when the set $\{t : \tilde{h}(t) = 0\}$ has Lebesgue measure zero. More recent work includes, among others, [21–24]. Johannes [25] estimated non-parametrically p when ϵ's distribution in (1) is estimated.

Hall and Meister [8] presented a new estimate for p using *ridging*, 'not involving kernels in any way', used also when $\tilde{h}$ has periodic zeros. Meister [26] proposed also an estimate for p using local polynomials when $\tilde{h}$ has periodic zeros. Under additional assumptions on either p or h, the estimates in [8, see page 1542, lines −3, −2] and in [26, see the Introduction] are optimal but the assumptions and the rates are different.

2. Notation, definitions and tools

All the functions used are defined in R^d and are measurable and integrable with real values; d ≥ 1. The densities are defined with respect to Lebesgue measure. When the domain of integration is R^d, it is omitted. For any function g, its Fourrier transform is $\tilde{g}$. The vectors $X,\ Y$ take values, respectively, in $\mathcal{X},\ \mathcal{Y}$, which are both sets in R^d. $C,\ c,\ C_1,\ C_2$ denote generic positive constants. For positive a, b, $a \sim b$ means $C_1 b \leq a \leq C_2 b$. Constants $a_n,\ b_n,\ \beta_n,\ \delta_n, \theta_n$ decrease to zero as n increases.

Distances between densities are needed to evaluate the errors $(\hat{p}_n - p)$ and $(f_{\hat{p}_n} - f_p)$.

Definition 2.1 (Distances): For densities p_1, p_2 defined in $\mathcal{Y}(\subset R^d)$ their L_u-distance is

$$||p_1 - p_2||_u = \left[\int_{\mathcal{Y}} |p_1(w) - p_2(w)|^u \mathrm{d}w\right]^{1/u},\ 1 \leq u < \infty.$$

The sup-norm (or L_∞)- distance is

$$||p_1 - p_2||_\infty = \sup_{w \in \mathcal{Y}} |p_1(w) - p_2(w)|.$$

The Hellinger distance is

$$H(p_1, p_2) = \left[\int_{\mathcal{Y}} (\sqrt{p_1(y)} - \sqrt{p_2(y)})^2 \mathrm{d}y\right]^{1/2}.$$

A well-known inequality between the L_1-distance and Hellinger distance is used:

$$||p_1 - p_2||_1 \leq 2H(p_1, p_2). \tag{4}$$

Notation and definitions needed to define $\mathcal{P}$ in (2) follow.

Notation: If $x = (x_1, \ldots, x_d) \in R^d$, $a \in R$ and $s = (s_1, \ldots, s_d)$ is a d-tuple of non-negative integers,

$$x^s = (x_1^{s_1}, \ldots, x_d^{s_d}),\ xs = x_1 s_1 + \ldots + x_d s_d,\ ax = (ax_1, \ldots, ax_d),\ [s] = s_1 + \ldots + s_d;$$

for $y \in R^d$,

$$|x - y| = \max\{|x_i - y_i|, i = 1, \ldots, \mathrm{d}\}.$$

For a real-valued function g defined in R^d let $g^{(s)}(x_0)$ denote the $[s]$-th order mixed partial derivative of g at x_0, i.e.

$$g^{(s)}(x_0) = \frac{\partial^{[s]} g(x_0)}{\partial x_1^{s_1} \ldots x_d^{s_d}}.$$

Definition 2.2: The modulus of continuity w of g is a function from R^+ with positive values such that

$$w(\delta) = \sup\{|g(x) - g(y)| : |x - y| \le \delta\},\ \delta > 0. \tag{5}$$

If the rth order mixed partial derivative of g has modulus of continuity w, then

$$|g^{(t)}(x) - g^{(t)}(y)| \le w(|x - y|),\ [t] = r. \tag{6}$$

Definition 2.3: Let $\mathcal{P}$ in (2) consist of densities defined on the same compact set $\mathcal{Y}$ ($\subset R^d$), that have all s-mixed order partial derivatives uniformly bounded, $0 \le [s] \le q$, with the q-th mixed order derivative having the same and known modulus of continuity w_q.

When

$$w_q(\delta) = L \cdot \delta^{\gamma},\ L > 0, \quad 0 < \gamma < 1,\ \tilde{q} = q + \gamma, \tag{7}$$

$\mathcal{P}$ is called $\tilde{q}$-smooth family of densities, ignoring L.

Kernels are introduced, used either to obtain $f_{\hat{p}_n}$ or upper bounds on $||\hat{p}_n - p||_2$. Let $K(x)$ be a symmetric function defined in R^d at least q times continuously differentiable with bounded Fourrier transform $\tilde{K}$ having compact support $[-M, M]^d$, $M > 0$, such that for $s \in (R^+)^d$,

$$\int K(x)\mathrm{d}x = 1, \quad \int x^s K(x)\mathrm{d}x = 0,\ [s] = 1, \ldots, q, \int (|x|^q + |x|^{q+1})K(x)\mathrm{d}x < \infty. \tag{8}$$

Kernel K can be obtained by taking d-fold products of Devroye's trapezoidal kernel [1] and making smooth enough the linear leg of the trapezoid [27]. For any positive number b_n, let

$$K_n(x) = b_n^{-d} K(x b_n^{-1}), \tag{9}$$

with b_n decreasing to 0 as n increases. If $X_1, X_2, \ldots, X_n$ are independent, identically distributed (*i.i.d.*) vectors in R^d with density f, the kernel estimate of f using K is

$$\hat{f}_n(x) = \frac{1}{n} \sum_{j=1}^{n} K_n(x - X_j). \tag{10}$$

Rates of convergence are obtained herein when $h(x-y)$ is Totally Positive. Thus, the notion of Total Positivity is introduced from [28], as well as other results needed.

Definition 2.4: A real function $Q(x,y)$ of two variables ranging over linearly ordered sets $\mathcal{X}$ and $\mathcal{Y}$, respectively, is said to be Totally Positive of order r (abbreviated TP_r) if for all

$$x_1 < x_2, < \ldots < x_m,\ y_1 < y_2 < \ldots < y_m, \quad x_i \in \mathcal{X}, y_i \in \mathcal{Y},\ 1 \le m \le r, \tag{11}$$

the determinant

$$\begin{vmatrix} Q(x_1,y_1) & Q(x_1,y_2) & \ldots & Q(x_1,y_m) \\ Q(x_2,y_1) & Q(x_2,y_2) & \ldots & Q(x_2,y_m) \\ \ldots & & & \\ Q(x_m,y_1) & Q(x_m,y_2) & \ldots & Q(x_m,y_m) \end{vmatrix} \ge 0. \tag{12}$$

Most often $\mathcal{X}$ and $\mathcal{Y}$ are either intervals of the real line or countable subsets of discrete values along the real line. When r is omitted in TP_r, total positivity holds for any value of r. Many density functions $Q(x,y)$ are totally positive (TP) with respect to a σ-finite measure, *with the variable y being a real parameter*. Examples include the exponential family, the normal and the non-central t-density [28, pages 19 and 20]. A potential TP density $Q(x,y)$ is $h(x-y)$ used in $h * p$, in (2), e.g. when h is the standard normal density.

Proposition 2.1 ([29, p. 34, 1968, Theorem 3.1 (a), p. 21]): *Let $Q(x,y)$ be TP_r, let μ denote a σ-finite measure such that $\int_{\mathcal{Y}} Q(x,y)\mathrm{d}\mu(y)$ exists for every $x \in \mathcal{X}$ and $\mu(U) > 0$ for each open set U for which $U \cap \mathcal{Y}$ is not empty. Suppose $p(y)$ is bounded, measurable and changes sign $J \le r-1$ times. Let*

$$f_p(x) = \int Q(x,y)p(y)\mathrm{d}\mu(y),$$

be well defined such that the integral converges absolutely, then $f_p(x)$ changes sign at most J times.

Remark 2.1: When $h(x-y)$ is totally positive, a MDE $f_{\hat{p}_n}$ is obtained and its L_1-upper-rate of convergence to f_p is calculated for densities $\mathcal{P}_J$ in R such that for all $p_1, p_2 \in \mathcal{P}_J$, $p_1 - p_2$ changes sign at most J times. The Minimum Distance criterion is used over a Vapnik–Chervonenkis class of sets (defined below) obtained via $\mathcal{P}_J$. Class $\mathcal{P}_J$ is less general than a class of densities $\tilde{\mathcal{P}}$ introduced by [30] when studying distinguishability of sets of distributions. $\tilde{\mathcal{P}}$ is used in [31] to obtain MDE of a density and its L_1-error rate in probability which, for normal densities with known variance, provides the $n^{-.5}$ rate for the estimate of the mean. This rate is not achievable with nonparametric methods based on parameter space discretization and the use of its metric entropy, due to a multiplicative factor $(\log n)^{\alpha}, \alpha > 0$, appearing because of the exponential bound used to evaluate the discrepancy of the empirical distribution from the population distribution/probability; see the convergence rates in Remark 4.5.

Definition 2.5 ([32]): Given a class $\mathcal{C}$ of subsets of a set V and a finite set U that is subset of V, let $\Delta^{\mathcal{C}}(U)$ be the number of different sets $A \cap U$ for $A \in \mathcal{C}$. Let

$$m^{\mathcal{C}}(n) = \max\{\Delta^{\mathcal{C}}(U) : U \text{ has } n \text{ elements}\},\ n = 1, 2, \ldots,$$

$$v(\mathcal{C}) = \begin{cases} \inf\{n : m^{\mathcal{C}}(n) < 2^n\} \\ \infty, \text{if } m^{\mathcal{C}}(n) = 2^n \text{ for all } n. \end{cases}$$

The class $\mathcal{C}$ will be called Vapnik–Chervonenkis (VC) class of exponent $v(\mathcal{C})$ if $v(\mathcal{C}) < \infty$.

3. Estimates $f_{\hat{p}_n}$ with $\hat{p}_n \in \mathcal{P}$ and convergence rates

Let $X_1, \ldots, X_n$ be a sample of d-dimensional vectors from unknown density $g \in \mathcal{G},\ d \geq 1$; $\mathcal{G}$ is a known family of densities, ρ is a distance for densities.

Definition 3.1: Let S_n be an estimate of $g(\in \mathcal{G})$.

a) S_n is uniformly consistent estimate for g in probability, with upper rate of convergence δ_n, if for every $\epsilon > 0$ there is $C(\epsilon)(> 1\ w.l.o.g.)$ such that

$$\sup_{g \in \mathcal{G}} P_g[\rho(S_n, g) > C(\epsilon)\delta_n] \leq \epsilon,\ \forall n \geq 1; \tag{13}$$

(13) is briefly denoted 'S_n has upper ρ-error rate, δ_n, in probability, $\rho(S_n, g) \leq C\delta_n$'.

b) The uniform upper risk rate of S_n is δ_n when there is constant $C_U(> 0)$ such that

$$\sup_{g \in \mathcal{G}} E_g \rho(S_n, g) \leq C_U \delta_n,\ \forall n \geq 1; \tag{14}$$

(14) is briefly denoted 'S_n has upper ρ-risk rate, δ_n, $E_g \rho(S_n, g) \leq C\delta_n$'.

P_g and E_g in (13), (14) denote, respectively, probability and expected value under g which is omitted in the sequel. $\mathcal{G}$ is determined from experts in the problem studied, in order to include the unknown density g; the rate of convergence δ_n in Definition 3.1 depends on $\mathcal{G}$.

For *any* estimate T_n of g in $\mathcal{G}$ with $T_n \notin \mathcal{G}$, a Minimum Distance Estimate (MDE) $\hat{g}_n \in \mathcal{G}$ is obtained with the same upper convergence rate as T_n.

Lemma 3.1: *Let $X_1, \ldots, X_n$ be a sample of d-dimensional vectors from unknown density g, element of a known family of densities $\mathcal{G}$, ρ is a distance; $d \geq 1$. Let T_n be an estimate of g, $T_n \notin \mathcal{G}$, such that the upper ρ-error rate of T_n is θ_n, either in probability or in risk.*

Define MDE $\hat{g}_n \in \mathcal{G}$:

$$\rho(T_n, \hat{g}_n) \leq inf\{\rho(T_n, g^*); g^* \in \mathcal{G}\} + \theta_n. \tag{15}$$

Then, the upper ρ-error rate of $\hat{g}_n$ is $3\theta_n$, either in probability or in risk.

Remark 3.1: The MDE $\hat{g}_n(\in \mathcal{G})$ always exists, whether achieving the value of the infimum, I_n, in (15) or another value in $(I_n, I_n + \theta_n]$; $\hat{g}_n$ is not necessarily unique and can be obtained, e.g. with ρ-discretization of $\mathcal{G}$. The difficulty of the minimization in (15) will depend on G and ρ; $\hat{g}_n$ is not an improvement of T_n but both $g, \hat{g}_n$ are elements of $\mathcal{G}$.

For the deconvolution problem (1)–(3), with $(\mathcal{G}, \rho)$ in (15) replaced by $(\mathcal{F}_{\mathcal{P}}, ||\cdot||_1)$ that is not totally bounded, kernel estimate $\hat{f}_n$ in (10) can be used for T_n in Lemma 3.1 to obtain *MDE* $f_{\hat{p}_n}$. However, for estimation on R^d in L_1-norm, [33] showed that consistent estimates cannot be built in a minimax sense. Under some additional condition, [34, Theorem 4, Remark 4] showed optimal estimators can be obtained that could be used as T_n in Lemma 3.1.

When $(\mathcal{F}_{\mathcal{P}}, ||\cdot||_1)$ is totally bounded, the intermediate estimate T_n is not needed. Direct approaches to obtain *MDE* $f_{\hat{p}_n}$ follow. The next Proposition from [35] is used as tool, with notation from the deconvolution problem (1)–(3) and taking into account its structure via $N_{\mathcal{F}_{\mathcal{P},d}}(a_n)$.

Proposition 3.1: *Let $X_1, \ldots, X_n$ be a sample in R^d from unknown density f_p, $p \in \mathcal{P}; d \geq 1$. Assume that $\mathcal{F}_{\mathcal{P},d} = \{f_p; p \in \mathcal{P}\}$ is L_1 totally bounded and let $N_{\mathcal{F}_{\mathcal{P},d}}(a_n)$ be the smallest number of L_1-balls of radius a_n needed to cover $\mathcal{F}_{\mathcal{P},d}$. Then, a MDE $f_{\hat{p}_n}$ can be constructed with upper L_1 error bound in probability*

$$C_1 a_n + C_2 \left(\frac{\log N_{\mathcal{F}_{\mathcal{P},d}}(a_n)}{n} \right)^{1/2}, \; C_1 > 0, \; C_2 > 0, \tag{16}$$

and upper-L_1-rate of convergence, a_n, to f_p in probability

$$a_n \sim \left(\frac{\log N_{\mathcal{F}_{\mathcal{P},d}}(a_n)}{n} \right)^{1/2}. \tag{17}$$

The centres of the $N_{\mathcal{F}_{\mathcal{P},d}}(a_n)$ balls covering $\mathcal{F}_{\mathcal{P},d}$ in Proposition 3.1 are $\mathcal{F}_{\mathcal{P},d}$'s elements and constitute an a_n-L_1-sieve which, in the deconvolution problem (1)–(3), depends on h and is used to construct $f_{\hat{p}_n}$. The Minimum Distance method to obtain $f_{\hat{p}_n}$ can be used also when model parameters, like h or the smoothness $\tilde{q}$ are not known. The unknown parameters are included in the *MDE* criterion and $N_{\mathcal{F}_{\mathcal{P},d}}(a_n)$ is increased. The approach is used in the next proposition with h assumed unknown, element of a family of densities $\mathcal{H}$, with f_p in (2) replaced by $f_{p,h}$ and $\mathcal{F}_{\mathcal{P}}$ in (3) replaced by

$$\mathcal{F}_{\mathcal{P},\mathcal{H}} = \{f_{p,h};\; p \in \mathcal{P}, h \in \mathcal{H}\}. \tag{18}$$

In the remaining Propositions in this section, the deconvolution structure (1),(2), (3) or (18) is used: in Propositions 3.2, 3.3 to show, respectively, $\mathcal{F}_{\mathcal{P},\mathcal{H}}$ and $\mathcal{F}_{\mathcal{P},d}$ are L_1-totally bounded and in Proposition 3.4 for the sets used in MDE to be Vapnik–Chervonenkis class.

Proposition 3.2: *In the deconvolution problem* (1),(2), (18), *assuming identifiability of $f_{p,h}$, let $X_1, \ldots, X_n$ be a sample from unknown density $f_{p,h}$, $p \in \mathcal{P}, h \in \mathcal{H}$. Assume that $\mathcal{P}$ and $\mathcal{H}$ are both L_1 totally bounded. Let $N_{\mathcal{P}}(a_n)$ and $N_{\mathcal{H}}(\xi_n)$ be, respectively the smallest numbers of L_1-balls of radius a_n and ξ_n needed to cover $\mathcal{P}$ and $\mathcal{H}$. Then, an MDE $\hat{h}_n * \hat{p}_n$ can be constructed with upper-L_1-rate of convergence* $\max\{a_n, \xi_n\}$ *to $f_{p,h}$ in probability, with*

$$a_n \sim \left(\frac{\log N_{\mathcal{P}}(a_n)}{n} \right)^{1/2}, \; \xi_n \sim \left(\frac{\log N_{\mathcal{H}}(\xi_n)}{n} \right)^{1/2}. \tag{19}$$

Remark 3.2: Upper error rates for $||f_{\hat{p}_n} - f_p||_1$, as those in Propositions 3.1 and 3.2, can be obtained also under weak dependence, with the mixing coefficient $\phi(r_n)$ appearing under the square-root in (16); with the proper choice of r_n the upper rate is the same with that in the *i.i.d.* case [36,37].

Particular examples of deconvolution problems (1)–(3) are now studied.

Proposition 3.3: *For the d-dimensional deconvolution problem* (1)–(3)*, assume that Y consists of d independent random variables, h is standard multivariate normal* $\mathcal{N}(0, I_d)$ *and* $\mathcal{P}$ *is the family of d-products of* $\tilde{q}$*-smooth densities, each having known support* $[-a, a], a \in R^+$; I_d *is unit matrix in* R^d, $d \geq 1$.

An MDE $f_{\hat{p}_n}$ *can be obtained with upper-*L_1*-rate of convergence* a_n *in probability,*

$$a_n \sim \{\frac{[\log(1/a_n)]^2}{n}\}^{1/2} \sim \frac{\log n}{\sqrt{n}}. \tag{20}$$

In the next proposition, p is element of family $\mathcal{P}_J$ described in Section 2 that has not been used often in the literature and a_n is obtained via [31], without using the metric entropy, $\log N_{\mathcal{F}_{\mathcal{P},d}}(a_n)$, as in (17).

Proposition 3.4: *For the deconvolution problem* (1)–(3) *in R, let h be such that* $Q(x, y) = h(x - y)$ *is Totally Positive (TP) and let* $\mathcal{P}_J$ *be a family of bounded and measurable densities that is* L_1*-separable (to avoid measurability problems), such that for every* p_1, p_2 *in* $\mathcal{P}_J$ *their difference* $(p_1 - p_2)$ *changes sign at most J times;* $0 < J < \infty$*. Assume also that the* σ*-finite measure determined via h and Lebesgue measure satisfy the conditions in Proposition* 2.1*. Then, an MDE* $f_{\hat{p}_n}$ *can be obtained with upper-*L_1*-rate of convergence in probability,*

(a) *when J is known,*

$$a_n \sim \frac{(\log n)^{.5}}{n^{.5}}, \tag{21}$$

(b) *when J is unknown,*

$$a_n \sim \frac{m_n^{.5}(\log n)^{.5}}{n^{.5}}, \tag{22}$$

with m_n *increasing to infinity as slow as it is wished.*

4. L_2-upper rates of convergence for $\hat{p}_n$, $\tilde{h} \neq 0$

For the deconvolution problem in R^d, let $X_1, \ldots, X_n$ be *i.i.d.* vectors with values in $\mathcal{X}(\subset R^d)$ and density f_p satisfying (2) with p defined on $\mathcal{Y}(\subset R^d)$; d ≥ 1. It is assumed that estimate $f_{\hat{p}_n}$ and an upper bound on the error $||f_{\hat{p}_n} - f_p||_1$ have been obtained, as described in the previous section, thus $\hat{p}_n$ is already available. The main result connecting this section with the previous is the upper bound for $||\hat{p}_n - p||_2$ in (26) that depends on $||f_{\hat{p}_n} - f_p||_1$, obtained via ψ_n introduced below.

The Assumptions:

($\mathcal{A}1$) h is known, $\tilde{h} \neq 0$, $||\tilde{h}||_2 < \infty$,
($\mathcal{A}2$) $\mathcal{Y}$ is compact,

($\mathcal{A}3$) $\mathcal{P}$ is the family of $\tilde{q}$-smooth densities (Definition 2.3),
($\mathcal{A}4$) $\mathcal{Y} \subset \mathcal{X} \subset R^d$, $d \geq 1$,
($\mathcal{A}5$) $f_{\hat{p}_n}$ is an estimate of f_p, obtained as described in Section 3, with upper L_1-error rate, a_n, in probability and in risk; $\hat{p}_n \in \mathcal{P}$.

Remark 4.1: When p is anisotropic, plug-in upper convergence rates can be obtained for $||\hat{p}_n^{(s)} - p^{(s)}||_2, [s] = 0, 1, \ldots, q$, that will depend on the least favourable anisotropic component. Assume, e.g. that $\mathcal{P}$ has q-derivatives and the modulus of continuity of $\frac{\partial^q p}{\partial x_i^q}$ is $w_{q,i}(x)$, $i = 1, \ldots, d$. Then, in (26) and thereafter $w_q(b_n)$ is replaced by $\max\{w_{q,i}(b_n); i = 1, \ldots, \mathrm{d}\}$, and wherever $q + \gamma$ appears it is replaced by $q + \min\{\gamma_i;\ i = 1, \ldots, \mathrm{d}\}$. When some of the parameters are unknown, e.g. either the smoothness q or the $\gamma_i, i = 1, \ldots, d$, or the deconvolution kernel h, discretization of the corresponding *enlarged* parameter space is used in the Minimum Distance Estimation criterion; see, e.g. Proposition 3.2.

Let $\tilde{h}$ and $\tilde{K}_n$ be, respectively, the Fourrier transforms of h and K_n; in (9), $K_n(x)$ is defined as $b_n^{-d} K(x b_n^{-1})$, with b_n a positive number to be determined. Since $\tilde{h} \neq 0$, let ψ_n be the inverse Fourrier transform of

$$\tilde{\psi}_n = \frac{\tilde{K}_n}{\tilde{h}}. \tag{23}$$

By the convolution theorem,

$$\psi_n * h = K_n. \tag{24}$$

An upper bound for $||\psi_n||_2$ is obtained. The set $[-M, M]^d$ is the support of $\tilde{K}$.

Lemma 4.1: *Under* ($\mathcal{A}1$),

$$\begin{aligned} ||\psi_n||_2 = C||\tilde{\psi}_n||_2 &\leq C \cdot \left[\int_{[-\frac{M}{b_n}, \frac{M}{b_n}]^d} |\tilde{K}(t b_n)|^2 |\tilde{h}(t)|^{-2} \mathrm{d}t \right]^{1/2} \\ &\leq C \cdot \frac{\sup_{t \in [-M/b_n, M/b_n]^d} |\tilde{h}(t)|^{-1}}{b_n^{5d}}. \end{aligned} \tag{25}$$

An upper bound for $||\hat{p}_n - p||_2$ is provided when $\tilde{h} \neq 0$.

Proposition 4.1: *Under assumptions* ($\mathcal{A}1$) $-$ ($\mathcal{A}5$),

$$\begin{aligned} ||\hat{p}_n - p||_2 &\leq C[b_n^q w_q(b_n) + ||\psi_n||_2 \, ||f_{\hat{p}_n} - f_p||_1] \\ &\leq C[b_n^q w_q(b_n) + \frac{\sup_{t \in [-M/b_n, M/b_n]^d} |\tilde{h}(t)|^{-1}}{b_n^{5d}} ||f_{\hat{p}_n} - f_p||_1]; \end{aligned} \tag{26}$$

$[-M, M]^d$ is $\tilde{K}$'s support, C is generic constant.

The next Lemma provides an upper bound for $||\hat{p}_n - p||_2$ in probability or in risk from the corresponding bound of $||f_{\hat{p}_n} - f_p||_1$.

Lemma 4.2: *If the upper ρ_1-error rate of $f_{\hat{p}_n}$ is a_n, in probability and/or in risk, and for the ρ_2-error of $\hat{p}_n$ and positive scalars λ_n, μ_n holds*

$$\rho_2(\hat{p}_n, p) \leq \lambda_n + \mu_n \rho_1(f_{\hat{p}_n}, f_p), \tag{27}$$

then, the upper ρ_2-error rate of $\hat{p}_n$ is $\lambda_n + \mu_n a_n$, respectively, in probability and/or in risk.

Remark 4.2: Usually, λ_n, μ_n in (27) depend on a parameter, e.g. b_n, to be determined such that $\lambda_n + \mu_n a_n$ is minimized. Lemma 4.2 can be used for any estimates S_n, U_n instead of $\hat{p}_n, f_{\hat{p}_n}$.

Models for h are now presented. Let $0 < C_1 \leq C_2 < \infty$, $|t| = (|t_1|, \ldots, |t_d|)$, $k > 0$, $\alpha_j \geq 0$, $\beta_j > .5, j = 1, \ldots, d$, $\bar{\alpha} = \frac{1}{d}\sum_{j=1}^d \alpha_j$, $\bar{\beta} = \frac{1}{d}\sum_{j=1}^d \beta_j$.

($\mathcal{M}1$) h is super-smooth when $\tilde{h} \neq 0$ and for large $|t|$-values, $d\bar{\alpha} > 0$,

$$C_1 \mathrm{e}^{-\sum_{j=1}^d \alpha_j |t_j|^k} \Pi_{j=1}^d |t_j|^{\beta_j} \leq |\tilde{h}(t_1, \ldots, t_d)| \leq C_2 \mathrm{e}^{-\sum_{j=1}^d \alpha_j |t_j|^k} \Pi_{j=1}^d |t_j|^{\beta_j}. \tag{28}$$

($\mathcal{M}2$) h is smooth when $\tilde{h} \neq 0$ and for large $|t|$-values

$$C_1 \Pi_{j=1}^d |t_j|^{-\beta_j} \leq |\tilde{h}(t_1, \ldots, t_d)| \leq C_2 \Pi_{j=1}^d |t_j|^{-\beta_j}. \tag{29}$$

Careful choice of b_n determines the least upper bound (26). When $\tilde{h}(t)$ varies exponentially as t increases, it determines the upper bound in (26). For algebraic variation of $\tilde{h}(t)$ as t increases, b_n satisfies

$$\frac{b_n^{q+.5d} w_q(b_n)}{\sup_{t \in [-M/b_n, M/b_n]^d} |\tilde{h}(t)|^{-1}} \sim a_n. \tag{30}$$

A small b_n-value satisfying (30) exists and is unique since when b_n decreases to zero, the numerator in the left side of (30) decreases to zero, the denominator increases to infinity and w_q, $\tilde{h}$ are continuous. Thus, the estimate $f_{\hat{p}_n}$ with the smallest rate a_n in ($\mathcal{A}5$) is preferred.

The upper error rates for $||\hat{p}_n - p||_2$ in probability and for $E||\hat{p}_n - p||_2$ are now given explicitly as function of a_n, the upper bound of $||f_{\hat{p}_n} - f_p||_1$ in probability or in risk in ($\mathcal{A}5$), for super-smooth and smooth h, using in (31)–(36) and in Examples 4.1 and 4.2 the brief notations for (a) and (b) in Definition 3.1. Note that only the lower bounds in (28) and (29) are used herein to obtain the upper error bound for $||\hat{p}_n - p||_2$.

Proposition 4.2: *Assume that ($\mathcal{A}1$) − ($\mathcal{A}5$) hold, in particular a_n is in ($\mathcal{A}5$).*

(i) For super-smooth h from model ($\mathcal{M}1$), an upper rate in probability on $\hat{p}_n$'s L_2-error is

$$||\hat{p}_n - p||_2 \leq C_{\bar{\alpha},d,k,M} \cdot (\log a_n^{-1})^{-q/k} w_q[C(\log a_n^{-1})^{-1/k}]. \tag{31}$$

When $w_q(b_n) = L \cdot b_n^{\gamma}, 0 < \gamma < 1, \tilde{q} = q + \gamma$,

$$||\hat{p}_n - p||_2 \leq C_{\bar{\alpha},d,k,M} \cdot (\log a_n^{-1})^{-\tilde{q}/k}. \tag{32}$$

The dimension d affects only constant $C_{\bar{\alpha},d,k,M}$.

(ii) For smooth h from model $(\mathcal{M}2)$, *an upper rate on* $||\hat{p}_n - p||_2$ *is obtained when* b_n *satisfies*

$$b_n^q w_q(b_n) \sim \frac{a_n}{b_n^{d\bar{\beta}+.5d}}.$$

When $w_q(b_n) = L \cdot b_n^{\gamma}$, $0 < \gamma < 1$, *an upper rate in probability on* $\hat{p}_n$*'s* L_2*-error is*

$$||\hat{p}_n - p||_2 \leq c_M a_n^{\tilde{q}/(\tilde{q}+d\bar{\beta}+.5d)}, \ \tilde{q} = q + \gamma. \tag{33}$$

(iii) When $E||f_{\hat{p}_n} - f_p||_1 \leq a_n$ *and* $w_q(b_n) = L \cdot b_n^{\gamma}$, *the upper rates in* (32) *and* (33) *hold also for* $E||\hat{p}_n - p||_2$.

Remark 4.3: Model $(\mathcal{M}1)$ can be enlarged, with k in (28) replaced by positive $k_j, j = 1, \ldots, d$. Then, upper bounds (31), (32) remain valid with $\max\{k_1, \ldots, k_d\}$ replacing k. In the proof, k of the upper bound in (A7) will be replaced by $\max\{k_1, \ldots, k_d\}$.

Upper rates on the L_2-error and risk of the derivative of $\hat{p}_n$ follow. Since $\hat{p}_n \in \mathcal{P}$, its derivative $\hat{p}_n^{(s)}$ is used to estimate $p^{(s)}$.

Corollary 4.1: *Assume* $(\mathcal{A}1) - (\mathcal{A}5)$ *hold,* δ_n *is the upper bound obtained in Proposition* 4.2, $w_q(b) = L \cdot b^{\gamma}, 0 < \gamma < 1$, $\tilde{q} = q + \gamma, L > 0$, $s = (s_1, \ldots, s_d)$ *is a d-tuple of non-negative integers,* $[s] = s_1 + \ldots + s_d \leq q$.
(i) *If* $||\hat{p}_n - p||_2 \leq \delta_n$ *in probability, then in probability*

$$||\hat{p}_n^{(s)} - p^{(s)}||_2 \leq C \cdot \delta_n^{\frac{\tilde{q}-[s]}{\tilde{q}}}. \tag{34}$$

(ii) *If the upper rate of* $E||\hat{p}_n - p||_2$ *is* δ_n, *then*

$$E||\hat{p}_n^{(s)} - p^{(s)}||_2 \leq C \cdot \delta_n^{\frac{\tilde{q}-[s]}{\tilde{q}}}. \tag{35}$$

The next result indicates the reason that, when h is super-smooth, estimates of p and $p^{(s)}$ are frequently minimax optimal.

Corollary 4.2: *Under the assumptions in Proposition* 4.2 *(a) (i) and Corollary 4.1 and if* $||f_{\hat{p}_n} - f_p||_1 \sim n^{-\delta}$ *in probability,* $0 < \delta < 1$,

$$||\hat{p}_n^{(s)} - p^{(s)}||_2 \leq C_{\tilde{\alpha},d,k,M}(\delta \log n)^{-(\tilde{q}-[s])/k}, \ [s] \geq 0. \tag{36}$$

If $E||\hat{f}_n - f_p||_1 \sim n^{-\delta}$, *the upper bound in* (36) *is valid for the risk* $E||\hat{p}_n^{(s)} - p^{(s)}||_2$.

Remark 4.4: When $d = 1$, $\hat{p}_n^{(s)}$ is risk minimax optimal for any $\delta > 0$ for the weighted L_2-distance [6] and the L_2-distance (see [9, e.g.]). The same holds for $d > 1$; see, e.g. [2], Theorem 3, Case B.

Remark 4.5: We searched the literature for density estimates of f_p.

For p defined on a compact subset in R, estimates for location-scale Gaussian mixtures have Hellinger error rates in probability $(\log n)^{\zeta}/n^{\delta}$, $0 < \delta \leq .5$, $\zeta > 0$ [38–40]. From (4)

these bounds hold also for L_1-distance and estimates with form $f_{\hat{p}_n}$ can be obtained via Lemma 3.1, with the same upper L_1-error rates. These rates and additional results in the literature, e.g. ([41, Theorems 2.1 and 2.2 and for a slowly increasing sequence of compact domains with lengths affecting multiplicatively the rates]), as well as (20)–(22) suggest to use $a_n \sim n^{-1/2}(\log n)^{\zeta},\ 0 < \zeta$.

Example 4.1: Assume $(\mathcal{A}1) - (\mathcal{A}5)$ with $a_n \sim n^{-1/2}(\log n)^{\zeta}$ in probability, $d = 1, w_q(b) = L{\cdot}b^{\gamma}, \gamma > 0, \tilde{q} = q + \gamma$. Then:
(a) for h the standard normal, $\tilde{h}(t) \sim \mathrm{e}^{-t^2}$ for large $|t|$, and from (32), (34) in probability

$$||\hat{p}_n^{(s)} - p^{(s)}||_2 \leq C(\log n)^{(\tilde{q}-[s])/2},\ [s] \geq 0.$$

(b) for h the Cauchy, $\tilde{h}(t) \sim \mathrm{e}^{-|t|}$ for large $|t|$, and from (32), (34) in probability

$$||\hat{p}_n^{(s)} - p^{(s)}||_2 \leq C(\log n)^{\tilde{q}-[s]},\ [s] \geq 0.$$

(c) for h the exponential, $\tilde{h}(t) \sim |t|^{-\beta}$ for large $|t|$, and from (33) in probability

$$||\hat{p}_n^{(s)} - p^{(s)}||_2 \leq C\frac{(\log n)^{\xi}}{n^{(\tilde{q}-[s])/(2\tilde{q}+2\beta+1)}},\ \xi = \zeta(\tilde{q} - [s])/(\tilde{q} + \beta + .5),\ [s] \geq 0.$$

The bound in (c) misses by the factor $(\log n)^{\xi}$ the weighted L_2-minimax rate [6] and the L_2-minimax rate (see, e.g. [9]).

The bounds in (a)–(c) remain valid when a_n is the risk rate.

Example 4.2: For the d-dimensional deconvolution in Proposition 3.3, the upper rate of convergence in probability of $f_{\hat{p}_n}$ to f_p is $\frac{\log n}{n^{.5}}$. Thus, the upper L_2-rates of convergence in probability to $p^{(s)}$ for super-smooth and smooth h are, respectively, $(\log n)^{(q-[s])/k}$ and $(\log n)^{\frac{2(\tilde{q}-[s])}{2\tilde{q}+2\sum_{i=1}^{d}\beta_i+d}} \cdot n^{-\frac{q-[s]}{2\tilde{q}+2\sum_{i=1}^{d}\beta_i+d}},\ [s] \geq 0.$

Remark 4.6: When h is smooth, we compare the upper L_2-risk rate herein with that of the lower L_u-risk bound provided by [3, p. 892–895] for the isotropic and bounded Nikolskii class, $\mathcal{N}_{r,d}(\tilde{q}, L)$, in the generalized deconvolution with density of the X's in R^d having form $(1-\alpha)p + \alpha(h * p); 0 \leq \alpha \leq 1,\ \mathrm{d} \geq 1,\ 2 \leq u < \infty, r$ is d-vector $(u, u, \ldots, u)$, $\tilde{q}$ and L as defined in (7). The rate of the lower bound is $\delta_n^{-\rho(\alpha)}$; $\rho(\alpha)$ depends on parameters $\beta(\alpha)$ and $\omega(\alpha)$ and on whether a parameter $\kappa_\alpha(u)$ is larger than $u \cdot \omega(\alpha)$ or not. With our notation and for $\alpha = 1$ that corresponds to the problem herein,

$$\beta(1) = \frac{\tilde{q}}{2\sum_{j=1}^{d}\beta_j + d}, \quad \omega(1) = \frac{u\tilde{q}}{2\sum_{j=1}^{d}\beta_j + d}$$

and

$$\kappa_1(u) = \omega(1)[2 + \frac{1}{\beta(1)}] - u = \frac{2u\tilde{q}}{2\sum_{j=1}^{d}\beta_j + d} = 2\omega(1).$$

Since $\kappa_1(u)$ is positive and $\kappa_1(u) \leq u\omega(1)$, for $u \geq 2$,

$$\rho(1) = \frac{\beta(1)}{2\beta(1)+1} = \frac{\tilde{q}}{2\sum_{j=1}^{d}\beta_j + d}/(2\frac{\tilde{q}}{2\sum_{j=1}^{d}\beta_j + d} + 1) = \frac{\tilde{q}}{2\tilde{q} + 2\sum_{j=1}^{d}\beta_j + d}.$$

The rate of the L_u-lower bound is $n^{-\frac{\tilde{q}}{2\tilde{q}+2\sum_{j=1}^{d}\beta_j+d}}$, $2 \le u < \infty$. When $a_n \sim (\log n)^{\zeta} n^{-.5}$, the rate of the plug-in upper L_2-error bound herein is $[\frac{(\log n)^{2\zeta}}{n}]^{-\frac{\tilde{q}}{2\tilde{q}+2\sum_{j=1}^{d}\beta_j+d}}$, missing the lower bound by a power of $\log n$. However, the exponents in both bounds coincide.

Acknowledgments

Many thanks are due to Professor Alexander Meister, Editor, the AE and a referee for comments and suggestions that improved the presentation of this work. In particular, very many thanks are due to the second referee, for comments and suggestions that have led to the improvement of the content of the paper and whose deep knowledge in Mathematical Statistics and sharp thinking made the refereeing process fun! Special thanks are also due to Miss Eleni Yatracos for improvements in the English writing. Many thanks, as usual, are due to the Department of Statistics and Applied Probability, National University of Singapore, for the warm hospitality during my summer visits where most of these results were obtained.

Disclosure statement

No potential conflict of interest was reported by the author.

References

[1] Devroye L. A note on the usefulness of superkernels in density estimation. Ann Statist. 1992;20:2037–2056.
[2] Comte F, Lacour C. Anisotropic adaptive kernel deconvolution. Ann Inst H Poincaré Probab Statist. 2013;49:569–609.
[3] Lepski OV, Willer T. Lower bounds in the convolution structure density model. Bernoulli. 2017;23:884–926.
[4] Rebelles G, Structural adaptive deconvolution under L_p-losses. *arXiv:1504.06246v2*.2015.
[5] Fan J. Adaptive local one-dimensional sub-problems with application to a deconvolution problem. Ann Statist. 1993;21:600–610.
[6] Fan J. Deconvolution with supersmooth distributions. Can J Stat. 1992;20(2):155–169.
[7] Fan J. On the optimal rates of convergence for nonparametric deconvolution problems. Ann Statist. 1991;19:1257–1272.
[8] Hall P, Meister A. A ridge-parameter approach to deconvolution. Ann Stat. 2007;35:1535–1558.
[9] Meister A. Deconvolution problems in nonparametric statistics. Berlin: Springer-Verlag; 2009. (Lecture notes in statistics; 193).
[10] Lounici K, Nickl R. Global uniform risk bound for wavelet deconvolution estimators. Ann Statist. 2011;39:201–231.
[11] Masry E. Multivariate probability density deconvolution for stationary random processes. IEEE Trans Inf Theory. 1991;37:1105–1115.
[12] Youndjé E, Wells MT. Optimal bandwidth selection for multivariate kernel deconvolution density estimation. TEST. 2008;17:138–162.
[13] Carrol RJ, Hall P. Optimal rates of convergence for deconvolving a density. J Am Stat Assoc. 1988;83:1184–1186.
[14] Devroye L. Consistent deconvolution in density estimation. Can J Stat. 1989;17:235–239.
[15] Hesse CH. Deconvolving a density from partially contaminated observations. J Multivariate Anal. 1995;55:246–260.
[16] Loh W- L, Zhang C-H. Global properties of kernel estimators for mixing densities in exponential family models for discrete variables. Stat Sin. 1996;6:561–678.
[17] Loh W- L, Zhang C-H. Estimating mixing densities in exponential family models for discrete variables. Scand J Stat. 1997;24:15–32.

[18] Pensky M, Vidakovic B. Adaptive wavelet estimator for nonparametric density deconvolution. Ann Stat. 1999;27:2033–2053.
[19] Stefanski LA, Carroll RJ. Deconvoluting kernel density estimators. Statistics. 1990;21:169–184.
[20] Zhang C-H. Fourier methods for estimating mixing densities and distributions. Ann Stat. 1990;18:806–831.
[21] Butucea C, Tsybakov AB. Sharp optimality in density deconvolution with dominating bias. I, II. Theory Probab Appl. 2007;52(24–39):237–249.
[22] Delaigle A, Gijbels I. Estimation of integrated squared density derivatives from a contaminated sample. J Royal Stat Soc B. 2002;64:869–886.
[23] Groeneboom P, Jongbloed G. Density estimation in the uniform deconvolution model. Statist Nederlandica. 2003;57:136–157.
[24] Meister A. Density estimation with normal measurement error with unknown variance. Stat Sin. 2006;16:195–211.
[25] Johannes J. Deconvolution with unknown error distribution. Ann Stat. 2009;37:2301–2323.
[26] Meister A. Deconvolution from Fourrier-oscillating error densities under decay and smoothness restrictions. Inverse Probl. 2008;24(1):015003.
[27] Devroye L, Personal communication. 2013.
[28] Karlin S. Total positivity vol. I. Stanford (CA): Stanford University Press; 1968.
[29] Karlin S. Total positivity, absorption probabilities and applications. Trans Am Math Soc. 1964;111:33–107.
[30] Hoeffding W, Wolfowitz J. Distinguishability of sets of distributions. Ann Math Statist. 1958;29:700–718.
[31] Yatracos YG. A note on L_1 consistent estimation. Can J Stat. 1988;16:283–292.
[32] Vapnik VN, Chervonenkis AY. On the uniform convergence of relative frequencies of events to their probabilities. Theory Probab Appl. 1971;16:264–280.
[33] Ibragimov IA, Khasminski RZ. More on estimation of the density of a distribution. Zap Nauchn Sem Leningrad Otdel Mat Inst Steklov (LOMI). 1981;108:72–88. (in Russian)
[34] Goldenshluger A, Lepski OV. On adaptive minimax density estimation on R^d. Probab Theory Related Fields. 2014;159:479–543.
[35] Yatracos YG. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. Ann Statist. 1985;13:768–774.
[36] Roussas GG, Yatracos YG, Minimum distance estimates with rates under ϕ-mixing. In: Pollard D, Torgersen E, Yang GL, editors. Festschrift for Lucien Le Cam: research Papers in probability and statistics. New York (NY): Springer; 1997. p. 337–345.
[37] Roussas GG, Yatracos YG. Minimum distance regression-type estimates with rates under weak dependence. Ann Inst Stat Math. 1996;48(2):267–281.
[38] Genovese CR, Wasserman L. Rates of convergence for the Gaussian mixture sieve. Ann Stat. 2000;28:1105–1127.
[39] Ghosal S, van der Vaart AW. Entropy and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. Ann Stat. 2001;29:1233–1263.
[40] Zhang C-H. Generalized maximum likelihood estimation of normal mixture densities. Stat Sin. 2009;19:1297–1318.
[41] Ibragimov IA. Estimation of analytic functions. Hayward (CA): IMS; 2001. (State of the art in probability and statistics. Festschrift for W. R. van Zwet.
[42] Lorentz GG. Approximation of functions. New York (NY): Chelsea Publishing Company; 1986.
[43] Yatracos YG. On the estimation of the derivatives of a function with the derivatives of an estimate. J Multivariate Anal. 1989;28:172–175.

Appendix.

Proof of Lemma 3.1.: From (15), for the estimate $\hat{g}_n \in \mathcal{G}$ it holds

$$\rho(\hat{g}_n, g) \leq \rho(\hat{g}_n, T_n) + \rho(T_n, g) \leq inf\{\rho(T_n, g^*); g^* \in \mathcal{G}\} + \theta_n + \rho(T_n, g) \leq 2\rho(T_n, g) + \theta_n. \tag{A1}$$

When T_n's risk is bounded by θ_n, it follows from (15) that

$$\sup_{g\in\mathcal{G}} E\rho(\hat{g}_n, g) \le 3\theta_n.$$

For the upper rate in probability, for $\epsilon > 0$ let $C(\epsilon)(> 1)$ be the constant in (13) such that

$$\sup_{g\in G} P[\rho(T_n, g) > C(\epsilon)\theta_n] \le \epsilon.$$

Then, from (A1),

$$\sup_{g\in G} P[\rho(\hat{g}_n, g) > C(\epsilon)3\theta_n] \le \sup_{g\in G} P[2\rho(T_n, g) + \theta_n > C(\epsilon)3\theta_n] \le \sup_{g\in G} P[\rho(T_n, g) > C(\epsilon)\theta_n]. \quad \blacksquare$$

Proof of Proposition 3.1.: Follows from [35, Theorem 1]. ■

Proof of Proposition 3.2.: Let $h_1^*, \ldots, h_{N_{\mathcal{H}}(\xi_n)}^*$ be a ξ_n-L_1-sieve for $\mathcal{H}$ and $p_1^*, \ldots, p_{N_{\mathcal{P}}(a_n)}^*$ be a a_n-L_1-sieve for $\mathcal{P}$. For $h \in \mathcal{H},\ p \in \mathcal{P}$, let $h_i^*,\ p_k^*$ be such that

$$||h - h_i^*||_1 \le \xi_n,\ ||p - p_k^*||_1 \le a_n. \tag{A2}$$

Then, using (A2) and Young's inequality it follows that,

$$||h * p - h_i^* * p_k^*||_1 \le ||h * p - h_i^* * p||_1 + ||h_i^* * p - h_i^* * p_k^*||_1 \le \xi_n + a_n.$$

Thus,

$$\{h_i^* * p_k^*, 1 \le i \le N_{\mathcal{H}}(\xi_n), 1 \le k \le N_{\mathcal{P}}(a_n)\}$$

is a $(a_n + \xi_n)$-L_1-sieve with cardinality $N_{\mathcal{P}}(a_n) \cdot N_{\mathcal{H}}(\xi_n)$ in the space

$$\{h * p,\ h \in \mathcal{H}, p \in \mathcal{P}\}.$$

Thus, from (16) the upper bound in probability of the *MDE* $\hat{h}_n * \hat{p}_n$ is

$$C_1(a_n + \xi_n) + C_2\left(\frac{\log[N_{\mathcal{P}}(a_n) \cdot N_{\mathcal{H}}(\xi_n)]}{n}\right)^{1/2}$$
$$\le C\left[a_n + \xi_n + \left(\frac{\log N_{\mathcal{P}}(a_n)}{n}\right)^{1/2} + \left(\frac{\log N_{\mathcal{H}}(\xi_n)}{n}\right)^{1/2}\right]. \quad \blacksquare$$

Proof of Proposition 3.3.: Since $\mathcal{P}$ is sup-norm compact [42, p. 153], it is also L_1-totally bounded and by Young's inequality, $\mathcal{F}_{\mathcal{P},d}$ is also L_1-totally bounded. Thus, for every $a_n > 0$ there is a Ca_n-L_1-sieve of densities in $\mathcal{F}_{\mathcal{P},d}$. An upper bound for the $\log N_{\mathcal{F}_{\mathcal{P},d}}(a_n)$ is obtained using the a_n-L_1-sieve for p continuous on $[-a, a]$, with the logarithm of the sieve's cardinality bounded by $C_1[\log(1/a_n)]^2$ ([39], Theorem 3.1, p. 1240, with known a, $\sigma = 1$ and $\gamma = .5$). In every a_n-L_1 ball with centre in this sieve, replace the centre by a density in $\mathcal{F}_{\mathcal{P},1}$ from the same ball, if it exists. The so-obtained densities are a $(2a_n)$-L_1 sieve for $\mathcal{F}_{\mathcal{P},1}$ with cardinality bounded by $C_1[\log(1/a_n)]^2$. Thus, d-products of these densities are a C_2a_n-L_1 sieve of $\mathcal{F}_{\mathcal{P},d}$ and

$$\log N_{\mathcal{F}_{\mathcal{P},d}}(C_2a_n) \le c[\log(1/a_n)]^2. \tag{A3}$$

The rate (20) follows from Proposition 3.1. ■

Proof of Proposition 3.4.: (a) Consider the L_1-separable subset $\mathcal{P}_J^* = \{p_1^*, p_2^*, \ldots, p_n^*, \ldots,\}$ of $\mathcal{P}_J$. For every $p \in \mathcal{P}_J$ denote by F_p the probability measure with density f_p. For $\beta_n = \frac{(\log n)^{.5}}{n^{.5}}$, there is

$p^* \in \mathcal{P}_J^*$ such that $||p - p^*||_1 \le \beta_n$ and $||f_p - f_{p^*}||_1 \le \beta_n$. If

$$\mathcal{S}_J = \{ \{x : h * p_i^*(x) > h * p_j^*(x)\},\ i \ne j\}, \tag{A4}$$

then, for $p_1,\ p_2 \in \mathcal{P}_J$,

$$||f_{p_1} - f_{p_2}||_1 \le 2\beta_n + ||f_{p_1^*} - f_{p_2^*}||_1 \le 2\beta_n + 2 \sup_{A \in \mathcal{S}_J} |F_{p_1^*}(A) - F_{p_2^*}(A)|$$

$$\le 6\beta_n + 2 \sup_{A \in \mathcal{S}_J} |F_{p_1}(A) - F_{p_2}(A)|. \tag{A5}$$

By Total Positivity of h for any r and from Proposition 2.1 for every $i \ne j, h * (p_i - p_j)$ has at most J changes of sign and the sets in $\mathcal{S}_J$ are unions of at most J disjoint intervals, thus $\mathcal{S}_J$ is VC class with exponent $2J + 1$. From [31, Theorem 2, p. 287, with $a_n = \beta_n,\ l_k = 2,\ v_k = 2J + 1,\ F_{a_k} = \mathcal{S}_J$] it follows that the upper L_1 rate of convergence of $f_{\hat{p}_n}$ is

$$a_n \sim \frac{(\log n)^{.5}}{n^{.5}}.$$

(b) When J is unknown, consider L_1-separable families of densities $\mathcal{P}_I$ with the same properties as $\mathcal{P}_\mathcal{J}$ and I the maximum number of sign changes for the densities' differences; $I \ge 1$. Observe that $\mathcal{P}_I \subset \mathcal{P}_L,\ I \le L$. Let I_n increase to infinity with n and assume *w.l.o.g* that it takes integer values. For $n \ge n_0,\ \mathcal{P}_J \subset \mathcal{P}_{I_n}$ and for $p_1, p_2 \in \mathcal{P}_{I_n}$, (A5) holds with $\beta_n = \frac{(2I_n+1)^{.25}(\log n)^{.5}}{n^{.5}}$ and S_J replaced by S_{I_n} with VC-exponent $2I_n + 1$. From [31, Theorem 2] it follows that the upper L_1 rate of convergence of $f_{\hat{p}_n}$ is

$$a_n \sim \frac{(2I_n + 1)^{.5}(\log n)^{.5}}{n^{.5}}.$$

■

Proof of Lemma 4.2.: By taking expected values in both sides of (27) and the supremum over all $p \in \mathcal{P}$ the upper bound follows. For the upper bound in probability, let $\epsilon > 0$, and let $C(\epsilon)(> 1)$ be such that

$$\sup_{p \in \mathcal{P}} P[\rho_1(f_{\hat{p}_n}, f_p) > C(\epsilon)\delta_n] \le \epsilon.$$

Then,

$$P[\rho_2(\hat{p}_n, p) > C(\epsilon)(\lambda_n + \mu_n\delta_n)] \le P[\lambda_n + \mu_n\rho_1(f_{\hat{p}_n}, f_p) > C(\epsilon)(\lambda_n + \mu_n\delta_n)]$$

$$\le P[\rho_1(f_{\hat{p}_n}, f_p) > C(\epsilon)\delta_n]$$

and the result follows by taking the supremum for $p \in \mathcal{P}$. ■

Lemma A.3: *Let g be a function defined on a set $\mathcal{Y}$ in R^d that has all s-mixed order partial derivatives uniformly bounded for $0 \le [s] \le q$, with the q-th derivative having modulus of continuity w_q. Then, for the kernel K satisfying* (8), *K_n defined in* (9) *and $\mathcal{Y}$ compact,*

$$||g - K_n * g||_u \le c b_n^q w_q(b_n),\ c > 0,\ u \ge 1. \tag{A6}$$

Proof of Lemma A.3.: The result follows from [43, p. 173, Proposition 1]. ■

Proof of Lemma 4.1.: For the Fourrier transform $\tilde{K}_n(x)$ it holds,

$$\tilde{K}_n(x) = C \int \mathrm{e}^{-ixy} b_n^{-d} K(y/b_n)\mathrm{d}y = C \int \mathrm{e}^{i(xb_n)yb_n^{-1}} K(yb_n^{-1}) d(yb_n^{-1}) = C\tilde{K}(xb_n).$$

Boundedness of $\tilde{K}$ and Parseval's identity imply that

$$||\psi_n||_2 = C||\tilde{\psi}_n||_2 = C\left[\int_{[-M/b_n, M/b_n]^d} \frac{|\tilde{K}(b_n t)|^2}{|\tilde{h}(t)|^2}\mathrm{d}t\right]^{.5} \le C\frac{\sup_{t \in [-M/b_n, M/b_n]^d} |\tilde{h}(t)|^{-1}}{b_n^{.5d}}.$$ ■

Proof of Proposition 4.1.:

$$\left[\int_{\mathcal{Y}} |\hat{p}_n(y) - p(y)|^2 \mathrm{d}y\right]^{1/2} \leq \left[\int_{\mathcal{Y}} |\hat{p}_n(y) - K_n * \hat{p}_n(y)|^2 \mathrm{d}y\right]^{1/2}$$
$$+ \left[\int_{\mathcal{X}} |K_n * \hat{p}_n(x) - K_n * p(x)|^2 \mathrm{d}x\right]^{1/2} + \left[\int_{\mathcal{Y}} |K_n * p(y) - p(y)|^2 \mathrm{d}y\right]^{1/2}$$
$$\leq C b_n^q w_q(b_n) + ||\psi_n * h * (\hat{p}_n - p)||_2 \leq C b_n^q w_q(b_n) + ||\psi_n||_2 \cdot ||f_{\hat{p}_n} - f_p||_1.$$

The first inequality is due to the triangular property of the $||\cdot||_2$-distance and to $\mathcal{Y} \subset \mathcal{X}$. The second inequality is due to Lemma A.3 and (24). The third inequality follows from Young's inequality for convolutions. The result follows from Lemma 4.1. ■

Proof of Proposition 4.2.: (i) When h follows the super-smooth model (28), the second term in the upper bound (26) has an exponential rate but the first term decreases at algebraic rate. Since

$$\sup_{t \in [-M/b_n, M/b_n]^d} |\tilde{h}(t)|^{-1} \leq C \cdot \mathrm{e}^{\sum_{j=1}^d \alpha_j M^k b_n^{-k}} \leq C \cdot \mathrm{e}^{\mathrm{d}\bar{\alpha} M^k b_n^{-k}}, \tag{A7}$$

the second term in upper bound (26) converges to zero as n increases if

$$\lim_{n\to\infty} \frac{\exp\{d\bar{\alpha} M^k b_n^{-k}\}}{b_n^{.5d}} a_n = 0 \iff \lim_{n\to\infty} d\bar{\alpha} M^k b_n^{-k} - .5d \log b_n - \log a_n^{-1} = -\infty. \tag{A8}$$

Choosing

$$b_n^k = \frac{4d\bar{\alpha} M^k}{\log a_n^{-1}} \quad \text{or} \quad b_n = \frac{(4d\bar{\alpha})^{1/k} M}{(\log a_n^{-1})^{1/k}}$$

(A8) holds and from Lemma 4.2 the terms in upper bound (26) are

$$b_n^q w_q(b_n) \sim (\log a_n^{-1})^{-q/k} w_q[C(\log a_n^{-1})^{-1/k}], \tag{A9}$$

$$\frac{\sup_{t\in[-M/b_n, M/b_n]^d} |\tilde{h}(t)|^{-1}}{b_n^{.5d}} a_n \leq a_n^{3/4} (\log a_n^{-1})^{5d/k}, \tag{A10}$$

with (A10) converging faster to 0 as n increases than (A9).

When $w_q(b_n) = L \cdot b_n^\gamma$, (A9) determines the upper convergence rate $(\log a_n^{-1})^{-(q+\gamma)/k}$.

(ii) When h follows the smooth model (29), both terms in upper bound (26) have algebraic rate. Since

$$\sup_{t \in [-M/b_n, M/b_n]^d} |\tilde{h}(t)|^{-1} \leq C \cdot \left(\frac{M}{b_n}\right)^{d\bar{\beta}}$$

and from Lemma 4.2 we choose b_n such that

$$b_n^q w_q(b_n) \sim \frac{a_n}{b_n^{d\bar{\beta}+.5d}}.$$

When $w_q(b_n) = L \cdot b_n^\gamma$, $\tilde{q} = q + \gamma$,

$$b_n^{\tilde{q}} \sim \frac{1}{b_n^{d\bar{\beta}} \cdot b_n^{5d}} a_n \text{ or } b_n \sim a_n^{1/(\tilde{q}+d\bar{\beta}+.5d)} \tag{A11}$$

and

$$||\hat{p}_n - p||_2 \leq c_M a_n^{\tilde{q}/(\tilde{q}+d\bar{\beta}+.5d)}.$$

(iii) Follows using the approach in (i) and (ii). ■

Proof of Corollary 4.1.: Follows along the lines in [43], Proposition 2, p. 174 and Remarks (i) and (ii) pages 174, 175, since p and $p^{(s)}$ have compact support. ■

Proof of Corollary 4.2.: The bounds are obtained by plugging $a_n \sim n^{-\delta}$ in the bounds in Proposition 4.2 (a) (i) and in (34) and (35). For densities in R, optimality for any $\delta > 0$ follows from the optimal rates in [5–7]. ■