The Annals of Statistics
1998, Vol. 26, No. 1, 335–362

LOG-DENSITY ESTIMATION IN LINEAR INVERSE PROBLEMS[1]

BY JA-YONG KOO AND HAN-YEONG CHUNG

Hallym University

We estimate a probability density function p which is related by a linear operator K to a density function q in sequences of regular exponential families based on a random sample from q. In this paper deconvolution and positron emission tomography are considered. The logarithm of the density function is approximated by basis functions consisting of singular functions of K. While direct maximum likelihood (or minimum Kullback–Leibler) density estimation in exponential families selects the parameters to match the moments of the basis functions to the sample moments, in the inverse problem the moment of each singular function is related to a corresponding moment of the direct problem by a factor given by a singular value λ_ν of K. Thus an appropriate analogue of the maximum likelihood estimate is obtained by matching moments with respect to p to $1/\lambda_\nu$ times the empirical moments associated with the sample from q. Bounds on the Kullback–Leibler distance between the true density and the estimators are obtained and rates of convergence are established for log-density functions having a measure of smoothness. The density estimator converges to the unknown density in the Kullback–Leibler sense and in the L_2-sense at a rate determined not only by the order of smoothness of the log-density and the dimension of data but also by the decay rate of the singular values of the operator. A minimax lower bound for deconvolution is provided under certain conditions. Numerical examples using simulated data are provided to illustrate the finite-sample performance of the proposed method for deconvolution and positron emission tomography.

1. Introduction. Suppose we observe a random sample $Y_1, \dots, Y_n$ from a density function $q(y)$, $y \in \mathscr{D} \subseteq \mathscr{R}^d$, which is related by a linear operator K to a density function $p(x)$, $x \in \mathscr{B} \subseteq \mathscr{R}^d$, that we wish to estimate. The linear operator equation $q = Kp$ is usually represented by an integral equation

$$q(y) = \int k(y, x) p(x)\, dx, \tag{1}$$

where k is known. In this paper, we will consider two interesting problems: deconvolution (where q is the convolution of p with a known density k) and positron emission tomography (PET) (where q is the Radon transform of p). This kind of indirect problem is referred to as a statistical inverse problem.

Received July 1994; revised March 1997.
[1]Supported by Non Directed Research Fund, Korea Research Foundation.
AMS 1991 *subject classifications*. 62G05, 62G07.
Key words and phrases. Linear inverse problems, SVD, exponential family, indirect likelihood, MILE, Kullback–Leibler distance, rate of convergence.

For inverse problems related to the Fredholm integral equation of the first kind, this is usually the case. The problem of solving such equations is often difficult since in cases which are of most interest scientifically, K is not invertible; that is, K^{-1} does not exist as a bounded linear operator so that a small perturbation of q may result in a large distortion of the solution p. These inverse problems are called ill-posed and this makes our inverse problem somewhat more difficult.

Consider the direct problem where the main interest is to estimate q. The approximation of log-densities in direct problems has been considered by many people. Related works include Neyman (1937), Crain (1974, 1976a, b, 1977), Leonard (1978), Silverman (1982), Mead and Papanicolaou (1984), Stone and Koo (1986), O'Sullivan (1988), Stone (1989, 1990, 1994), Kooperberg and Stone (1991, 1992), Barron and Sheu (1991; [BS] hereafter), Kooperberg (1995), Koo (1996) and Koo and Kim (1996). Estimates of density functions based on exponential families have an advantage as they are automatically positive and integrate to 1. For other traditional methods of nonparametric density estimation, such as kernel estimators and orthogonal series expansions of the density rather than the log-density, refer to Devroye and Györfi (1985) and Silverman (1986).

There is considerable interest in statistical inverse problems; the following literature on statistical inverse problems may not be a complete list. Deconvolution has been considered by Mendelsohn and Rice (1982), Carroll and Hall (1988), Stefanski and Carroll (1990), Fan (1991, 1993) and Efromovich (1997). PET (positron emission tomography) has been considered by Vardi, Shepp and Kaufman (1985), Jones and Silverman (1989), Johnstone and Silverman (1990, 1991), Bickel and Ritov (1995) and O'Sullivan (1995). Donoho (1993) considered wavelet methods for recovery of objects, such as signals, densities and spectra, from noisy and indirect data; Donoho (1994) addressed the minimax risk in estimating a linear functional of an unknown object from indirect data; Donoho (1995) developed a wavelet–vaguelette decomposition (WVD) for linear inverse problems. Silverman, Jones, Nychka and Wilson (1990), Vardi and Lee (1993), Eggermont and LaRiccia (1995) and Koo and Park (1996) applied the EM algorithm to linear inverse problems. O'Sullivan (1986), Nychka and Cox (1989), Koo (1993) and Kolaczyk (1996) studied linear inverse problems in regression frameworks.

The approach taken here is to seek a solution with p in an exponential family determined by functions from the singular-value decomposition (SVD) of the operator K. In this way positivity and integrability (to 1) of the estimate are ensured and it is possible to determine the convergence rates for sufficiently smooth densities p following the general method of [BS]. The difference is that while direct maximum likelihood (or minimum Kullback–Leibler) density estimation in exponential families selects the parameters to match the moments of the basis functions to the sample moments, in the inverse problem the moment of each singular function is related to a corresponding moment of the direct problem by a factor given by a singular value

λ_ν of the operator K. Thus an appropriate analogue of the maximum likelihood estimate is obtained by matching moments with respect to p to $1/\lambda_\nu$ times the empirical moments associated with the sample from q.

It is shown that the proposed density estimator $\hat{p}_n$ converges to p in the Kullback–Leibler sense at a rate determined not only by the order of smoothness r of $\log p$ and the dimension d of x but also by the decay rate s of the singular values. A minimax lower bound is provided for deconvolution to show our estimator is asymptotically optimal, where the optimal rate of convergence has the form $n^{-2r/(2r+2s+1)}$ or $(\log n)^{-2r/s}$ accordingly as the characteristic function of the contaminating noise decays algebraically or exponentially. In the case of PET, the rate has the form $n^{-2r/(2r+3)}$ or $n^{-r/(r+2)}$ depending on the smoothness condition for $\log p$.

Simulation results for deconvolution and PET are provided to show the finite-sample performance of $\hat{p}_n$ having a fixed number of basis functions. The EM algorithm is known to be slow in implementation, and the kernel-type estimators (KE's) for deconvolution in Stefanski and Carroll (1990) and Fan (1991) or the orthogonal series estimators (OSE's) for PET in Jones and Silverman (1989) and Johnstone and Silverman (1990, 1991) may not have the positivity for some n, especially where p is close to zero such as at tails. For positivity problems, see Jones and Silverman (1989) and Stefanski and Carroll (1990) and our simulation result in Section 6. However, for asymptotic analysis, we consider the class of densities which are bounded away from zero and infinity, in which case the probability that OSE's or KE's take nonpositive values will tend to zero as $n \to \infty$. One can consider modified versions of OSE's and KE's to guarantee the positivity and the property of integration to 1; Efromovich (1997) employed the nonnegative projection in L_2 for this purpose.

In continuation of the numerical work of this paper, computer simulation is being used to determine the finite-sample performance of inference based on log-density estimation with SVD. Important advantages of computer simulation are that attractive and mathematically unwieldy modifications can be studied and that the effect of the ill-posedness by comparing $\hat{p}_n$ with the estimate based on data from p which is not observable in practice can be seen. In our investigation, we have focused on a selection rule of choosing basis functions in a data-dependent manner.

The basic idea of this paper is similar to Johnstone and Silverman [(1990); [JS] hereafter] and Donoho (1995), although they did not consider the positivity constraint: their proposal is to form SVD/WVD coefficients of the empirical data and to operate on these coefficients. It is believed that WVD is a promising topic for future investigation due to its remarkable local adaptivity.

The paper is organized as follows. In Section 2 we describe SVD for deconvolution and PET. Section 3 proposes the log-density estimation based on SVD. Asymptotic results on rates of convergence are stated in Section 4 and proved in Section 7. A minimax lower bound for deconvolution is provided in Section 5 and the proof of it is also given in Section 7. Section 6 contains numerical examples for deconvolution and PET using simulated data.

2. SVD for deconvolution and PET. We begin with a brief review of SVD and its significance. Let $\mathscr{G}$ and $\mathscr{H}$ be Hilbert spaces and let $K: \mathscr{G} \to \mathscr{H}$ be a bounded linear operator. Let $\langle \ , \ \rangle$ stand equally for the inner products of $\mathscr{G}$ and $\mathscr{H}$. Then under suitable conditions there exist orthonormal sets of functions $\{\phi_\nu\}$ in $\mathscr{G}$ and $\{\psi_\nu\}$ in $\mathscr{H}$, and (possibly complex) numbers $\{\lambda_\nu\}$, the singular values of K, such that the following hold:

1. given p in $\mathscr{G}$, $Kp = \sum_\nu \lambda_\nu \langle p, \phi_\nu \rangle \psi_\nu$;
2. the ϕ_ν's span the orthogonal complement of the kernel of K;
3. the ψ_ν's span the range of K;
4. and $K\phi_\nu = \lambda_\nu \psi_\nu$ for all ν.

If no λ_ν is zero, we have the reproducing formula

$$p = \sum_\nu \frac{\langle Kp, \psi_\nu \rangle}{\lambda_\nu} \phi_\nu.$$

For example, if K^*K is a compact operator, we can choose $\{\phi_\nu\}$ as its eigenfunctions; λ_ν^2, the corresponding eigenvalues; and $\psi_\nu = K\phi_\nu / \langle K\phi_\nu, K\phi_\nu \rangle^{1/2}$, the normalized image.

The point here is that in the bases ϕ_ν and ψ_ν, K is diagonal and if a singular value λ_ν is small, then it will be difficult to recover reliably the component of an unknown function p along the corresponding ϕ_ν based on observations from Kp since noise encountered in estimation of the component of p along ϕ_ν will be amplified by a factor of λ_ν^{-1}. There are several forms dealing with this instability such as the windowed SVD method which includes the tapered orthogonal series method, quadratic regularization and iterative damped backprojection; refer to Donoho (1995) for more details on these methods and various examples of SVD.

2.1. *Circular deconvolution.* Suppose that the observations Y_j, $j = 1, \dots, n$, are the sum of two independent and identically distributed components X_j and Z_j. We desire to estimate the unknown density p of the X_j using the observed data Y_j whose unknown density is q. The density function k of the additive contaminating noise Z_j is assumed known; in addition, X_j and Z_j are assumed to be independent. For simplicity, we assume that the X_j and Z_j take values in the unit circle; see Section 5 for the noncircular case. This classical model of circular data or so-called wrapped distribution has also been considered by Johnstone and Silverman (1991) and Efromovich (1997), among others. We assume that $\mathscr{B}$ and $\mathscr{D}$ are the unit circle, and the dominating measures μ and σ of p and q, respectively, are the usual Lebesgue measure on the unit circle; $\mathscr{G}$ and $\mathscr{H}$ are the spaces of functions which are square-integrable with respect to the Lebesgue measure. The density functions p and q of X_j and Y_j, respectively, are related by the convolution equation

$$q(y) = \int_0^1 k(y-x)p(x)\,dx, \tag{2}$$

where all arithmetic on the arguments of k and p is performed modulo 1. Let us observe that

$$|q(y)| \le \sup_{x\in\mathscr{B}} p(x) \int_0^1 k(y-x)\,dx = \sup_{x\in\mathscr{B}} p(x). \tag{3}$$

Let $\sum_{\mathscr{Z}} \lambda_\nu \phi_\nu$ be the formal Fourier expansion of k with $\phi_\nu(x) = e^{2\pi i \nu x}$. By standard calculations the convolution mapping has SVD given by singular functions $\phi_\nu(x) = \psi_\nu(x) = e^{2\pi i\nu x}$, with singular values λ_ν, $\nu \in \mathscr{Z}$.

2.2. *PET*. We describe SVD for an idealized version of PET described in [JS]. Give the name *detector space* to the space $\mathscr{D}$ of all possible unordered pairs of points on a detector ring, and call *brain space* a disc $\mathscr{B}$ in the plane enclosed by the detector ring. Here $\mathscr{B}$ is $\{(x_1, x_2)\colon x_1^2 + x_2^2 \le 1\}$ in Cartesian coordiantes or $\{(u,v)\colon 0 \le u \le 1,\ 0 \le v < 2\pi\}$ in polar coordinates and $\mathscr{D}$ is $\{(y_1, y_2)\colon 0 \le y_1 \le 1,\ 0 \le y_2 < 2\pi\}$. Define a dominating measure μ on brain space to be $d\mu(x_1, x_2) = \pi^{-1}\,dx_1\,dx_2$ or, equivalently, $d\mu(u,v) = \pi^{-1}u\,du\,dv$, and on detector space, a dominating measure σ by $d\sigma(y_1, y_2) = 2\pi^{-2}(1-y_1^2)^{1/2}\,dy_1\,dy_2$; $\mathscr{G}$ is the space $L_2(\mathscr{B}, \mu)$ of functions on brain space that are square-integrable with respect to the dominating measure μ. Correspondingly, $\mathscr{H}$ is the space $L_2(\mathscr{D}, \sigma)$ of detector-space functions square-integrable relative to σ.

Now suppose an emission takes place at (X_1, X_2) and that the corresponding photon pair has trajectory at angle Ω; taking $0 \le \Omega \le \pi$ for definiteness, the joint probability density with respect to $dx_1\,dx_2\,d\omega$ on $\mathscr{B}$ and $0 \le \omega \le \pi$ is given by $p_{X_1, X_2, \Omega}(x_1, x_2, \omega) = \pi^{-2}p(x_1, x_2)$. Now change variables by setting

$$\begin{aligned} Y_1 &= |X_1 \cos\Omega + X_2 \sin\Omega|,\\ Y_2 &= \begin{cases} \Omega, & \text{if } X_1\cos\Omega + X_2\sin\Omega \ge 0,\\ \Omega + \pi, & \text{otherwise,}\end{cases}\\ T &= -X_1\sin\Omega + X_2\cos\Omega; \end{aligned} \tag{4}$$

the variables (Y_1, Y_2) are the coordinates of the detected photon pair. Integrating out the unobserved variable T, we obtain the joint density with respect to $dy_1\,dy_2$,

$$p_{Y_1, Y_2}(y_1, y_2) = \pi^{-2}\int_{-\sqrt{1-y_1^2}}^{\sqrt{1-y_1^2}} p(y_1\cos y_2 - t\sin y_2,\ y_1\sin y_2 + t\cos y_2)\,dt.$$

The observable density q with respect to σ in detector space is given by Kp with K the Radon operator; specifically,

$$(5)\qquad \begin{aligned}(Kp)(y_1, y_2) &= \tfrac{1}{2}\left(1 - y_1^2\right)^{-1/2} \int_{-\sqrt{1-y_1^2}}^{\sqrt{1-y_1^2}} p(y_1 \cos y_2 \\ &\qquad - t \sin y_2, y_1 \sin y_2 + t \cos y_2)\, dt.\end{aligned}$$

Introducing the Dirac delta function δ, (5) can be written as (1), where

$$(6)\qquad k(y, x) = \tfrac{1}{2}\left(1 - y_1^2\right)^{-1/2} \delta(x_1 \cos y_2 + x_2 \sin y_2 - y_1).$$

It can be seen that

$$(7)\qquad |Kp(y_1, y_2)| \le \sup_{x \in \mathscr{B}} p(x) \tfrac{1}{2}\left(1 - y_1^2\right)^{-1/2} \int_{-\sqrt{1-y_1^2}}^{\sqrt{1-y_1^2}} dt = \sup_{x \in \mathscr{B}} p(x).$$

To describe the SVD of the Radon operator, we need double indices, specifically $\nu \in \mathscr{N} = \{\nu = (\nu_1, \nu_2)\colon\ \nu_2 = 0, 1, 2, \dots;\ \nu_1 = \nu_2, \nu_2 - 2, \dots, -\nu_2\}$. In brain space, an orthonormal basis for $L_2(\mathscr{B}, \mu)$ is given by $\phi_\nu(u, v) = (\nu_2 + 1)^{1/2} Z_{\nu_2}^{|\nu_1|}(u) \exp(i\nu_1 v)$, $\nu \in \mathscr{N}$, $(u, v) \in \mathscr{B}$, where $Z_{\nu_2}^{|\nu_1|}$ denotes the Zernike polynomial of degree ν_2 and order $|\nu_1|$. The corresponding orthonormal functions in $L_2(\mathscr{D}, \sigma)$ are $\psi_\nu(y_1, y_2) = U_{\nu_2}(y_1)\exp(i\nu_1 y_2)$ for $\nu \in \mathscr{N}$ and $(y_1, y_2) \in \mathscr{D}$, where $U_{\nu_2}(\cos y_1) = \sin(\nu_2 + 1)y_1 / \sin y_1$ are the Chebyshev polynomials of the second kind. Then, we have $K\phi_\nu = \psi_\nu$ with singular values λ_ν specified by $\lambda_\nu = (\nu_2 + 1)^{-1/2}$, $\nu \in \mathscr{N}$. Refer to Deans (1983) for the properties of the Zernike and the Chebyshev polynomials.

In the PET problem, we have $X_1, \dots, X_n$, which are n independent unobservable observations of emissions in brain space from the density p, and $Y_1, \dots, Y_n$, which are corresponding observable observations in detector space drawn from the density q.

3. Log-density estimation based on SVD. This section describes the log-density estimation based on SVD of the operator K. We relate our method to the other two popular principles: maximum entropy method in the problem of moments [Mead and Papanicolaou (1984)] and the EM algorithm for deconvolution.

3.1. *Maximum entropy method for the moment problem.* In the classical moment problem, one seeks a positive density p from knowledge of its power moments $a_j = \int x^j p(x)\, dx$, $j = 0, 1, 2, \dots$. In practice, only a finite number of moments, say $J + 1$, are usually available. Clearly then there exists an infinite variety of functions whose first $J + 1$ moments coincide and a unique reconstruction of p is impossible.

The maximum entropy approach offers a definite procedure for the construction of a sequence of approximations. Introducing appropriate Lagrange multipliers $\theta_0, \theta_1, \dots, \theta_J$, one ends up with the solution of the form $p_J(x) = \exp(\sum_{j=1}^{J} \theta_j x^j - \theta_0)$. Assuming $a_0 = 1$, the Lagrange multipliers should satisfy a system of equations:

$$\exp(\theta_0) = \int \exp\left(\sum_{j=1}^{J} \theta_j x^j\right) dx \quad \text{and} \tag{8}$$

$$a_j = \frac{\int x^j \exp\left(\sum_{j=1}^{J} \theta_j x^j\right) dx}{\int \exp\left(\sum_{j=1}^{J} \theta_j x^j\right) dx}, \qquad j = 1, \dots, J.$$

For numerical purposes, one introduces a function $\Gamma(\theta)$,

$$\Gamma(\theta) = \sum_{j=1}^{J} \theta_j a_j - \log \int \exp\left(\sum_{j=1}^{J} \theta_j x^j\right) dx, \tag{9}$$

where the a_j's are the actual numerical values of the known moments. Stationary points of the function $\Gamma(\theta)$ in (9) are solutions to the equation $\partial \Gamma(\theta)/\partial \theta_j = 0$, which is precisely equation (8).

For statisticians, a_j's are given in the form of empirical moments, that is, $\hat{a}_j = n^{-1}\sum_{m=1}^{n} X_m^j$, where $X_1, \dots, X_n$ form a random sample from p. Then the maximum-entropy solution is the density estimator $\hat{p}_J$ matching each empirical moment $\hat{a}_j$ to $\int x^j \hat{p}_J(x)\, dx$, where $\hat{p}_J$ belongs to the exponential family

$$\left\{ p_\theta \colon p_\theta(x) = \frac{\exp\left(\sum_{j=1}^{J} \theta_j x^j\right)}{\int \exp\left(\sum_{j=1}^{J} \theta_j x^j\right) dx} \right\}.$$

We have a question before we give our method of estimation: how to give a density estimator when we are given a set of statistics $\{\hat{b}_j\}$ whose expected values are the same as $\{\hat{a}_j\}$?

3.2. *Missing data formulation for deconvolution.* Deconvolution problem may be formulated in terms of missing data, which enables the application of the EM algorithm. Suppose the density of X has the form $p_\theta = \exp\{\sum_{\nu \in \mathscr{J}} \theta_\nu \phi_\nu - c(\theta)\}$, where $c(\theta) = \log \int \exp\{\sum \theta_\nu \phi_\nu(x)\}\, dx$ and $\{\phi_\nu \colon \nu \in \mathscr{J}\}$ is a set of functions. For identifiability, a constant function is not included in $\{\phi_\nu \colon \nu \in \mathscr{J}\}$. We also assume that p_θ satisfies the linear operator equation (1) with $k(\cdot - x)$ the conditional density of Y given $X = x$. Then the joint distribution of (X, Y) is specified by k and p_θ such that $k(y - x)p_\theta(x)$ is the density of (X, Y). In this context one may speak of X as the missing part of (X, Y).

The EM algorithm [Dempster, Laird and Rubin (1977)] is an iterative procedure for selecting an estimator of an unknown parameter θ when a part of the sample is missing. It is especially appropriate to applications involving exponential families. For a random sample $(X_1, Y_1), \dots, (X_n, Y_n)$ from the distribution of (X, Y), the maximum likelihood estimator maximizes the (unobserved) likelihood function $l_u(\theta) = n^{-1}\sum_{j=1}^{n} \log\{k(Y_j - X_j)p_\theta(X_j)\}$. If

$Y_1, \dots, Y_n$ are merely available, the (observed) log-likelihood is given by

$$l_o(\theta) = n^{-1} \sum_{j=1}^{n} \log \int k(Y_j - x) p_\theta(x)\, dx. \tag{10}$$

The two steps of the EM algorithm can be expressed as follows:

1. E step—calculate

$$\phi_\nu^{(m)} = n^{-1} \sum_{j=1}^{n} \left\{ \frac{\int \phi_\nu(x) k(Y_j - x) p_\theta(x)\, dx}{\int k(Y_j - x) p_\theta(x)\, dx} \right\};$$

2. M step—obtain $\theta^{(m+1)}$ as the solution of $\int \phi_\nu(x) p_\theta(x)\, dx = \phi_\nu^{(m)}$.

A stationary point θ^* of the EM algorithm satisfies

$$\int \phi_\nu(x) p_{\theta^*}(x)\, dx = \frac{1}{n} \sum_{j=1}^{n} \left\{ \frac{\int \phi_\nu(x) k(Y_j - x) p_{\theta^*}(x)\, dx}{\int k(Y_j - x) p_{\theta^*}(x)\, dx} \right\}. \tag{11}$$

If θ^* is the true parameter, then (11) can be written as $E_{\theta^*}\phi_\nu(X) = n^{-1}\sum_{j=1}^{n} E_{\theta^*}\{\phi_\nu(X_j) \mid Y_j\}$. Since $E_{\theta^*}[E_{\theta^*}\{\phi_\nu(X) \mid Y\}] = E_{\theta^*}\phi_\nu(X)$, the EM algorithm may be interpreted as a method matching each $E_\theta \phi_\nu(X)$, $\nu \in \mathscr{J}$, to an unbiased estimator of it.

3.3. *Definition of estimators*. When K is either the convolution operator or the Radon operator, $\phi_0(x) = 1$ for $x \in \mathscr{B}$, $\psi_0(y) = 1$ for $y \in \mathscr{D}$ and $\lambda_0 = 1$. Assume that the densities p and q have singular function series representation $p = \sum a_\nu \phi_\nu$ and $q = \sum b_\nu \phi_\nu$, where $a_\nu = \langle p, \phi_\nu \rangle = \int_{\mathscr{B}} p(x)\phi_\nu(x)\mu(dx)$ and $b_\nu = \langle q, \psi_\nu \rangle = \int_{\mathscr{D}} q(y)\phi_\nu(y)\sigma(dy)$. The relation $q = Kp$ gives

$$b_\nu = \lambda_\nu a_\nu. \tag{12}$$

Relation (12) is essential in constructing a density estimator $\hat{p}_n$ based on SVD guaranteeing that $\hat{p}_n$ is a *bona fide* density in the sense that $\hat{p}_n$ is nonnegative and integrates to 1.

An index set $\mathscr{J}_n$ is a subset of $\mathscr{Z}^d \setminus \{0\}$ and J_n denotes the number of elements in $\mathscr{J}_n$. For a subset $\mathscr{J}$ of $\mathscr{Z}^d$, let $\sum_{\mathscr{J}}$ denote the summation over $\mathscr{J}$. Let Θ_n be the collection of J_n-dimensional vectors $\theta = (\theta_\nu)_{\mathscr{J}_n}$, where $(\theta_\nu)_{\mathscr{J}_n}$ is a J_n-dimensional vector of elements θ_ν, $\nu \in \mathscr{J}_n$. The exponential family based on singular functions $\{\phi_\nu \colon \nu \in \mathscr{J}_n\}$ is defined by

$$p_\theta(x) = \exp\left\{ \sum_{\mathscr{J}_n} \theta_\nu \phi_\nu(x) - c_n(\theta) \right\} \quad \text{for } x \in \mathscr{B} \text{ and } \theta \in \Theta_n, \tag{13}$$

where $c_n(\theta) = \log \int_{\mathscr{B}} \exp\{\sum \theta_\nu \phi_\nu(x)\}\mu(dx)$.

Before we propose our estimators, let us consider the case when we have a random sample $X_1, \dots, X_n$ from the distribution with density p. The log-likelihood function based on the exponential family (13) is defined by $l_u(\theta) = \sum_{\mathscr{J}_n} \theta_\nu \bar{\phi}_\nu - c_n(\theta)$, $\theta \in \Theta_n$, where $\bar{\phi}_\nu = n^{-1}\sum_{j=1}^{n} \phi_\nu(X_j)$. Given X_j's, we define by $\tilde{p}_n = p_{\tilde{\theta}_n}$ the maximum likelihood estimator (MLE) of p, where $\tilde{\theta}_n$ maximizes $l_u(\theta)$. The MLE $\tilde{p}_n$ should satisfy the likelihood equation $\langle \phi_\nu, p_\theta \rangle = \overline{\phi}_\nu$ for $\nu \in \mathscr{J}_n$; that is, $\tilde{p}_n$ is an estimator matching each $\int_{\mathscr{B}} \phi_\nu(x)\tilde{p}_n(x)\mu(dx)$ to an unbiased estimator $\overline{\phi}_\nu$ of $\int_{\mathscr{B}} \phi_\nu(x)p(x)\mu(dx)$.

Now we define density estimators for our inverse problems. Since X_j's are not observable, we replace $\overline{\phi}_\nu$ by $\overline{\psi}_\nu/\lambda_\nu$, where $\overline{\psi}_\nu = n^{-1}\sum_{j=1}^{n} \psi_\nu(Y_j)$. We introduce the indirect likelihood

$$l(\theta) = \sum_{\mathscr{J}_n} \theta_\nu \frac{\overline{\psi}_\nu}{\lambda_\nu} - c_n(\theta), \qquad \theta \in \Theta_n. \tag{14}$$

It should be emphasized the $l(\theta)$ in (14) is not necessarily interpretable as a likelihood; it is an object function to be optimized for the definition of our density estimators. We define by $p_{\hat{\theta}_n}$ the maximum indirect likelihood estimator (MILE) of p based on incomplete data $Y_1, \dots, Y_n$, where $\hat{\theta}_n$ maximizes $l(\theta)$ over $\theta \in \Theta_n$. Let us note that the MILE $\hat{p}_n$ should satisfy the equation

$$\langle \phi_\nu, p_\theta \rangle = \overline{\psi}_\nu/\lambda_\nu \quad \text{for } \nu \in \mathscr{J}_n. \tag{15}$$

Remark 1. From (15), the MILE can be motivated as an estimator matching each of $\int_{\mathscr{B}} \phi_\nu(x)p_\theta(x)\mu(dx)$, $\nu \in \mathscr{J}_n$, to an unbiased estimator $\overline{\psi}_\nu/\lambda_\nu$ of $\int_{\mathscr{B}} \phi_\nu(x)p(x)\mu(dx)$ based on the incomplete data alone. For the direct problem, Stone (1989, 1990, 1994), [BS] and Koo and Kim (1996) have shown asymptotic properties of exponential family density estimators based on several basis functions. Since $E_p \overline{\phi}_\nu = E_q(\overline{\psi}_\nu/\lambda_\nu)$ from (12), we may hope that we can investigate the asymptotic behavior of $\hat{p}_n$ by a modification of these methods.

Remark 2. For deconvolution, it can be shown that the function $l_o(\theta)$ given by (10) increases at each iteration of the EM algorithm, even when the density of X does not belong to the exponential family based on the singular functions ϕ_ν, $\nu \in \mathscr{J}_n$ [Koo and Park (1996)]. Under fairly general conditions the EM algorithm will converge to a local maximum of $l_o(\theta)$ [Wu (1983)]. Since the concavity of $l_u(\theta)$ in θ does not in general imply concavity of $l_o(\theta)$ in θ, there is no guarantee that such a local maximum point is unique or that it is the global maximum point. However, since the Hessian matrix of $c_n(\theta)$ is positive definite, $\hat{\theta}_n$ is unique if it exists.

4. Asymptotic results. In this section we state asymptotic results on sequences of exponential families based on SVD. From now on we let $M, M_1, M_2, \dots$ denote positive constants which are independent of n.

In our subsequent analysis, we place a constraint on the unknown density p over $\mathscr{B}$ by assuming $\log p$ lies in a particular class $\mathscr{F}$. For reasons of mathematical tractability, this class is taken to be a particular ellipsoid $\mathscr{F}$ in the Hilbert space $\mathscr{G} = L_2(\mathscr{B}, \mu)$. Let $|x| = (\sum_{j=1}^{d} x_j^2)^{1/2}$ for $x \in \mathscr{R}^d$.

Smoothness class for deconvolution. We let $\mathscr{F}(r, M)$ be the nonparametric class of functions f in $\mathscr{G}$ such that $f = \sum_{\mathscr{Z}} f_\nu \phi_\nu$ satisfies the smoothness condition

$$\sum_{\mathscr{Z}} (1+|\nu|)^{2r}|f_\nu|^2 \le M \tag{16}$$

for a positive constant M.

Smoothness classes for PET. We transform the index set $\mathscr{N}$ into the lattice orthant $\mathscr{N}' = \{(\nu_1', \nu_2'): \nu_1' \ge 0, \nu_2' \ge 0\}$ by the change of variables $\nu_1' = (\nu_1+\nu_2)/2$, $\nu_2' = (\nu_1-\nu_2)/2$. A function $f \in \mathscr{G}$ can be represented as $f = \sum_{\mathscr{N}'} f_\nu \phi_\nu$, where

$$\phi_\nu(u,v) = (\nu_1+\nu_2+1)^{1/2} Z_{\nu_1+\nu_2}^{|\nu_1-\nu_2|}(u)\exp(i(\nu_1-\nu_2)v) \quad \text{for } \nu \in \mathscr{N}'.$$

We consider the ellipsoids

$$\mathscr{F}(r,M) = \left\{ f: \sum_{\mathscr{N}'} (1+|\nu|)^{2r}|f_\nu|^2 \le M \right\} \tag{17}$$

and

$$\mathscr{F}_{JS}(r,M) = \left\{ f: \sum_{\mathscr{N}'} (1+\nu_1)^r(1+\nu_2)^r|f_\nu|^2 \le M \right\} \tag{18}$$

for a threshold M.

REMARK 3. The positive integer r in (16)–(18) can be thought of as a measure of smoothness of functions in such ellipsoids. Let $L_2(I)$ be the Hilbert space of square-integrable functions on $I=[0,1]$, and let $\|\cdot\|_2$ denote the usual L_2-norm therein. For integer m and $f \in L_2(I)$, let $D^m f$ denote the derivative of order m, and let $\mathscr{W}_2^r = \{f \in L_2(I): D^r f \in L_2\}$ be the corresponding Sobolev space on I. For a function $f \in L_2(I)$, define by f_ν the classical Fourier coefficient which is given by the usual inner product of f and ϕ_ν, where $\phi_\nu(x) = e^{2\pi i\nu x}$ for $x \in I$. The nonparametric class of functions given by $\{f: \sum_{\mathscr{Z}} (1+|\nu|)^{2r}|f_\nu|^2 \le M\}$ can be identified by a periodic Sobolev class in the L_2-sense, that is, $\mathscr{W}_2(r,M) = \{f \in \mathscr{W}_2^r: \|D^r f\|_2 \le M,\ D^m f(0) = D^m f(1),\ m = 0,\dots,r\}$ [Nussbaum (1985)]. On the other hand, f belongs to the set $\mathscr{F}_{JS}(r,M)$ if f has r weak derivatives with respect to the modified dominating measure $d\mu_{r+1}(x) = (r+1)(1-|x|^2)^r\, d\mu(x)$; refer to Proposition 2.2 of [JS] for the proof of this fact. The characterization of $\mathscr{F}(r,M)$ given by (17) appears to be quite different from that of $\mathscr{F}_{JS}(r,M)$.

Ill-posedness. It is assumed that the singular values satisfy

$$|\lambda_\nu| \ge d_1(1+|\nu|)^{-s} \quad \text{for } \nu \in \mathscr{Z}^d, \tag{19}$$

where d_1 is a positive constant and s a nonnegative constant. We refer to s as the *order* of K.

REMARK 4. Condition (19) excludes the case where $\lambda_\nu = 0$ for some $\nu \in \mathscr{Z}^d$, in which case the density function p is not identifiable and, hence, not estimable. The constant s in (19) can be thought of as a measure of ill-posedness of the inverse problem; the larger s is, the more difficult the given inverse problem. If the Fourier coefficients of the density k satisfy $|\lambda_\nu| \asymp (1 + |\nu|)^{-s}$ in our deconvolution problem, then the order of convolution operator is s. Polyá's criterion [Feller (1971), page 509] shows that this is the Fourier expansion of a probability density k. In the idealized PET problem, the order of the Radon operator is given by $s = 1/2$ since it can be shown that there exists a positive constant d_1 such that $\lambda_\nu = (1 + \nu_2)^{-1/2} \geq d_1(1 + |\nu|)^{-1/2}$ for $\nu \in \mathscr{N}$ and $\lambda_\nu = (1 + \nu_1 + \nu_2)^{-1/2} \geq d_1(1 + |\nu|)^{-1/2}$ for $\nu \in \mathscr{N}'$. The relatively slow decay of the singular values suggests that the costs of indirect observation in the PET problem are not inordinately large.

Index set. Let N_n denote a positive integer depending on sample size n. The index set for deconvolution is chosen by

$$\mathscr{J}_n = \{\nu \in \mathscr{Z}: 0 < |\nu| \leq N_n\}.$$

For PET, the index set is chosen as

$$\mathscr{J}_n = \begin{cases} \{\nu \in \mathscr{N}': 0 < |\nu| \leq N_n\}, & \text{when } \log p \in \mathscr{F}(r, M), \\ \{\nu \in \mathscr{N}': 1 < (\nu_1 + 1)(\nu_2 + 1) \leq N_n\}, & \text{when } \log p \in \mathscr{F}_{JS}(r, M). \end{cases}$$

The relative entropy (Kullback–Leibler divergence) between two densities p_1 and p_2 defined on $\mathscr{B}$ is denoted by

$$D(p_1 \| p_2) = \int p_1(x) \log\left(\frac{p_1(x)}{p_2(x)}\right) \mu(dx).$$

In addition to the L_2 loss function, we mainly use the entropy-based loss function since the use of exponential family density estimation is natural with this loss function (see [BS] and references therein). Let $\|\cdot\|_\infty$ and $\|\cdot\|_2$ denote L_∞- and L_2-norms, respectively, with respect to μ. Let $\mathscr{J}_n^0 = \{0\} \cup \mathscr{J}_n$ and define $\mathscr{S}_n$ to be the linear space spanned by singular functions ϕ_ν for $\nu \in \mathscr{J}_n^0$; that is, $\mathscr{S}_n = \{\sum_{\mathscr{J}_n^0} \theta_\nu \phi_\nu\}$. Define A_n such that $\|s_n\|_\infty \leq A_n \|s_n\|_2$ for all $s_n \in \mathscr{S}_n$; let $\Delta_n = \inf_{s_n \in \mathscr{S}_n} \|f - s_n\|_2$ and $\gamma_n = \inf_{s_n \in \mathscr{S}_n} \|f - s_n\|_\infty$ be L_2 and L_∞ degrees of approximation of $f = \log p$ by a truncated singular-function series $s_n \in \mathscr{S}_n$.

Information projection. Consider the equation

$$\langle \phi_n, p \rangle = \langle \phi_n, p_\theta \rangle, \qquad \theta \in \Theta_n, \tag{20}$$

where ϕ_n is the J_n-dimensional vector of elements ϕ_ν, $\nu \in \mathscr{J}_n$, and $\langle \phi_n, h \rangle$ denotes the vector $(\langle \phi_\nu, h \rangle)_{\mathscr{J}_n}$ for any function h. By the Pythagorean-like identity (4.2) of [BS], the solution θ_n^* to (20) uniquely minimizes $D(p \| p_\theta)$ over $\theta \in \Theta_n$. Let $p_n^* = p_{\theta_n^*}$ if θ_n^* exists. We refer to p_n^* as the information projection of p.

For the asymptotic results when $\log p \in \mathscr{F}(r, M)$, we need the following condition:

(A1) $r \geq 1$ for deconvolution and $r \geq 2$ for PET.

The following theorem shows that θ_n^* exists with $\langle \phi_n, p \rangle = \langle \phi_n, p_{\theta_n^*} \rangle$ and that there is an upper bound on the approximation error $D(p \| p_n^*)$. For this task we set $\varepsilon_n = 4M_1^2 \exp(4\gamma_n + 1) A_n \Delta_n$, where M_1 is the positive constant to be given in Lemma 1 satisfying $M_1^{-1} \le p \le M_1$.

THEOREM 1. *Suppose that* (A1) *holds and* $\varepsilon_n \le 1$. *Then, for* $\log p \in \mathscr{F}(r, M)$, *the information projection* p_n^*, *achieving the minimum* $D(p \| p_n^*)$, *exists and satisfies the following*:

(i) $\|\log p/p_n^*\|_\infty \le 2\gamma_n + \varepsilon_n$;
(ii) $D(p \| p_n^*) \le (M_1/2)\exp(\gamma_n)\Delta_n^2$.

In the following theorem, we show that the MILE $\hat{p}_n$ exists except on a set whose probability is less than a preassigned value and that the estimation error $D(p_n^* \| \hat{p}_n)$ converges to zero in probability at the rate $N_n^{2s} J_n/n$. For this theorem, we set $\delta_n = 4d_1^{-1} M_1^{3/2} \exp(2\gamma_n + \varepsilon_n + 1) N_n^s A_n \sqrt{J_n/n}$.

THEOREM 2. *Suppose that* $\log p \in \mathscr{F}(r, M)$, $\varepsilon_n \le 1$ *and* $\delta_n \le 1$. *Then, under* (A1), *for every* $M_2 \le \delta_n^{-2}$, *there is a set of probability less than* $1/M_2$ *such that outside this set the MILE exists and satisfies the following*:

(i) $\|\log p_n^*/\hat{p}_n\|_\infty \le M_2^{1/2}\delta_n$;
(ii) $D(p_n^* \| \hat{p}_n) \le M_2 M_3 N_n^{2s}(J_n/n)$, where $M_3 \ge 2d_1^{-2} M_1 \exp(2\gamma_n + \varepsilon_n + \tau)$ and $\tau = \delta_n M_2^{1/2} \le 1$.

By combining theorems 1 and 2, we obtain an asymptotic result for the MILE. Let $a_n \asymp b_n$ mean that $\inf a_n/b_n > 0$ and $\sup a_n/b_n < \infty$.

THEOREM 3. *Suppose that* $\log p \in \mathscr{F}(r, M)$ *and that* (A1) *holds. Then, choosing* $N_n \asymp n^{1/(2r+2s+d)}$, *we have the following*:

(i) $D(p \| \hat{p}_n) = O_P(n^{-2r/(2r+2s+d)})$;
(ii) $\|\log p/\hat{p}_n\|_\infty = o_p(1)$;
(iii) $\|p - \hat{p}_n\|_2^2 = O_P(n^{-2r/(2r+2s+d)})$.

For the asymptotic result for PET when $\log p \in \mathscr{F}_{JS}(r, M)$, we assume the following condition:

(A2) $r \ge 3$.

THEOREM 4. *Suppose that* $\log p \in \mathscr{F}_{JS}(r, M)$. *Under* (A2), *we have that, for* $N_n \asymp n^{1/(r+2)}$, *the following hold*:

(i) $D(p \| \hat{p}_n) = O_P(n^{-r/(r+2)})$;
(ii) $\|p - \hat{p}_n\|_2^2 = O_P(n^{-r/(r+2)})$.

REMARK 5. For PET one may want to note the difference of the rates of convergence in Theorems 3 and 4. The rate $n^{-2r/(2r+3)}$ is same as the rate given by Donoho (1995), although a different nonparametric class of functions is considered; the rate $n^{-r/(r+2)}$ is same as the rate in [JS] although we assume that $\log p \in \mathscr{F}_{JS}(r, M)$ rather than $p \in \mathscr{F}_{JS}(r, M)$.

REMARK 6. In the next section we show the minimaxity of MILE for deconvolution, where the rates of convergence depend on the smoothness of the contaminating noise. We anticipate that the rates in Theorem 3 are also minimax lower bounds for other inverse problems, including PET.

REMARK 7. Stone (1990) considered large-sample inference for logspline models when $\log p$ belongs to the Hölder class. Koo and Kim (1996) addressed the minimaxity of log-density estimation based on wavelets over the Besov space which includes the Sobolev space and the Hölder class as a special case. Barron and Yang (1996) obtained minimaxity for the direct problem when p belongs to nonparametric classes including the multivariate Hölder class. It would be worthwhile to extend our results via the WVD of Donoho (1995) to linear inverse problems over other classes of functions such as Besov spaces.

5. More results on deconvolution. This section addresses the minimaxity of the MILE for circular deconvolution. To find a minimax lower bound, we follow the popular approach: (a) specify a subproblem and (b) use the difficulty of the subproblem as a lower bound. Especially, we will use the method of Koo (1993) where basis functions are used for both lower and upper bounds. This idea was inspired by Stone (1980, 1982), Ibragimov and Has'minskii (1981), Birgé (1983), Donoho and Liu (1991a, b) and [JS].

5.1. *Minimaxity for circular deconvolution.* The difficulty of deconvolution depends on the smoothness of the distribution of the error variable Z and on the smoothness of p. We classify the smoothness of error distributions into two classes following Fan (1991). The characteristic function of Z is denoted by $\chi_Z(t) = E\exp(itZ)$. We will call the distribution of a random variable Z ordinary smooth of order s if its characteristic function $\chi_Z(t)$ satisfies

$$d_1|t|^{-s} \le |\chi_Z(t)| \le d_2|t|^{-s} \quad \text{as } |t| \to \infty, \tag{21}$$

for a positive constant s. We will call the distribution of a random variable Z super smooth of order s if its characteristic function $\chi_Z(t)$ satisfies

$$d_1|t|^{s_0}\exp\left(-|t|^s/d_0\right) \le |\chi_Z(t)| \le d_2|t|^{s_1}\exp\left(-|t|^s/d_0\right) \quad \text{as } |t| \to \infty, \tag{22}$$

where s and d_0 are positive constants. Here the positive constants d_0, d_1 and d_2 and real s_0 and s_1 will have no effect on explored convergence.

Consider an unknown distribution P_p which depends on p with $\log p \in \mathscr{F}(r, M)$. Let $\hat{p}_n$, $n \ge 1$, denote estimators of p, $\hat{p}_n$ being based on $Y_1, \dots, Y_n$ from the distribution P_q or, equivalently, P_p. Let $\{b_n\}$ be a sequence of positive constants. It is called a lower rate of convergence for p in a relative entropy sense if

$$\lim_{c\to 0} \liminf_{n} \inf_{\hat{p}_n} \sup_{\log p \in \mathscr{F}(r,M)} P_p\big(D(p\|\hat{p}_n) \ge cb_n\big) = 1,$$

here $\inf_{\hat{p}_n}$ denotes the infimum over all possible estimators $\hat{p}_n$. The sequence is said to be an achievable rate of convergence for p in a relative entropy sense if there is a sequence $\{\hat{p}_n\}$ of estimators such that

$$\lim_{c\to \infty} \limsup_{n} \sup_{\log p \in \mathscr{F}(r,M)} P_p\big(D(p\|\hat{p}_n) \ge cb_n\big) = 0. \tag{23}$$

It is called an optimal rate of convergence for p if it is a lower and an achievable rate of convergence. If $\{b_n\}$ is the optimal rate of convergence and $\{\hat{p}_n\}$ satisfies (23), the estimators $\hat{p}_n$, $n \ge 1$, is said to be asymptotically optimal.

To develop upper bounds, we assume that the following holds:

(A3) $\chi_Z(t) \neq 0$ for any t.

For the circular deconvolution model, we have the following asymptotic optimality of our MILE's. According to Theorem 3, $\{n^{-2r/(2r+2s+1)}\}$ is an achievable rate of convergence for the ordinary smooth case.

THEOREM 5. *Suppose that* (A1) *and* (A3) *hold.*

(a) *If Z is ordinary smooth in the sense of* (21), *then* $\{n^{-2r/(2r+2s+1)}\}$ *is a lower rate of convergence and the MILE achieves this rate of convergence by choosing* $N_n \asymp n^{1/(2r+2s+1)}$.

(b) *If Z is super smooth in the sense of* (22), *then* $\{(\log n)^{-2r/s}\}$ *is a lower rate of convergence and the MILE achieves this rate of convergence by choosing* $N_n \asymp (\log n)^{1/s}$.

REMARK 8. It follows from the argument used in the proof of lower rates of convergence that the rates in Theorem 5 are also lower rates of convergence when the Kullback–Leibler divergence is replaced by the L_2-norm. As in Theorems 1–3, one can show the MILE for the super smooth case achieves the same rate of convergence in L_2-norm.

5.2. *Noncircular deconvolution.* In circular deconvolution, Z takes values only in the unit circle, whereas in noncircular case X still takes values in the unit interval and Z may take values in $\mathscr{R}$. Then Y may take values in $\mathscr{R}$ and the density q of Y is given by

$$q(y) = \int_0^1 k(y-x)p(x)\,dx \quad \text{for } y \in \mathscr{R}, \tag{24}$$

where k is the density of Z. The ill-posedness for the noncircular deconvolution problem is also determined by the smoothness of the distribution of the error variable Z. By the smoothness of the error distribution, again we mean the decay rate of $|\chi_Z(t)|$ as $|t| \to \infty$. Refer to Fan (1991) for specific examples of these distributions.

Since $\phi_\nu(x) = e^{2\pi i \nu x}$ is not square-integrable on $\mathscr{R}$, it cannot be a singular function for the noncircular convolution operator. However, independence of X and Z and the property of ϕ_ν provide us unbiased estimators of $E\phi_\nu(X)$'s which are sufficient to give an MILE $\hat{p}_n$ of p. It follows from the relation $E\phi_\nu(Y) = \chi_Z(2\pi\nu)E\phi_\nu(X)$ that an unbiased estimator of $E\phi_\nu(X)$ is given by $\overline{\phi}_\nu/\chi_Z(2\pi\nu)$. Similarly, the MILE $\hat{p}_n$ of p is $p_{\hat{\theta}_n}$, where $\hat{\theta}_n$ is the maximizer of the indirect likelihood (14), which is given by $\sum_{0<|\nu|\le N_n} \theta_\nu \overline{\phi}_\nu/\chi_Z(2\pi\nu) - c_n(\theta)$.

For the noncircular case, we have same rates of convergence as in the circular case, which is stated in the following theorem.

THEOREM 6. *Suppose that* (A1) *and* (A3) *hold.*

(a) *If* $|\chi_Z(t)| \ge d_1|t|^{-s}$ *as* $|t| \to \infty$ *for a positive constant* d_1 *and a nonnegative constant* s, *we have* $D(p\|\hat{p}_n) = O_P(n^{-2r/(2r+2s+1)})$ *for* $N_n \asymp n^{1/(2r+2s+1)}$.

(b) *If* $|\chi_Z(t)| \ge d_1|t|^{-s_0}\exp(-|t|^s/d_0)$ *as* $|t| \to \infty$ *for some positive constants* d_0, d_1, s *and real* s_0, *then we have that* $D(p\|\hat{p}_n) = O_P((\log n)^{-2r/s})$ *by choosing* $N_n \asymp (\log n)^{1/s}$.

REMARK 9. Theorem 6 can be proved by the argument used to prove Theorems 3 and 5; it follows from (24) that (3) is still true for the noncircular case, and Lemma 2 in Section 7 follows from the inequality $E_q|\psi_\nu(Y)|^2 \le M_1\int_{\mathscr{D}} \sigma(dy) = M_1$.

6. Simulation. The finite-sample performance of MILE's having a fixed number of basis functions is illustrated using simulated data for deconvolution and PET. The problem of choosing basis functions for MILE's in a data-dependent way would be an important problem for future investigation.

6.1. *Deconvolution.* The exponential family

$$p_\theta(x) = \exp\left\{\sum_{\nu=1}^{2} \theta_\nu \cos(2\pi\nu x) + \sum_{\nu=1}^{2} \theta_{\nu+2} \sin(2\pi\nu x) - c(\theta)\right\}, \quad 0 \le x \le 1, \tag{25}$$

is taken for deconvolution, where $c(\theta)$ is the normalizing constant. Since our ultimate interest is in the densities rather than the parameters, we do not use the conventional basis functions $\{\sqrt{2}\cos(2\pi\nu x), \sqrt{2}\sin(2\pi\nu x)\}$ for convenience in implementation. The Newton–Raphson method as in Koo and Park (1996) is employed to maximize the indirect likelihood; the Gaussian quadrature `gauleg.f` in Press, Teukolsky, Vetterling and Flannery (1992) is used for the computations of various quantities during the Newton–Raphson iterations such as $c(\theta)$ or the Hessian matrix of $c(\theta)$. To provide an approximation to the unknown density function on the real line, we scale the data to the interval $[0.1, 0.9]$ and find the preliminary MILE on the interval $[0, 1]$. The final answer is then scaled back to the original interval.

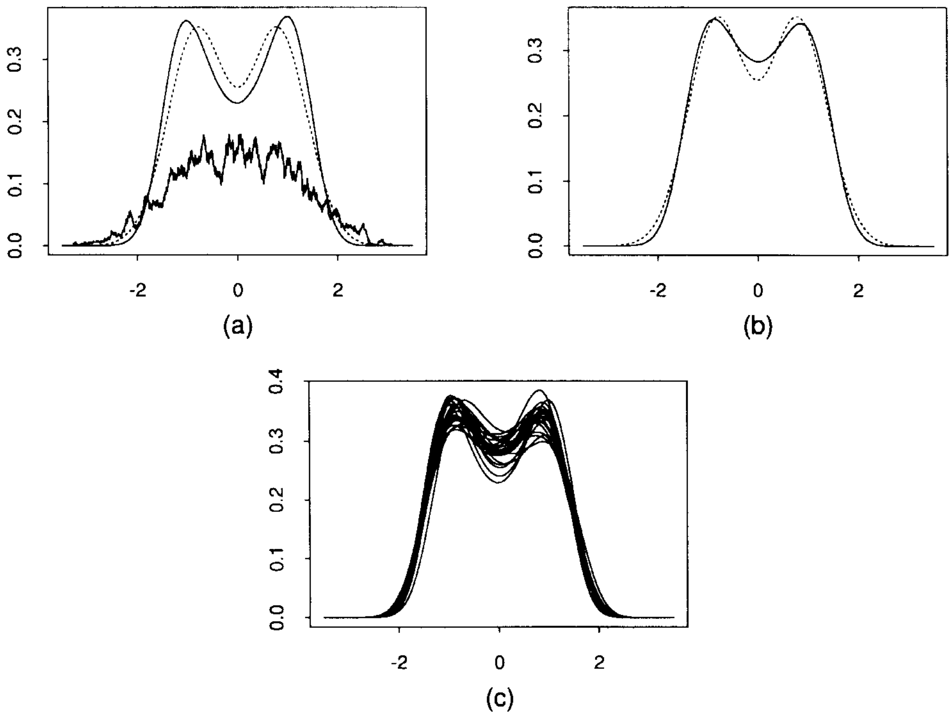

FIG. 1. *The Stefanski–Carroll bimodal density with* $n = 2500$: (a) *an MILE*; (b) *the mean of* 25 *repetitions*; (c) *an overlap plot of* 25 *repetitions* (*solid line*, *MILE*; *dotted line*, *truth*).

To compare the performance of MILE's with other estimators, we have generated X_j's from a bimodal density of the form $p(x) = 0.5 \times N(x; -(2/3)^{1/2}, 1/3) + 0.5N(x; (2/3)^{1/2}, 1/3)$. Here $N(\cdot; a, b^2)$ is the density function of a normal distribution with mean a and variance b^2. Normal measurement error with variance $1/3$ has been considered so that q is unimodal. The sample size $n = 2500$ and 25 repetitions have been performed. Figure 1(a) shows an estimate, Figure 1(b) displays the mean of the 25 estimates and Figure 1(c) gives a good idea of the variability inherent to the estimators. The wiggly line in Figure 1(a) is the kernel density estimate of q which is included only as a descriptor of $Y_1, \dots, Y_n$; it is rescaled in order not to interfere other plots. We can note that the performance of MILE looks much better than that of Stefanski and Carroll (1990) and similar to that of Koo and Park (1996).

6.2. *PET*. Since we work with real densities, we may identify the complex bases with equivalent real orthonormal bases in a standard fashion as in [JS]. The exponential family in brain space is chosen by $p_\theta(u, v) = \exp\{\Sigma_{\mathscr{J}} \theta_\nu \tilde{\phi}_\nu(u, v) - c(\theta)\}$ for $(u, v) \in \mathscr{B}$, where $\mathscr{J} = \{\nu\colon\ \nu_2 = 1, \dots, B;\ \nu_1 =$

$\nu_2, \nu_2 - 2, \dots, -\nu_2\}$, $c(\theta) = \log \int_{\mathscr{B}} \exp(\sum_\nu \theta_\nu \tilde{\phi}_\nu)\, d\mu$ and

$$\tilde{\phi}_\nu = \begin{cases} \sqrt{2}\ \mathrm{Re}\, \phi_\nu, & \text{if } \nu_1 > 0, \\ \phi_{(0,\nu_2)}, & \text{if } \nu_1 = 0, \\ \sqrt{2}\ \mathrm{Im}\, \phi_\nu, & \text{if } \nu_1 < 0. \end{cases}$$

In brain space, an algorithm of computing the Zernike polynomials is necessary. The Zernike polynomials are related to the more general Jacobi polynomials [Deans (1983)] such that the recurrence relation (4.5.14) of Jacobi polynomials in Press, Teukolsky, Vetterling and Flannery (1992) is used for the computation of the Zernike polynomials. In detector space, we need to compute the Chebyshev polynomial of the second, for which the recurrence relation in Appendix C of Deans (1983) is adopted. To maximize the indirect likelihood, the Newton–Raphson method is implemented as in Koo (1996), where an iterated Gaussian quadrature rule based on `gauleg.f` is used for the computation of various quantities which are necessary during the Newton–Raphson iterations.

Figure 2 illustrates a simulation example for the idealized PET. The density function p of (X_1, X_2) is the truncation at $\mathscr{B}$ of

$$\tfrac{1}{3}N_2\Big(x; (0, -0.3), \big(\tfrac{1}{8}\big)^2\Big) + \tfrac{1}{3}N_2\Big(x; (0.3, 0.3), \big(\tfrac{1}{8}\big)^2\Big) \\ + \tfrac{1}{3}N_2\Big(x; (-0.3, 0.3), \big(\tfrac{1}{8}\big)^2\Big),$$

where $N_2(x; a, b^2) = N(x_1; a_1, b^2)N(x_2; a_2, b^2)$ for $a = (a_1, a_2)$. Figure 2(a) shows the true density, Figure 2(b) displays an MILE based on $X_1, \dots, X_n$ (which are observable in simulation but unobservable in practice) and Figure 2(c) illustrates an MILE using data $Y_1, \dots, Y_n$ from $q = Kp$, where Y_j's are generated according to the formula described in Section 2.2. For this example, the sample size $n = 6400$ and $B = 4$. As a comparison, Figure 2(d) shows an OSE which has the form $(1 + \sum_{\mathscr{J}} \hat{a}_\nu \tilde{\phi}_\nu)/\pi$, where $\hat{a}_\nu = n^{-1} \sum_j \tilde{\psi}_\nu(Y_j)/\lambda_\nu$ with $\tilde{\psi}_\nu$ the real version of ψ_ν, and $B = 7$. Since $B = 4$ means 14 parameters and $B = 7$ implies 35 parameters, the MILE gives a much more parsimonious reconstruction of p than the OSE. The OSE with $B = 4$ for the same data is unimodal and we need about 65 parameters ($B = 10$) to identify the trimodal structure reasonably well.

7. Proof of asymptotic results. In this section, we prove asymptotic results in Sections 4 and 5 supposing that $\log p \in \mathscr{F}(r, M)$ or $\log p \in \mathscr{F}_{JS}(r, M)$ and that (A1)–(A3) hold. Since we use several lemmas in [BS], we write Lemma BSi to denote Lemma i in [BS]. The method of proof is an extension of [BS] to the multivariate case with multiindex. Since $\{\phi_\nu\}$ and $\{\psi_\nu\}$ are fixed for a given operator K, we should prove our results under the assumption that $\{\phi_\nu\}$ and $\{\psi_\nu\}$ are orthonormal with respect to the dominating measures μ and σ, respectively. Observe that Lemmas BS1–BS5 are still true for multivariate density estimation using multiindex such as in our case.

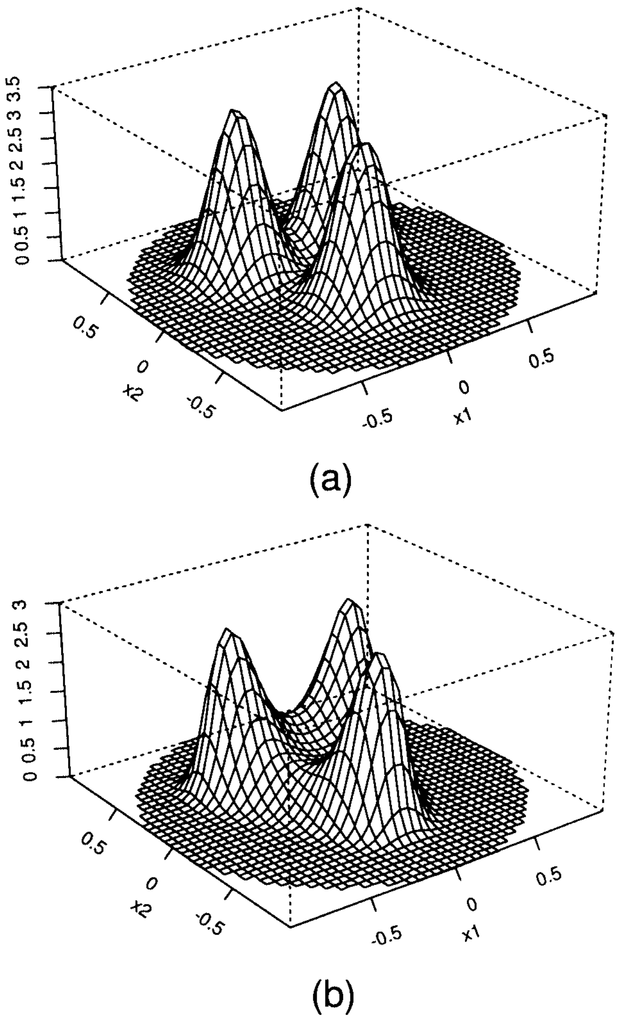

FIG. 2. *The trimodal density for PET*: (a) *the plot of* p; (b) *an MLE based on* X_j*'s*; (c) *an MILE with* 14 *parameters based on* Y_j*'s*; (d) *an OSE with* 35 *parameters based on* Y_j*'s.*

All integrals are understood to be with respect to the dominating measure μ unless stated otherwise; $\|\alpha\|$ is the Euclidean norm of a vector $\alpha \in \Theta_n$. For $f = \log p$, let $f_\nu = \langle f, \phi_\nu \rangle$ and let $s_n(f) = \sum_{\mathscr{J}_n^0} f_\nu \phi_\nu$ denote the truncated singular-function series which is assumed to satisfy the given L_2- and L_∞-bounds on the error $f - s_n(f)$. Let C denote a positive constant which is independent of n and is not necessarily equal at each appearance of it.

7.1. *A technical lemma.* We develop upper and lower bounds for p.

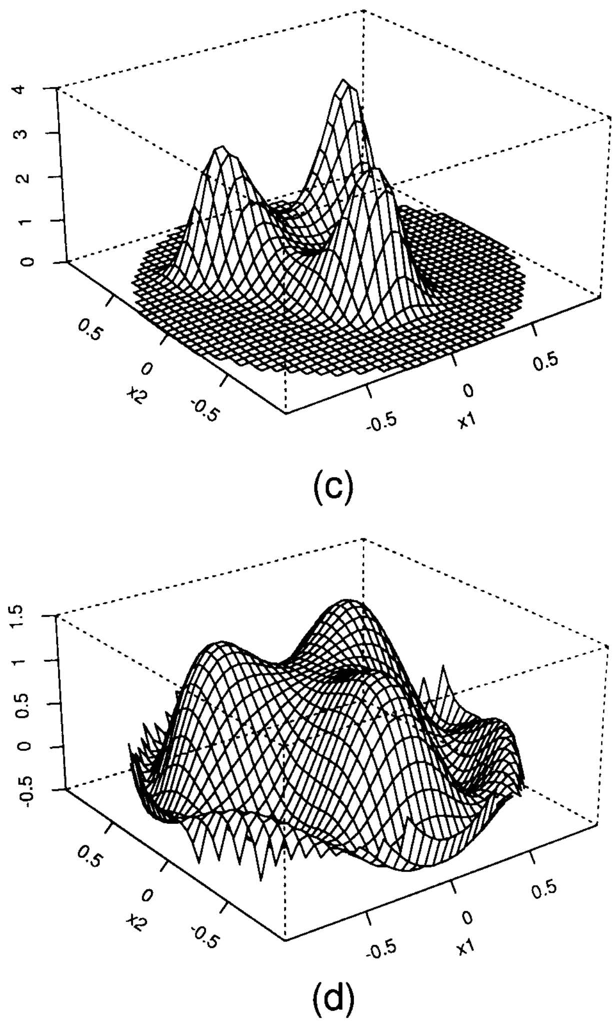

FIG. 2. *Continued.*

LEMMA 1. *Under* (A1), $M_1^{-1} \le p \le M_1$.

PROOF. Consider the case of deconvolution. By the Cauchy–Schwarz inequality, we obtain that

$$|f(x)|^2 \le \sum_{\mathscr{Z}} |f_\nu|^2 (1 + |\nu|)^{2r} \sum_{\mathscr{Z}} |\phi_\nu(x)|^2 (1 + |\nu|)^{-2r} \le M \sum_{\mathscr{Z}} (1 + |\nu|)^{-2r}.$$

Since $r \ge 1$, the series $\sum_{\mathscr{Z}} (1 + |\nu|)^{-2r}$ is convergent. This completes the proof for deconvolution by choosing $M_1 > 1$ such that $(\log M_1)^2 \ge M \sum_{\mathscr{Z}} (1 + |\nu|)^{-2r}$.

Now consider the case with PET. Since Zernike polynomials satisfy $|Z_{\nu_2}^{\nu_1}(u)| \leq Z_{\nu_2}^{\nu_1}(1) = 1$ for $0 \leq u \leq 1$ (see [JS]),

$$|\phi_\nu| \leq \sqrt{1 + \nu_1 + \nu_2} \quad \text{for } \nu \in \mathscr{N}'. \tag{26}$$

Applying the Cauchy–Schwarz inequality, we have that

$$|f(x)|^2 \leq M \sum_{\mathscr{N}'} (1 + |\nu|)^{-2r+1}.$$

Let us observe that

$$\begin{aligned}
\sum_{\mathscr{N}'} (1 + |\nu|)^{-2r+1} &\leq C \int_{\mathscr{R}^2} (1 + |x|)^{-2r+1}\, dx \\
&= C \int_0^\infty dt \int_{|x|=t} (1 + |x|)^{-2r+1}\, dx \\
&= C \int_0^\infty (1 + t)^{-2r+1}\, dt \int_{|x|=t} dx \\
&= C \int_0^\infty (1 + t)^{-2r+1} t\, dt,
\end{aligned}$$

which is convergence under (A1). This completes the proof of Lemma 1 by choosing $M_1 > 1$ such that $(\log M_1)^2 \geq MC\int_0^\infty (1 + t)^{-2r+1} t\, dt$. □

7.2. *Proof of Theorem* 1. The first task is to show that θ_n^* exists with $\langle \phi_n, p_n^* \rangle = \langle \phi_n, p \rangle$ and that $\log p/p_n^*$ is bounded by a constant when n is large. For this task, set $\alpha_n^* = \langle \phi_n, p \rangle$ and $\alpha_n = \langle \phi_n, p_{\beta_n} \rangle$, where $\beta_n = (f_\nu)_{\mathscr{J}_n} \in \Theta_n$. The entries in the vector $\alpha_n^* - \alpha_n$ are seen to be coefficients in the $L_2(\mu)$ orthogonal projection of $p - p_{\beta_n}$ onto $\mathscr{S}_n$. By Bessel's inequality, Lemma 1 and Lemma BS2, we have

$$\begin{aligned}
\| \alpha_n^* - \alpha_n \|^2 \leq \| p - p_{\beta_n} \|_2^2 &\leq M_1 \int \frac{(p - p_{\beta_n})^2}{p} \\
&\leq M_1^2 \exp\big(2\|f - s_n(f)\|_\infty - 2\{f_0 + c_n(\beta_n)\}\big) \|f - s_n(f)\|_2^2 \\
&\leq M_1^2 \exp(4\gamma_n) \Delta_n^2.
\end{aligned}$$

For the last inequality we have used the fact that $|c_n(\beta_n) + f_0|$ is not greater than $\|f - s_n(f)\|_\infty$, since $c_n(\beta_n) + f_0$ is seen to equal $\log \int \exp(s_n(f) - f) p$. From this same fact it is seen that $\|\log p/p_{\beta_n}\|_\infty \leq 2\|f - s_n(f)\|_\infty = 2\gamma_n$. By this and Lemma 1, $\|\log p_{\beta_n}\|_\infty \leq \log M_1 + 2\gamma_n$. Now apply Lemma BS5 with $\theta_0 = \beta_n$, $\alpha_0 = \alpha_n$, $\alpha = \alpha_n^*$, $q = 1$ and $b = \exp(\|\log p_{\beta_n}\|_\infty) \leq M_1 \exp(2\gamma_n)$. If $M_1 \exp(2\gamma_n)\Delta_n \leq 1/(4ebA_n)$, that is, if $\varepsilon_n \leq 1$, then from Lemma BS5 we may conclude that the solution θ_n^* to the equation $\int \phi_n p_\theta = \alpha_n$ exist and that $\|\log p_n^*/p_{\beta_n}\|_\infty \leq \varepsilon_n$. So by the triangle inequality, we obtain $\|\log p/p_n^*\|_\infty \leq 2\gamma_n + \varepsilon_n$, which verifies Theorem 1(i), and

$$\|\log p_n^*\|_\infty \leq 2 \log M_1 + \gamma_n + \varepsilon_n. \tag{27}$$

By Lemma 1 and Lemma BS1, we have

$$D(p\|p_n^*) \le D(p\|p_{\beta_n}) \le \frac{1}{2}\exp(\|f - s_n(f)\|_\infty) M_1 \|f - s_n(f)\|_2^2$$
$$\le \frac{M_1}{2}\exp(\gamma_n)\Delta_n^2.$$

This completes the proof of Theorem 1. □

7.3. *Proof of Theorem* 2. To prove Theorem 2, we need the following lemma.

LEMMA 2. $E_q \sum_{\mathscr{J}_n} \{\bar{\psi}_\nu - E_q \psi_\nu(Y)\}^2 \le M_1 J_n/n$.

PROOF. By Lemma 1, (3) and (7) we have that $q \le M_1$. Hence

$$\sum_{\mathscr{J}_n} E_q\Big\{\bar{\psi}_\nu - E_q\psi_\nu(Y)\Big\}^2 = \frac{1}{n}\sum_{\mathscr{J}_n} E_q\Big\{\psi_\nu(Y) - E_q\psi_\nu(Y)\Big\}^2$$
$$\le \frac{1}{n}\sum_{\mathscr{J}_n} E_q \psi_\nu^2(Y)$$
$$\le \frac{1}{n}\sum_{\mathscr{J}_n} M_1 \int_{\mathscr{D}} \psi_\nu^2(y)\sigma(dy) = M_1\frac{J_n}{n}.$$

This completes the proof of Lemma 2. □

For the proof of Theorem 2, we have to show that $D(p_n^*\|\hat{p}_n)$ is small with high probability. Let $\alpha_n^* = \int \phi_n p_n^*$, which is the same as $\int \phi_n p = (E_q\psi_\nu(Y)/\lambda_\nu)_{\mathscr{J}_n}$. Also let $\hat{\alpha}_n = (\bar{\psi}_\nu/\lambda_\nu)_{\mathscr{J}_n}$. Whenever a solution $\hat{\theta}_n \in \Theta_n$ to the equation $\int \phi_n p_\theta = \hat{\alpha}_n$ exists, we recognize $\hat{p}_n = p_{\hat{\theta}_n}$ as an MILE. With these choices $\|\hat{\alpha}_n - \alpha_n^*\|^2 = \sum_{\mathscr{J}_n}\{\bar{\psi}_\nu - E_q\psi_\nu(Y)\}^2/|\lambda_\nu|^2$. By Chebyshev's inequality, $\|\hat{\alpha}_n - \alpha_n^*\|^2 \le d_1^{-2} M_1 M_2 N_n^{2s} J_n/n$ except on a set whose probability satisfies

$$P\left[\sum_{\mathscr{J}_n}\frac{\Big\{\bar{\psi}_\nu - E_q\psi_\nu(Y)\Big\}^2}{|\lambda_\nu|^2} > d_1^{-2}\frac{M_1M_2N_n^{2s}J_n}{n}\right]$$
$$\le P\left[\sum_{\mathscr{J}_n}\Big\{\bar{\psi}_\nu - E_q\psi_\nu(Y)\Big\}^2 > \frac{M_1M_2J_n}{n}\right]$$
$$\le \frac{n}{M_1M_2J_n}E_q\left[\sum_{\mathscr{J}_n}\Big\{\bar{\psi}_\nu - E_q\psi_\nu(Y)\Big\}^2\right] \le \frac{1}{M_2}.$$

Here the first inequality is due to the assumption on $\{\lambda_\nu\}$ and the third is due to Lemma 2. Now apply Lemma BS5 with $\theta_0 = \theta_n^*$, $\alpha_0 = \alpha_n^*$, $\alpha = \hat{\alpha}_n$, $q = 1$ and $b = \exp(\|\log p_n^*\|_\infty)$, where b is not greater than $M_1\exp(2\gamma_n + \varepsilon_n)$ by

(27). If $d_1^{-1}(M_1 M_2 N_n^{2s} J_n/n)^{1/2} \le 1/(4ebA_n)$, that is, if $\delta_n^2 \le 1/M_2$, then except on the set above (whose probability is less than $1/M_2$) the conditions of Lemma BS5 are satisfied, whence $\hat{\theta}_n$ exists and (i) $\|\log p_n^*/\hat{p}_n\|_\infty \le 4be^\tau A_n \|\hat{\alpha}_n - \alpha_n^*\| \le M_2^{1/2}\delta_n$ and (ii) $D(p_n^* \| \hat{p}_n) \le 2b\exp(\tau)\|\hat{\alpha}_n - \alpha_n^*\|^2 \le 2d_1^{-2} M_1 M_2 \exp(2\gamma_n + \varepsilon_n + \tau) N_n^{2s} J_n/n$. Here τ satisfies $4ebA_n\|\hat{\alpha}_n - \alpha_n^*\| \le \tau \le 1$. The L_∞-norm of $\log p_n^*/\hat{p}_n$ has just been shown to be less than or equal to $M_2^{1/2}\delta_n$ and the estimation error satisfies $D(p_n^* \| \hat{p}_n) \le M_2 M_3 N_n^{2s} J_n/n$, except on a set whose probability is less than $1/M_2$. Thus the proof of Theorem 2 is complete. □

7.4. *Proof of Theorem* 3. To prove Theorem 3, we need bounds on A_n, Δ_n and γ_n.

LEMMA 3. (i) *For deconvolution*, $A_n = \sqrt{2N_n + 1}$, $\Delta_n = O(N_n^{-r})$ *and* $\gamma_n = O(N_n^{-(r-1/2)})$.

(ii) *For PET*, $A_n = CN_n^{3/2}$, $\Delta_n = O(N_n^{-r})$ *and* $\gamma_n = O(N_n^{-(r-3/2)})$.

PROOF. Refer to [BS] for the proof of (i). Consider the case of PET. To determine A_n, choose any element $s_n = \sum_{\mathscr{I}_n^0} \theta_\nu \phi_\nu$ in $\mathscr{S}_n$. By the Cauchy–Schwarz inequality and (26), we have that, uniformly in $x \in \mathscr{B}$,

$$
\begin{aligned}
|s_n(x)| &\le \left(\sum_{\mathscr{I}_n^0} |\phi_\nu(x)|^2 \right)^{1/2} \left(\sum_{\mathscr{I}_n^0} |\theta_\nu|^2 \right)^{1/2} \\
&\le \left(\sum_{\mathscr{I}_n^0} (1 + \nu_1 + \nu_2) \right)^{1/2} \|s_n\|_2 \\
&\le CN_n^{3/2} \|s_n\|_2 .
\end{aligned}
$$

Let $\mathscr{I}_n^c = \{\nu \in \mathscr{N}' : |\nu| > N_n\}$. Since $(1 + N_n)^{2r} \sum_{\mathscr{I}_n^c} |f_\nu|^2 \le \sum_{\mathscr{I}_n^c} (1 + |\nu|)^{2r} |f_\nu|^2 < M$, we have the bound on Δ_n. It follows from the Cauchy–Schwarz inequality that the error $|f(x) - s_n(f)(x)|^2$ is bounded by

$$
\begin{aligned}
\sum_{\mathscr{I}_n^c} (1 + |\nu|)^{2r} |f_\nu|^2 \sum_{\mathscr{I}_n^c} (1 + |\nu|)^{-2r+1} &\le M \sum_{\mathscr{I}_n^c} (1 + |\nu|)^{-2r+1} \\
&\asymp \int_{N_n}^\infty dt \int_{|x|=t} (1 + |x|)^{-2r+1}\, dx \\
&\asymp N_n^{-2r+3}.
\end{aligned}
$$

This completes the proof of Lemma 3. □

PROOF OF THEOREM 3. Choose $N_n \asymp n^{1/(2r+2s+d)}$. By Lemma 3 and (A1), $\gamma_n = o(1)$, $\varepsilon_n = O(A_n \Delta_n) = o(1)$ and $\delta_n = O(N_n^s A_n \sqrt{J_n/n}) = o(1)$. Therefore p_n^* exists and $\hat{p}_n$ exists in probability for sufficiently large n. It follows from Lemma 3 that $\Delta_n^2 \asymp n^{-2r/(2r+2s+d)}$ and $N_n^{2s} J_n/n \asymp n^{-2r(2r+2s+d)}$. Consequently, from Theorems 1 and 2, we obtain the desired result of Theorem 3 as follows. Since the Kullback–Leibler loss decomposes into a sum of approximation error and estimation error by Lemma BS3: $D(p\|\hat{p}_n) = D(p\|p_n^*) + D(p_n^*\|\hat{p}_n)$, we can verify Theorem 3(i). By the triangle inequality, we have that $\|\log p/\hat{p}_n\|_\infty = O_P(2\gamma_n + \varepsilon_n + \delta_n) = o_p(1)$, which is the desired result of Theorem 3(ii). It follows from Lemmas BS1 and BS2 that $\|p - \hat{p}\|_2^2 = O_P(D(p\|\hat{p}_n))$, which implies Theorem 3(iii). Now the proof of Theorem 3 is complete. □

7.5. *Proof of Theorem* 4. By the Cauchy–Schwarz inequality and (26), we have that

$$\begin{aligned} |f(x)|^2 &\le \sum_{\mathscr{N}'} |f_\nu|^2 (1+\nu_1)^r (1+\nu_2)^r \sum_{\mathscr{N}'} (1+\nu_1+\nu_2)(1+\nu_1)^{-r}(1+\nu_2)^{-r} \\ &\le M \sum_{\mathscr{N}'} (1+\nu_1)^{-r+1/2}(1+\nu_2)^{-r+1/2}, \end{aligned}$$

which is convergent under (A2). Hence, we have the following lemma as in Lemmas 1 and 2.

LEMMA 4. *There exists* M_1 *such that* $M_1^{-1} \le p \le M_1$ *and* $E_q\{\bar{\psi}_\nu(Y) - E_q\psi_\nu(Y)\}^2 \le M_1/n$.

LEMMA 5. *We have* (i) $A_n = CN_n$, (ii) $\Delta_n = O(N_n^{-r/2})$ *and* (iii) $\gamma_n = O(N_n^{-(r-2)/2})$.

PROOF. For A_n, choose any element $s_n = \sum_{\mathscr{I}_n^0} \theta_\nu \phi_\nu$ in $\mathscr{S}_n$. It follows from Lemma (4.3) of [JS] that

$$\sum_{\mathscr{I}_n^0} (1+\nu_1+\nu_2) \le M_4 N_n^2. \tag{28}$$

By the Cauchy–Schwarz inequality, (26) and (28), we have that, uniformly in $x \in \mathscr{B}$,

$$\begin{aligned} |s_n(x)| &\le \Big(\sum_{\mathscr{I}_n^0} |\phi_\nu(x)|^2\Big)^{1/2} \Big(\sum_{\mathscr{I}_n^0} |\theta_\nu|^2\Big)^{1/2} \\ &\le \Big(\sum_{\mathscr{I}_n^0} (1+\nu_1+\nu_2)\Big)^{1/2} \|s_n\|_2 \\ &\le M_4^{1/2} N_n \|s_n\|_2, \end{aligned}$$

which shows (i). Let $\mathscr{I}_n^c = \{\nu \in \mathscr{N}' : (\nu_1+1)(\nu_2+1) > N_n\}$. Since

$$\begin{aligned} N_n^r \sum_{\mathscr{I}_n^c} |f_\nu|^2 &\le \sum_{m > N_n} \sum_{(\nu_1+1)(\nu_2+1)=m} (\nu_1+1)^r(\nu_2+1)^r |f_\nu|^2 \\ &\le \sum_{\mathscr{I}_n^c} (\nu_1+1)^r(\nu_2+1)^r |f_\nu|^2 \le M, \end{aligned}$$

we have (ii). It follows from (26) that the error $|f(x) - s_n(f)(x)|^2$ is bounded by

$$M \sum_{\mathscr{I}_n^c} (1+\nu_1)^{-r+1}(1+\nu_2)^{-r+1} = \sum_{m > N_n} m^{-r+1} = O\big(N_n^{-r+2}\big),$$

which proves (iii). This completes the proof of Lemma 5. □

PROOF OF THEOREM 4. Define α_n^* and $\hat{\alpha}_n$ as in the proof of Theorem 2. By Lemma 4, (7) and (28), we have

$$\begin{aligned} P&\left[\sum_{\mathscr{I}_n} \frac{\big\{\bar{\psi}_\nu - E_q \psi_\nu(Y)\big\}^2}{|\lambda_\nu|^2} > \frac{M_1 M_2 M_4 N_n^2}{n}\right] \\ &\le \frac{n}{M_1 M_2 M_4 N_n^2} E_q\left[\sum_{\mathscr{I}_n} \big\{\bar{\psi}_\nu - E_q \psi_\nu(Y)\big\}^2 (1+\nu_1+\nu_2)\right] \\ &\le \frac{1}{M_2}. \end{aligned}$$

Now choose $N_n \asymp n^{1/(r+2)}$ such that $\Delta_n^2 \asymp n^{-r/(r+2)}$ and $N_n^2/n \asymp n^{-r/(r+2)}$. By Lemma 5, $\gamma_n = o(1)$, $\varepsilon_n = O(A_n \Delta_n) = o(1)$ and $\delta_n = O(A_n\sqrt{N_n^2/n}) = o(1)$. It follows from the argument used to prove Theorem 3 that $D(p\|\hat{p}_n) = O_P(n^{-r/(r+2)})$ and that $\|p - \hat{p}_n\|_2^2 = O_P(n^{-r/(r+2)})$. This completes the proof of Theorem 4. □

7.6. *Proof of Theorem* 5.

Ordinary smooth case. Since we have shown that the MILE for deconvolution achieves the rate $\{n^{-2r/(2r+2s+1)}\}$ in Theorem 3, it remains to show that it is a lower rate of convergence. For a positive integer N_n, let $V_n = \{v\colon v = 1,\dots,N_n\}$. Define g_{nv} for $v \in V_n$ by

$$g_{nv} = N_n^{-r-1/2}\big(\phi_{N_n+v} + \phi_{-N_n-v}\big).$$

Given a $\{0,1\}$-valued sequence $\tau_n = (\tau_{nv})_{V_n}$, set

$$p_{\tau_n} = 1 + M_5 \sum_{V_n} \tau_{nv} g_{nv}$$

for a constant M_5 which will be determined below. Now choose N_n such that $N_n \asymp n^{1/(2r+2s+1)}$. Let $\mathscr{F}_n$ denote the collection of all functions p_{τ_n} as τ_n varies over the 2^{N_n} possible sequences. The following lemma shows that $\mathscr{F}_n \subset \mathscr{F}(r, M)$ for sufficiently large n.

LEMMA 6. *There is a positive constant M_5 such that, for large n, $\mathscr{F}_n$ is a subset of $\mathscr{F}(r, M)$.*

PROOF. Let $g_{\tau_n} = \sum_{V_n} \tau_{nv} g_{nv}$. Let us note that, for $j = 0,\dots,r-1$,

$$\|D^j g_{\tau_n}\|_\infty \le \sum_{V_n} \|D^j g_{nv}\|_\infty \le C N^{-r+j+1/2} \tag{29}$$

and

$$\|D^r p_{\tau_n}\|_2 \le M_5\sqrt{2}\,(4\pi)^r. \tag{30}$$

It follows from (A1) and (29) that, for large n,

$$C^{-1} \le p_{\tau_n} \le C. \tag{31}$$

By formula (5.35) in Barndorff-Nielsen and Cox (1989) and (29)–(31), we can choose M_5 such that

$$\|D^r \log p_{\tau_n}\|_2 \le M.$$

This completes the proof of Lemma 6. □

By (29) and Lemma BS2,

$$D(p_1\|p_2) \ge C\int (p_1 - p_2)^2 \ge CN_n^{-2r-1} \quad \text{for } p_1 \ne p_2 \in \mathscr{F}_n.$$

It follows from Lemma 3.1 of Koo (1993) that there exists a subset $\mathscr{F}_n^*$ of $\mathscr{F}_n$ such that, for large n,

$$\begin{aligned} D(p_1\|p_2) &> CN_n^{-2r} \quad \text{for } p_1 \ne p_2 \in \mathscr{F}_n^* \quad \text{and} \\ \log\{\#(\mathscr{F}_n^*) - 1\} &> 0.27N_n, \end{aligned} \tag{32}$$

where $\#(\mathscr{F}_n^*)$ denotes the cardinality of $\mathscr{F}_n^*$. Observe that, when n is large,

$$\begin{aligned} C^{-1} \le Kp \le C \quad &\text{for } p \in \mathscr{F}_n^* \quad \text{and} \\ \|Kp_1 - Kp_2\|_2 \le CN_n^{-r-s} \quad &\text{for } p_1, p_2 \in \mathscr{F}_n^*. \end{aligned} \tag{33}$$

By Jensen's inequality

$$D(p_1\|p_2) \le \log\int \frac{p_1^2}{p_2} \le \int \frac{(p_1 - p_2)^2}{p_2} \quad \text{for any densities } p_1, p_2. \tag{34}$$

It follows from (33) and (34) that

$$D(Kp_1\|Kp_2) \le \int \frac{(Kp_1 - Kp_2)^2}{Kp_2} \le CN_n^{-2r-2s} \quad \text{for all } p_1 \text{ and } p_2 \in \mathscr{F}_n^*.$$

Now using Fano's lemma [Birgé (1983)] as in Koo (1993), we have the desired result for the ordinary smooth case.

Super smooth case. Construct $\mathscr{F}_n^*$ as in the ordinary smooth case. In the same manner, we can show that

$$D(Kp_1\|Kp_2) \le CN_n^{-2r+2s_1} \exp\{-2(2\pi N_n)^s/d_0\} \quad \text{for all } p_1, p_2 \in \mathscr{F}_n^*. \tag{35}$$

Choose N_n such that $2\pi N_n = (d_0/2)^{1/s}(\log n + C\log\log n)^{1/s} + a_n$ for $C > (2s_1 - 2r - 1)/s$ and $0 \le a_n < 1$. Now, applying Fano's lemma with (32) and (35) as in Koo (1993), we obtain $(\log n)^{-2r/s}$ is a lower rate of convergence.

To prove that the MILE for deconvolution achieves this lower rate of convergence, let us note that $\|\hat{\alpha}_n - \alpha_n^*\|^2 = O_P[n^{-1}N_n^{1-s_0}\exp\{2(4\pi N_n)^s/d_0\}]$. Here $\hat{\alpha}_n$ and α_n^* denote the same quantities as in the proof of Theorems 1 and 2. Now choose N_n such that $N_n = (4\pi)^{-1}(d_0/4)^{1/s}(\log n)^{1/s} + a_n$ for $0 \leq a_n < 1$, then $N_n^{-2s_0}\exp\{2(4\pi N_n)^s/d_0\}J_n/n = o(n^{-1/3})$. Let us note that $\gamma_n = o(1)$, $\varepsilon_n = o(1)$, $\delta_n = CN_n^{-s_0}\exp\{(4\pi N_n)^s/d_0\}A_n\sqrt{J_n/n} = o(1)$ under (A1). By the same argument used to prove Theorem 3, we have $D(p\|\hat{p}_n) = O_P((\log n)^{-2r/s})$. This completes the proof of Theorem 5. $\square$

Acknowledgments. Both authors are very grateful to an Associate Editor and two reviewers for their helpful remarks, which improved the presentation and the results of the paper substantially. They would like to take this opportunity to thank Professor D. L. Donoho, who introduced the first author to inverse problems (the positivity constraint was one of the main issues of a lecture by Donoho on inverse problems at the University of California, Berkeley), and Professor I. M. Johnstone for helpful discussions on simulation for PET. The original version of this paper was finished in the hospitable environment of the Department of Statistics at Stanford University.

REFERENCES

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1989). *Asymptotic Techniques for Use in Statistics*. Chapman and Hall, London.

BARRON, A. R. and SHEU, C.-H. (1991). Approximation of density functions by sequences of exponential families. *Ann. Statist.* **19** 1347–1369.

BARRON, A. R. and YANG, Y. (1996). Information theoretic determination of minimax rates of convergence. Technical report, Dept. Statistics, Yale Univ.

BICKEL, P. J. and RITOV, Y. (1995). Estimating linear functionals of a PET image. *IEEE Transactions on Medical Imaging* **14** 81–87.

BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237.

CARROLL, R. J. and HALL, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* **83** 1184–1186.

CRAIN, B. R. (1974). Estimation of distributions using orthogonal expansion. *Ann. Statist.* **2** 454–463.

CRAIN, B. R. (1976a). Exponential models, maximum likelihood estimation, and the Haar condition. *J. Amer. Statist. Assoc.* **71** 737–740.

CRAIN, B. R. (1976b). More on estimation of distribution using orthogonal expansions. *J. Amer. Statist. Assoc.* **71** 741–745.

CRAIN, B. R. (1977). An information theoretic approach to approximating a probability distribution. *SIAM J. Appl. Math.* **32** 339–346.

DEANS, S. R. (1983). *The Radon Transform and Some of Its Applications*. Wiley, New York.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric Density Estimation: The L_1 View*. Wiley, New York.

DONOHO, D. L. (1993). Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data. *Proc. Sympos. Appl. Math.* **47** 173–205.

DONOHO, D. L. (1994). Statistical estimation and optimal recovery. *Ann. Statist.* **22** 238–270.

DONOHO, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition. *Appl. Comput. Harm. Anal.* **2** 101–126.

DONOHO, D. L. and LIU, R. C. (1991a). Geometrizing rates of convergence II. *Ann. Statist.* **19** 633–667.

DONOHO, D. L. and LIU, R. C. (1991b). Geometrizing rates of convergence III. *Ann. Statist.* **19** 668–701.

EFROMOVICH, S. (1997). Density estimation for the case of supersmooth measurement error. *J. Amer. Statist. Assoc.* **92** 526–535.

EGGERMONT, P. P. P. and LARICCIA, V. N. (1995). Maximum smoothed likelihood density estimation for inverse problems. *Ann. Statist.* **23** 199–220.

FAN, J. (1991). On the optimal rate of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19** 1257–1272.

FAN, J. (1993). Adaptively local one-dimensional subproblem with applications to a deconvolution problem. *Ann. Statist.* **21** 600–610.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Application* **2**, 2nd ed. Wiley, New York.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer, New York.

JOHNSTONE, I. M. and SILVERMAN, B. W. (1990). Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.* **18** 251–280.

JOHNSTONE, I. M. and SILVERMAN, B. W. (1991). Discretization effects in statistical inverse problems. *J. Complexity* **7** 1–34.

JONES, M. C. and SILVERMAN, B. W. (1989). An orthogonal series density estimation approach to reconstructing positron emission tomography images. *J. Appl. Statist.* **16** 177–191.

KOLACZYK, E. D. (1996). A wavelet shrinkage approach to tomographic image reconstruction. *J. Amer. Statist. Assoc.* **91** 1079–1090.

KOO, J.-Y. (1993). Optimal rates of convergence for nonparametric statistical inverse problems. *Ann. Statist.* **21** 590–599.

KOO, J.-Y. (1996). Bivariate *B*-splines for tensor logspline density estimation. *Comput. Statist. Data Anal.* **21** 31–42.

KOO, J.-Y. and KIM, W.-C. (1996). Wavelet density estimation by approximation of log-densities. *Statist. Probab. Lett.* **26** 271–278.

KOO, J.-Y. and PARK, B. U. (1996). *B*-spline deconvolution based on the EM algorithm. *J. Statist. Comput. Simulation* **54** 275–288.

KOOPERBERG, C. (1995). Density estimation for bivariate survival data. Technical Report 296, Dept. Statistics, Univ. Washington.

KOOPERBERG, C. and STONE, C. J. (1991). A study of logspline density estimation. *Comput. Statist. Data Anal.* **12** 327–347.

KOOPERBERG, C. and STONE, C. J. (1992). Logspline density estimation for censored data. *J. Comput. Graph. Statist.* **1** 301–328.

LEONARD, T. (1978). Density estimation stochastic processes and prior information (with discussion). *J. Roy. Statist. Soc. Ser. B* **40** 113–146.

MEAN, J. R. and PAPANICOLAOU, N. (1984). Maximum entropy in the problem of moments. *J. Math. Phys.* **25** 2404–2417.

MENDELSOHN, J. and RICE, J. (1982). Deconvolution of microfluorometric histograms with *B*-splines. *J. Amer. Statist. Assoc.* **77** 748–753.

NEYMAN, J. (1937). "Smooth" test for goodness of fit. *Scand. Actuar. J.* **20** 149–199.

NUSSBAUM, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L_2. *Ann. Statist.* **13** 984–997.

NYCHKA, D. W. and COX, D. D. (1989). Convergence rates for regularized solutions of integral equations from discrete noisy data. *Ann. Statist.* **17** 556–572.

O'SULLIVAN, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statist. Sci.* **1** 502–527.

O'SULLIVAN, F. (1988). Fast computation of fully automated log-density and log-hazzard estimators. *SIAM J. Sci. Statist. Comput.* **9** 363–379.

O'SULLIVAN, F. (1995). A study of least squares and maximum likelihood of image reconstruction in positron emission tomography. *Ann. Statist.* **23** 1267–1300.

PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T. and FLANNERY, B. P. (1992). *Numerical Recipes in FORTRAN: The Art of Scientific Computing*, 2nd ed. Cambridge Univ. Press.

SILVERMAN, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10** 795–810.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

SILVERMAN, B. W., JONES, M. C., NYCHKA, D. W. and WILSON, J. D. (1990). A smoothed EM algorithm to indirect estimation problems, with particular reference to stereology and emission tomography. *J. Roy. Statist. Soc. Ser. B* **52** 271–324.

STEFANSKI, L. and CARROLL, R. J. (1990). Deconvoluting kernel density estimators. *Statistics* **21** 169–184.

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

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

STONE, C. J. (1989). Uniform error bounds involving logspline models. In *Probability, Statistics and Mathematics: Papers in Honor of Samual Karlin* (T. W. Anderson, K. B. Athreya and D. L. Iglehart, eds.) 335–355. Academic Press, New York.

STONE, C. J. (1990). Large sample inference for logspline model. *Ann. Statist.* **18** 717–741.

STONE, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22** 118–184.

STONE, C. J. and KOO, C. Y. (1986). Logspline density estimation. *Contemp. Math.* **59** 1–15.

VARDI, Y. and LEE, D. (1993). From image deblurring to optimal investments: maximum likelihood solutions for positive linear inverse problems. *J. Roy Statist. Soc. Ser. B* **55** 569–612.

VARDI, Y., SHEPP, L. A. and KAUFMAN, L. (1985). A statistical model for positron emission tomography. *J. Amer. Statist. Assoc.* **80** 8–37.

WU, C. F. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11** 95–103.

DEPARTMENT OF STATISTICS
HALLYM UNIVERSITY
CHUNCHON, KANGWON-DO 200-702
KOREA
E-MAIL: jykoo@sun.hallym.ac.kr