Statistica Sinica Preprint No: SS-2018-0173	
Title	Fast Nonparametric Maximum Likelihood Density Deconvolution Using Bernstein Polynomial
Manuscript ID	SS-2018-0173
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0173
Complete List of Authors	Zhong Guan
Corresponding Author	Zhong Guan
E-mail	zguan@iusb.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0173

Statistica Sinica **24**(201?), 000-000

Fast Nonparametric Maximum Likelihood Density Deconvolution Using Bernstein Polynomials

Zhong Guan
Indiana University South Bend

Abstract: We proposed a new maximum approximate likelihood method for deconvoluting a continuous density on a finite interval in additive measurement error models with a known error distribution. The proposed method uses an approximate Bernstein polynomial model, that is, a finite mixture of specific beta distributions. The change-point detection method is used to choose an optimal model degree. Based on a contaminated sample of size n, under an assumption satisfied by, among others, the generalized normal error distribution, the optimal rate of convergence of the mean integrated squared error is proved to be $\mathcal{O}(n^{-1+5/k}\log n)$ if the underlying unknown density admits an approximate Bernstein polynomial model of degree m within χ^2-divergence of order $\mathcal{O}(m^{-k})$, with $k > 5$. Simulations show that the small-sample performance of the proposed estimator is better than that of the deconvolution kernel density estimator. The proposed method is illustrated using a real-data application.

Key words and phrases: Bernstein polynomial model, Beta mixture model; Deconvolution; Density estimation; Kernel density; Measurement error model; Model selection.

1 Introduction

Because nonparametric models specify almost nothing about a population distribution, they are not wrong. Box (1976) describes how almost all almost all non-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0173

parametric and semiparametric models are not working models, and thus are not useful. Therefore, all useful models are parametric, and they are useful because they provide ways to retrieve information about the population. Nonparametric density estimation is a difficult task in statistics. It is even more difficult when the data are contaminated, which is common for big data. For each fixed x on the support of a density f in a nonparametric model, the information for the one-dimensional parameter $f(x)$ is zero (see Bickel et al. (1998)). Ibragimov and Khasminskii (1983) also showed that there is no such nonparametric model, even with smoothness assumptions, for which this information is positive. Properly reducing the dimensionality from infinity to being finite is necessary. To estimate an unknown smooth function as the nonparametric component of a nonparametric and semiparametric model, as statisticians do in empirical likelihood, we usually approximate it using a step function. Then, we parameterize it using the jump sizes of the step function, which is actually a parametric multinomial model. This approach gives an efficient estimate of the underlying cumulative distribution function. Because this approximate model is discrete, we have to use a kernel or some other method to smooth the “discrete density” to obtain a “continuous” density estimate. Instead of approximating the underlying distribution as a discrete distribution and then smoothing it, Guan (2016) proposed using a Bernstein polynomial model, namely, a mixture of beta distributions, to directly and smoothly estimate the underlying distribution using a maximum approximate Bernstein likelihood method. This method parameterizes the underlying distribution using the coefficients of the Bernstein polynomial, and differs from other Bernstein polynomial smoothing methods (Vitale, 1975) that use an empirical distribution to estimate these coefficients.

The extremely slow optimal rate of convergence of the nonparametric deconvolution kernel density estimate really limits its scope of application. The kernel density is a technique used to smooth the empirical distribution, which is a discretization and a parametrization of the continuous underlying distribution. Any infinite-dimensional nonparametric model is not a working model. In order to solve a nonparametric statistical problem, we have to use an approximate finite-dimensional model. A typical example is the empirical likelihood method, which approximates the unknown underlying distribution function using a step function with jumps only at the observed data. However, better smooth approximations exist and can be used as approximate finite-dimensional models for estimating the density directly.

Based on an independent and identically distributed (i.i.d.) sample $x_1, \ldots, x_n$, without measurement errors, from a population with density f, the kernel density $\tilde{f}_{\mathrm{K}}(x) = (nh)^{-1}\sum_{j=1}^{n} K_h(x - x_j)$ is the convolution of the scaled kernel $K_h(\cdot) = K(\cdot/h)/h$ and the empirical density. It has expectation $\mathrm{E}\{\tilde{f}_{\mathrm{K}}(x)\} = (K_h * f)(x) = \int K_h(x-y)f(y)dy$. Thus $\tilde{f}_{\mathrm{K}}$ is an "unbiased" estimator of the *convolution* $(K_h * f)$, rather than f. No matter how the kernel K and the bandwidth h are chosen, there is always trade-off between the bias and the variance. In this context, Guan (2016) proposed a new nonparametric maximum likelihood estimate for a density that is assumed to be a smooth function with a positive lower bound on a

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0173

known compact support. This method, unlike the traditional Bernstein polynomial smoothing method, approximately parameterizes the underlying density f, after truncation and transformation to $[0, 1]$, using a Bernstein-type polynomial. This polynomial is actually a mixture $f_m(x, \boldsymbol{p}) = \sum_{i=0}^{m} p_i\beta_{mi}(x)$ of beta densities β_{mi} with shape parameters $(i + 1, m - i + 1)$, for $i \in \mathbb{I}_0^m$, and unknown mixing probabilities $\boldsymbol{p} = \boldsymbol{p}_m = (p_0, \ldots, p_m)$; here, and in what follows, $\mathbb{I}_m^n = \{m, \ldots, n\}$, for integers $m \le n$. However, the Bernstein polynomial smoothing method (e.g.,see Vitale, 1975; Tenbusch, 1994) uses an empirical distribution to estimate $f(i/m)$, for $i \in \mathbb{I}_0^m$, which, divided by $m + 1$, are the coefficients of the classical Bernstein polynomials with degree of approximation at best $\mathcal{O}(m^{-1})$, no matter how smooth f is. Lorentz (1963) has shown that better coefficients p_i exist such that $f_m(x, \boldsymbol{p})$ achieves a much better degree of approximation. Like the empirical likelihood and even the bootstrap methods, Guan's (2016) method is a special case of the sieve method (Grenander, 1981; Wong and Shen, 1995; Shen, 1997), in the sense that we are estimating a finite-dimensional parameter in a dense subspace of the infinite-dimensional parameter space. This new estimator has been shown to have an almost parametric rate of convergence in the mean integrated squared error(MISE). Therefore, an accelerated density deconvolution using the Bernstein polynomial model is expected.

In a noisy data situation, let X and ε be independent random variables with densities f and g, respectively. We are interested in estimating the density f based on the contaminated data $y_1, \ldots, y_n$, which are i.i.d. observations of $Y = X + \varepsilon$.

This is an additive measurement error model with measurement error ε. This is common in practice. A simple example is that of rounding off errors with a known uninform distribution on $[-0.5/10^d, 0.5/10^d]$, for some integer d. Usually in this case, the errors are ignored if d is large. However, in some situations, ignoring the measurement errors can result in serious bias in a statistical inference.

The present study focuses on the additive measurement error model in which the error density g is assumed to be *known* or *can be estimated*. The density ψ of Y is the convolution of f and g; that is, $\psi(y) = (f * g)(y) = \int g(y-x)f(x)dx$. Thus $y_1, \ldots, y_n$ is a sample from ψ, which is a mixture of the translated $g(y-x)$, with unknown mixing density $f(x)$. Based on the contaminated data, a nonparametric estimator $\hat{f}_{\mathrm{F}}$, also known as the deconvolution kernel density estimator, of f (e.g., see Carroll and Hall, 1988; Devroye, 1989; Stefanski and Carroll, 1990) is obtained using the inverse Fourier transform with the aid of a kernel density estimation. Briefly, let $\hat{\psi}_{\mathrm{K}}$ be a kernel density estimate of ψ based on y_i. Let $\mathcal{F}(\varphi)$ denote the Fourier transform of φ. Because $\mathcal{F}(\psi) = \mathcal{F}(g)\mathcal{F}(f)$, we can estimate $\mathcal{F}(f)$ using $\mathcal{F}(\hat{\psi}_{\mathrm{K}})/\mathcal{F}(g)$ and obtain $\hat{f}_{\mathrm{F}}$ using the inverse Fourier transform.

The properties of the above deconvolution method have been studied extensively by, among others, Zhang (1990), Fan (1991, 1992), and Efromovich (1997). The kernel density deconvolution with heteroscedastic errors, that is, the ε_i have different densities g_i, is considered by Delaigle and Meister (2008). Without assuming that f has a *compact support*, the optimal rates of convergence for the nonparametric deconvolution are extremely slow, especially for supersmooth error distributions, including the normal distribution (e.g., see Carroll and Hall, 1988; Fan, 1991, 1992). Specifically, assuming that f has k bounded derivatives, Carroll and Hall (1988) proved that (i) if g is normal, then the best rate of any estimator of f is $\mathcal{O}((\log n)^{-k/2})$; (ii) if g is gamma with shape s, then the optimal rate is $\mathcal{O}(n^{-k/(2k+2s+1)})$; (iii) if g is double-gamma, that is, the absolute error is gamma with shape s, then the optimal rate is $\mathcal{O}(n^{-k/[2k+4(s-\lfloor s/2 \rfloor)]})$, where $\lfloor \cdot \rfloor$ is the floor function; and (iv) if g has compact support and infinitely many derivatives, then the optimal rate is slower than $\mathcal{O}(n^{-\eta})$, for any $\eta > 0$. However, if f has a compact support, then the density estimation based on the Bernstein polynomial model and the uncontaminated data $x_1, \ldots, x_n$ can achieve an almost parametric rate, such as $\mathcal{O}(\log^2 n/n)$ for the MISE (Guan, 2017). Guan (2016) showed a similar result under a differnt set of conditions. The kernel-type estimators for analytic densities can attain this type of rate (Stepanova, 2013). Under some regularity conditions, the best rate achievable by the parametric density estimate is $\mathcal{O}(n^{-1})$. Juditsky and Lambert-Lacroix (2004) identified a minimax rate of $\mathcal{O}(n^{-2k/(2k+1)})$ for the Hölder class of order k of univariate density functions, even restricted to $[0, 1]$ (see Ibragimov and Khasminskii, 1981). This is also the minimax rate for Sobolev class densities (e.g., see Schipper, 1996). In this study we consider a smaller class of density functions. For example, the density of a beta distribution with a non-integer shape is a member of the Hölder class, but does not belong to the class we study. Even if the errors have a supersmooth error distribution, such as the normal distribution, we expect that a nonparametric estimator of f based on the contaminated

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0173

data $y_1, \ldots, y_n$ and the Bernstein polynomial model to achieve a faster rate than $\mathcal{O}((\log n)^{-\eta})$, $\eta > 0$. Although Fan (1992) has shown that a nonparametric deconvolution with normal errors can be as good as a kernel density estimate based on uncontaminated data if the noise level decreases as the sample size increases, an accelerated denconvolution is still desirable. Recently, Delaigle and Hall (2014) proposed an improved kernel method on $\hat{f}_{\mathrm{F}}$ to speed up the convergence, assisted by a "close to being correct" parametric guess of f. The assumption of a known error density g is discussed by Horowitz and Markatou (1996), Efromovich (1997, 1999), and Neumann (1997).

We show that, using Bernstein-type polynomial, we can approximate the convolution density ψ in the additive measurement error model using a mixture model of known components but unknown mixture proportions. Consequently, we can estimate f using an approximate, maximum likelihood method. The resulting density estimate attains a much better convergence rate. This method differs from those in the literature, because it does not use Fourier transforms, and it can be viewed as a nearly parametric approach to a nonparametric density deconvolution. Like any finite-mixture model, this approximate model differs from classical parametric models because the number of parameters, the degree of the polynomial, is also unknown.

The remainder of the paper is organized as follows. In Section 2, we introduce and validate the Bernstein polynomial model for a nonparametric density estimation and density deconvolution. Methods for finding the maximum approximate likeli-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0173

hood estimates using an expectation-maximization (EM) algorithm and choosing the optimal degree m are also given in this section. Large-sample results are presented in Section 2.4. Simulation studies are conducted in Section 3 to compare the finite-sample performance of the proposed method and that of its competitors. We conclude this paper in Section 4. Additional simulation results, a real-data application, and the proofs are given in the online Supplementary Material.

2 Main Results

2.1 Approximate Bernstein Polynomial Model

Assume that the density f is continuous on $[0,1]$. Then, we have (Bernstein, 1912, 1932)

$$f(u) \approx B_m^f(u) = \frac{1}{m+1}\sum_{i=0}^{m} f(i/m)\beta_{mi}(u).$$

The best degree of approximation of f by B_m^f is $\mathcal{O}(m^{-1})$, no matter how smooth f is. From Lorentz (1963), if f has a positive lower bound and has higher continuous derivatives, then there exists a much better approximation $P_m^f(u) = \sum_{i=0}^m c_i\beta_{mi}(u)$, for $c_i \ge 0$. This is called a polynomial with positive coefficients, and enjoys a better degree of approximation than $\mathcal{O}(m^{-1})$. The coefficients c_i are refinements of those $f(i/m)/(m+1)$ in the classical Bernstein polynomial. Therefore, we have an approximate Bernstein polynomial model,

$$f_m(u;\boldsymbol{p}) = \sum_{i=0}^{m} p_i\beta_{mi}(u),$$

where $\boldsymbol{p} = (p_0, \ldots, p_m)^{\mathrm{T}} \in \mathbb{S}_m = \big\{(p_0, \ldots, p_m)^{\mathrm{T}} : p_i \geq 0, \quad \sum_{i=0}^m p_i = 1\big\}$, the m-simplex. The density ψ can be approximated by

$$\psi_m(y; \boldsymbol{p}) = (g * f_m)(y) = \sum_{i=0}^m p_i \psi_{mi}(y),$$

where $\psi_{mi}(y) = (g * \beta_{mi})(y) = \int_0^1 g(y-x)\beta_{mi}(x)dx$, for $i \in \mathbb{I}_0^m$. Therefore, the convolution ψ is approximately parameterized as a mixture of $\psi_{mi} = g * \beta_{mi}$, for $i \in \mathbb{I}_0^m$.

2.2 Maximum Likelihood Estimate

For a given m, the approximate Bernstein likelihood of $y_1, \ldots, y_n$ is defined as

$$\mathscr{L}(\boldsymbol{p}) = \prod_{j=1}^n \sum_{i=0}^m p_i (g * \beta_{mi})(y_j) \approx \prod_{j=1}^n \psi(y_j).$$

Thus, the approximate loglikelihood is $\ell(\boldsymbol{p}) = \sum_{j=1}^n \log \sum_{i=0}^m p_i (g * \beta_{mi})(y_j)$. The maximizer $\hat{\boldsymbol{p}}$ of $\ell(\boldsymbol{p})$ is called the maximum approximate Bernstein likelihood estimator (MABLE) of $\boldsymbol{p} = (p_0, \ldots, p_m)^{\mathrm{T}}$, the unknown mixture proportions. Then, we obtain an estimator $\hat{f}_{\mathrm{B}}(x) = f_m(x; \hat{\boldsymbol{p}})$ of f for an optimal degree m. The consequent density estimator $\hat{f}_{\mathrm{B}}(x)$ is called the MABLE of f. It is not surprising that $\hat{f}_{\mathrm{B}}$ outperforms the kernel density estimators, which do not take advantage of the conditions imposed on f in this study.

The EM algorithm (Dempster et al., 1977; Wu, 1983; Redner and Walker, 1984) applies to find $\hat{\boldsymbol{p}}$, leading to the iteration

$$p_i^{(s+1)} = p_i^{(s)} S_{mn}^{(i)}(\boldsymbol{p}^{(s)}), \quad i \in \mathbb{I}_0^m;\ s \in \mathbb{I}_0^\infty, \tag{1}$$

where

$$S_{mn}^{(i)}(\boldsymbol{p}) = \frac{\partial \ell(\boldsymbol{p})}{\partial p_i} = \frac{1}{n}\sum_{j=1}^{n} \frac{\psi_{mi}(y_j)}{\psi_m(y_j;\boldsymbol{p})}, \quad i \in \mathbb{I}_0^m.$$

Starting with positive initial $p_i^{(s)} > 0$, the convergence of $\boldsymbol{p}^{(s)} = (p_0^{(s)}, \ldots, p_m^{(s)})^{\mathrm{T}}$ to $\hat{\boldsymbol{p}}$ as $s \to \infty$ is guaranteed by Theorem 4.2 of Redner and Walker (1984). It is also clear that the resulting $\hat{f}_{\mathrm{B}}(x) = f_m(x;\hat{\boldsymbol{p}})$ is a bona fide density, because $\hat{\boldsymbol{p}} \in \mathbb{S}_m$.

In a nonparametric model, f is totally unspecified. If we have no information about the support of f, we can only estimate f as a density with support $[x_{(1)}, x_{(n)}]$, where $x_{(1)}$ and $x_{(n)}$ are the minimum and maximum order statistics, respectively, of a sample of size n from f. If the support S of f differs from [0, 1] and we can find a finite interval $[a,b] \subset S$, such that $[y_{(1)}, y_{(n)}] \subset [a,b]$ and $F(b) - F(a) \approx 1$, then we let $y_j^* = (y_j - a)/(b-a) = x_j^* + \varepsilon_j^*$, where $x_j^* = (x_j - a)/(b-a)$ and $\varepsilon_j^* = \varepsilon_j/(b-a)$. The densities of x_j^* and ε_j^* are $f^*(x) = (b-a)f[a+(b-a)x]$ and $g^*(x) = (b-a)g[(b-a)x]$, respectively. Let $\hat{f}_{\mathrm{B}}^*$ be an estimate of f^* based on y_j^*. Then, we can estimate f by $\hat{f}_{\mathrm{B}}(x) = \hat{f}_{\mathrm{B}}^*[(x-a)/(b-a)]/(b-a)$. Because $x_j = y_j - \varepsilon_j$ and the error distribution is known, we can choose (a,b) by properly extending $(y_{(1)}, y_{(n)})$, for example, $(a,b) = (y_{(1)} - \zeta\sigma_\varepsilon, y_{(n)} + \zeta\sigma_\varepsilon)$, for some $\zeta > 0$, where σ_ε is the standard deviation of the error ε.

2.3 Model Degree Selection

Denote the sample mean and variance of $y_1, \ldots, y_n$ as $\bar{y}$ and s^2, respectively. Because $\mu_0 = \mathrm{E}(\varepsilon) = 0$ and $\sigma_0^2 = \mathrm{E}(\varepsilon^2)$ are known, we estimate $\mu = \mathrm{E}(X)$ and $\sigma^2 = \mathrm{var}(X)$ by $\hat{\mu} = \bar{y}$ and $\hat{\sigma}^2 = s^2 - \sigma_0^2$, respectively. As in Guan (2016), we estimate the

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0173

lower bound m_b for m by $\hat{m}_b = \max\{\lceil\hat{\mu}(1-\hat{\mu})/\hat{\sigma}^2-3\rceil, 1\}$. Based on $\hat{m}_b$ we choose an appropriate m_0 and a large positive integer K to form $\mathcal{M} = \{m_i = m_0 + i,\, i \in \mathbb{I}_0^K\}$. Denote $\ell_i = \ell(\hat{\boldsymbol{p}}_{m_i})$, for $i \in \mathbb{I}_1^K$. Because $f_m(u;\boldsymbol{p}_m)$ is nested in $f_{m+1}(u;\boldsymbol{p}_{m+1})$ (Guan, 2016), $\psi_m(u;\boldsymbol{p}_m)$ is in $\psi_{m+1}(u;\boldsymbol{p}_{m+1})$. Thus, $u_i = \ell_i - \ell_{i-1}$, for $i \in \mathbb{I}_1^K$, are nonnegative. From our real-data analysis and extensive simulation studies we observed that for large K, the optimal degree m_q corresponds to a change-point q such that $\{u_{q+1},\ldots,u_K\}$ have a smaller mean and variance than those of $\{u_1,\ldots,u_q\}$. We can treat $u_1,\ldots,u_K$ as exponential observations. The change-point q can be estimated (see §1·4 of Csörgő and Horváth, 1997) by $\hat{q} = \arg\max_{1\le q<K}\{R(q)\}$, where $R(q) = K\log[(\ell_K-\ell_0)/K] - q\log[(\ell_q-\ell_0)/q] - (K-q)\log[(\ell_K-\ell_q)/(K-q)]$. Having obtained $\hat{\boldsymbol{p}}_m = (\hat{p}_0,\ldots,\hat{p}_m)^{\mathrm{T}}$, we can use $p_i^{(0)} = [i\hat{p}_{i-1} + (m-i+1)\hat{p}_i]/(m+1)$, for $i \in \mathbb{I}_0^{m+1}$, as the initial guess for iteration (1) for $\hat{\boldsymbol{p}}_{m+1}$.

2.4 Asymptotic Results

2.4.1 Information matrices

The information matrix for model $\psi_m(y;\boldsymbol{p})$ is $\mathcal{I}_m(f,g) = [I_m^{ij}(f,g)]$, where

$$I_m^{ij}(f,g) = \int_{S_\psi} \frac{\psi_{mi}(y)\psi_{mj}(y)}{\psi(y)}dy, \quad i,j \in \mathbb{I}_0^m,$$

and $S_\psi = \{y \in R : \psi(y) > 0\}$. The information matrix for $f_m(x;\boldsymbol{p})$ is $\mathcal{I}_m(f,\delta) = [I_m^{ij}(f,\delta)]$, where δ is the Dirac delta, and $I_m^{ij}(f,\delta) = \int_0^1[\beta_{mi}(x)\beta_{mj}(x)/f(x)]dx$, for $i,j \in \mathbb{I}_0^m$. It is easy to see that $\mathcal{I}_m(f,\delta) - \mathcal{I}_m(f,g)$ is positive-definite; that is,

$\mathcal{I}_m(f,\delta) \geq \mathcal{I}_m(f,g)$. Note that $\beta_{mi}(x) \leq m+1$. If $f \geq b_0 > 0$, then

$$I_m^{ij}(f,g) = \int_{S_\psi} \frac{\psi_{mi}(y)\psi_{mj}(y)}{\psi(y)} dy \leq \frac{m+1}{b_0}, \quad i,j \in \mathbb{I}_0^m. \tag{2}$$

Define an approximate Fisher information $\mathcal{J}_m(\boldsymbol{p};f,g) = [J_m^{ij}(\boldsymbol{p};f,g)]$, where

$$J_m^{ij}(\boldsymbol{p};f,g) = \int_{S_\psi} \frac{\psi_{mi}(y)\psi_{mj}(y)}{\psi_m^2(y;\boldsymbol{p})} \psi(y) dy, \quad i,j \in \mathbb{I}_0^m.$$

At $\boldsymbol{p} = \boldsymbol{p}_0$, $\mathcal{J}_m(\boldsymbol{p}_0;f,g)$ is an approximation of the "ultimate" Fisher information $\mathcal{I}_m(f,g)$, where $\boldsymbol{p}_0$ satisfies the assumption (A.1) or (A.2) below.

2.4.2 Assumptions

We show our asymptotic results assuming $f_m(u;\boldsymbol{p}) = \sum_{i=0}^m p_i\beta_{mi}(u)$ 1s an approximate model rather than an exact parametric model. We denote the chi-square divergence (χ^2-distance) between densities φ and ψ by

$$\chi^2(\varphi\|\psi) = \int_{S_\psi} \frac{[\varphi(y)-\psi(y)]^2}{\psi(y)} dy \equiv \int_{S_\psi} \left[\frac{\varphi}{\psi}(y) - 1\right]^2 \psi(y) dy.$$

Proposition 1. *For the divergence $\chi^2(h*g\|f*g)$ between convolutions with the same g, we have the following results. (i) $\chi^2(h*g\|f*g) \leq \chi^2(h\|f)$. In particular, if $g = \delta$, the Dirac delta, then $\chi^2(h*g\|f*g) = \chi^2(h\|f)$. (ii) The divergence $D^2(h\|f) = \chi^2(h*g\|f*g)$ is also a divergence of h from f; that is, $\chi^2(h*g\|f*g) \geq 0$ and $\chi^2(h*g\|f*g) = 0$ iff $h(x) = f(x)$, almost everywhere.*

Define $D^2(\boldsymbol{p}) = \chi^2(\psi_m(\cdot;\boldsymbol{p})\|\psi) = \int(\psi_m^2/\psi)(y)dy - 1$. We need the following assumptions for the asymptotic properties of $\hat{f}_{\mathrm{B}}$ (proved in the appendix):

(A.1). *There exist $\boldsymbol{p}_0 \in \mathbb{S}_m$ and $k > 0$ such that, uniformly in y such that $\psi(y) > 0$,*

$$\frac{\psi_m(y; \boldsymbol{p}_0) - \psi(y)}{\psi(y)} = \mathcal{O}(m^{-k/2})$$

and, thus, $D^2(\boldsymbol{p}_0) = \mathcal{O}(m^{-k})$.

(A.2). *There exist $\boldsymbol{p}_0 \in \mathbb{S}_m$ and $k > 0$ such that, uniformly in $x \in (0, 1)$,*

$$\frac{f_m(x; \boldsymbol{p}_0) - f(x)}{f(x)} = \mathcal{O}(m^{-k/2})$$

and, thus, $\chi^2(f_m(\cdot; \boldsymbol{p}_0) \| f) = \mathcal{O}(m^{-k})$.

Clearly, (A.2) is a stronger assumption, and implies (A.1) because ψ_m and ψ are convolutions of f_m and f with g, respectively.

2.4.3 Sufficient Conditions for Assumptions (A.1) and (A.2)

Let $C^{(r)}[0, 1]$ be the class of functions that have an rth continuous derivative $f^{(r)}$ on $[0, 1]$. A function f is said to be *α-Hölder continuous* with $\alpha \in (0, 1]$ if $|f(x)-f(y)| \leq C|x-y|^{\alpha}$, for some constant $C > 0$. For a density function g on $(-\infty, \infty)$, let $\psi(y)$ and $\psi_m(y; \boldsymbol{p})$ be the convolutions of g with f and f_m, respectively. We have the following result.

Theorem 1. *Suppose that $f(x) = x^a(1-x)^b f_0(x)$ is a density on $[0, 1]$, a and b are nonnegative integers, $f_0 \in C^{(r)}[0, 1]$, $r \geq 0$, $f_0(x) \geq b_0 > 0$, and $f_0^{(r)}$ is α-Hölder continuous, with $\alpha \in (0, 1]$. Then, both assumptions (A.1) and (A.2) are fulfilled, with $k = r + \alpha$.*

This is a generalization of the result of Lorentz (1963), which requires a positive lower bound for f.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0173

2.4.4 Rate of Convergence

Theorem 2. *Under Assumption (A.1) with $k > 0$, as $n \to \infty$, with probability one, the maximum value of $\ell(\boldsymbol{p})$ with $m = \mathcal{O}(n^{1/k})$ is attained at $\hat{\boldsymbol{p}}$ in the interior of $\mathbb{B}_m(r_n) = \{\boldsymbol{p} \in \mathbb{S}_m : D^2(\boldsymbol{p}) \le r_n\}$, where $r_n = n^{-1}\log n$ and the mean χ^2-distance between $\hat{\psi}_m(\cdot) = \psi_m(\cdot; \hat{\boldsymbol{p}})$ and $\psi(\cdot)$ satisfies*

$$D^2(\hat{\boldsymbol{p}}) = \int_{S_\psi} \frac{[\psi_m(y;\hat{\boldsymbol{p}}) - \psi(y)]^2}{\psi(y)} dy < \frac{\log n}{n}, \quad a.s. \tag{3}$$

Remark 1. *If $g = \delta$, the Dirac delta, then under (A.1), with $m = \mathcal{O}(n^{1/k})$, (3) is true for $\psi = f$ and $\psi_m = f_m$.*

Remark 2. *If $M_0 = \max_{-\infty \le y \le \infty} \psi(y) < \infty$, then Theorem 2 implies*

$$\|\hat{\psi}_m - \psi\|_2^2 = \int_{-\infty}^{\infty} [\psi_m(y;\hat{\boldsymbol{p}}) - \psi(y)]^2 dy \le M_0^{-1} \frac{\log n}{n}, \ a.s.$$

By (A.1) we also have, for some constant C,

$$(\hat{\boldsymbol{p}} - \boldsymbol{p}_0)^{\mathrm{T}} \mathcal{I}_m(f,g)(\hat{\boldsymbol{p}} - \boldsymbol{p}_0) = \int_{S_\psi} \frac{[\psi_m(y;\hat{\boldsymbol{p}}) - \psi_m(y;\boldsymbol{p}_0)]^2}{\psi(y)} dy \le C \frac{\log n}{n}, \ a.s.$$

Although the chi-square divergence (3) is also a divergence of $\hat{f}_{\mathrm{B}}$ from f, we are more interested in the rate of convergence for $\mathrm{MISE}(\hat{f}_{\mathrm{B}}, f) = \mathrm{E}\int_0^1 [\hat{f}_{\mathrm{B}}(x) - f(x)]^2 dx$ or $\mathrm{E}\chi^2(\hat{f}_{\mathrm{B}} \| f) = \mathrm{E}\int_0^1 [\hat{f}_{\mathrm{B}}(x) - f(x)]^2 / f(x) dx$, which is affected by the error density g. From (A.1) and (A.2), it suffices to investigate the rate of the (weighted) integrated squared error of $\hat{f}_{\mathrm{B}}(x) = f_m(x;\hat{\boldsymbol{p}})$ as an estimator of $f_m(x;\boldsymbol{p}_0)$:

$$\int_0^1 \frac{[f_m(x;\hat{\boldsymbol{p}}) - f_m(x;\boldsymbol{p}_0)]^2}{f(x)} dx = (\hat{\boldsymbol{p}} - \boldsymbol{p}_0)^{\mathrm{T}} \mathcal{I}_m(f,\delta)(\hat{\boldsymbol{p}} - \boldsymbol{p}_0).$$

Theorem 3. *Suppose that the sufficient conditions of Theorem 1 are fulfilled, with $a = b = 0$ and $k = r + \alpha > 4$. Then, a.s., with $m = \mathcal{O}(n^{1/k})$,*

$$(\hat{\boldsymbol{p}} - \boldsymbol{p}_0)^{\mathrm{T}} \mathcal{I}_m(f, \delta)(\hat{\boldsymbol{p}} - \boldsymbol{p}_0) = \mathbb{R}_n^{\mathrm{T}} \Omega_m(f, g) \mathbb{R}_n, \tag{4}$$

$$(\hat{\boldsymbol{p}} - \boldsymbol{p}_0)^{\mathrm{T}} \mathcal{I}_m(1, \delta)(\hat{\boldsymbol{p}} - \boldsymbol{p}_0) = \mathbb{R}_n^{\mathrm{T}} \tilde{\Omega}_m(f, g) \mathbb{R}_n, \tag{5}$$

where $\Omega_m(f,g) = \mathcal{I}_m^{-1}(f,g)\mathcal{I}_m(f,\delta)\mathcal{I}_m^{-1}(f,g)$, $\tilde{\Omega}_m(f,g) = \mathcal{I}_m^{-1}(f,g)\mathcal{I}_m(1,\delta)\mathcal{I}_m^{-1}(f,g)$, and $\mathbb{R}_n = \mathcal{I}_m(f,g)(\hat{\boldsymbol{p}} - \boldsymbol{p}_0)$ satisfies

$$\|\mathbb{R}_n\|^2 = \mathcal{O}\big(n^{-1+4/k} \log n\big). \tag{6}$$

Remark 3. *In addition to the conditions of Theorem 3, if the largest eigenvalues of $\Omega_m(f,g)$ and $\tilde{\Omega}_m(f,g)$ are of order $\mathcal{O}(m^\gamma)$, for some $\gamma < k - 4$, then, a.s., the rates in (4) and (5) are $\mathcal{O}(n^{-(k-4-\gamma)/k} \log n)$.*

In many cases, including the supersmooth error densities, we have the following result.

Theorem 4. *In addition to the conditions of Theorem 3, with $k = r + \alpha > 5$, if the error density g is nonvanishing, nonincreasing on $(0, \infty)$, and nondecreasing on $(-\infty, 0)$, then the largest eigenvalues λ_m and $\tilde{\lambda}_m$ of $\Omega_m(f,g)$ and $\tilde{\Omega}_m(f,g)$, respectively, satisfy $\lambda_m = \mathcal{O}(m)$ and $\tilde{\lambda}_m = \mathcal{O}(m)$. Therefore, with $m = \mathcal{O}(n^{1/k})$, a.s.,*

$$\int_0^1 \frac{[f_m(x;\hat{\boldsymbol{p}}) - f(x)]^2}{f(x)} dx = \mathcal{O}\big(n^{-(k-5)/k} \log n\big), \tag{7}$$

$$\int_0^1 [f_m(x;\hat{\boldsymbol{p}}) - f(x)]^2 dx = \mathcal{O}\big(n^{-(k-5)/k} \log n\big). \tag{8}$$

Remark 4. *From the proof of this theorem, we see that the condition that g is nonvanishing can be relaxed to require*

$$C_0 = \min\{g^2(-1), g^2(1)\} + \int_{-\infty}^{-1} g^2(y)dy + \int_1^{\infty} g^2(y)dy > 0.$$

Thus, even if g has a compact support $[-1,1]$, such that $C_0 = \min\{g^2(-1), g^2(1)\} > 0$, we still have a rate $\mathcal{O}(n^{-(k-5)/k}\log n)$, which is better than $\mathcal{O}(n^{-\eta})$, for some $\eta > 0$, whenever $k > 5$.

3 Simulation

In this section, we conduct a simulation to compare the finite-sample performance of the proposed estimator $\hat{f}_{\mathrm{B}}$ with the parametric deconvolution $\hat{f}_{\mathrm{P}}$, the Fourier transform estimator $\hat{f}_{\mathrm{F}}$ based on contaminated data, and the kernel density $\tilde{f}_{\mathrm{K}}$ based on uncontaminated data. The data $x_1, \ldots, x_n$ of size $n = 100, 200, 400$ are generated from f. The errors ε follow normal $\mathrm{N}(0, \sigma_0^2)$ and $\mathrm{L}(0, \sigma_0)$, for some selected σ_0. Only when σ_0 is "small" relative to the standard deviation σ of X can we obtain an applicable estimate of f, even for a parametric deconvolution. For instance, if both X and ε are normal, then the maximum likelihood estimates of $\mu = \mathrm{E}(X)$ and $\sigma^2 = \mathrm{var}(X)$ are, respectively, $\hat{\mu} = \bar{y}$ and $\hat{\sigma}^2 = \max\{(n-1)s^2/n - \sigma_0^2, 0\}$, where $\bar{y}$ and s^2 are the sample mean and sample variance, respectively, of $y_1, \ldots, y_n$. If $\sigma < \sigma_0$ and n is not large, then $\hat{\sigma}^2$ could also be zero because $\mathrm{pr}\{(n-1)s^2/n < \sigma_0^2\} = \mathrm{pr}[\chi_{n-1}^2 < n/\{1 + (\sigma/\sigma_0)^2\}]$ is not small. Therefore the parametric deconvolution $\hat{f}_{\mathrm{P}}(x)$ may be degenerate because $\hat{\sigma}^2$ could be zero, even

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0173

if $\sigma_0 \leq \sigma$.

We used the R package decon (Wang and Wang, 2011), which implements the methods of Fan (1991, 1992), Delaigle and Gijbels (2004), and Delaigle and Meister (2008) to calculate $\hat{f}_{\mathrm{F}}$. The "dboot2" method was used to choose an optimal bandwidth h (see Delaigle and Gijbels, 2004 and Wang and Wang, 2011 for details).

After R Monte Carlo runs, we obtained $\hat{f}^{(1)}, \ldots, \hat{f}^{(R)}$ for the estimator $\hat{f}$. We then approximated the point-wise mean squared error at $x \in [a, b]$, $\mathrm{pMSE}[\hat{f}(x)] = \mathrm{E}[\hat{f}(x) - f(x)]^2 = \mathrm{var}[\hat{f}(x)] + \mathrm{Bias}^2[\hat{f}(x)]$, by $\widehat{\mathrm{pMSE}}[\hat{f}(x)] = \hat{\sigma}^2[\hat{f}(x)] + \{\hat{\mu}[\hat{f}(x)] - f(x)\}^2$, where $\hat{\mu}[\hat{f}(x)]$ and $\hat{\sigma}^2[\hat{f}(x)]$ are the sample mean and the sample variance, respectively, of $\hat{f}^{(1)}(x), \ldots, \hat{f}^{(R)}(x)$. The MISE $\mathrm{MISE}(\hat{f}) = \mathrm{E}\int[\hat{f}(x) - f(x)]^2 dx$ is estimated by $\widehat{\mathrm{MISE}}(\hat{f}) = \sum_{i=1}^{N} \widehat{\mathrm{pMSE}}[\hat{f}(t_i)]\Delta t$, where $\Delta t = (b-a)/N$, $t_i = a + i\Delta t$, for $i \in \mathbb{I}_0^N$, and $N = 512$.

In the simulation results shown in Table 1, as in Fan (1992), we have (i) unimodal $f = \mathrm{N}(0, 1)$ truncated by $[a, b] = [-7, 7]$, and (ii) bimodal $f = 0.6\mathrm{N}(-2, 1^2) + 0.4\mathrm{N}(2, 0.8^2)$ truncated by $[a, b] = [-7, 7]$. The errors ε are taken from $\mathrm{N}(0, \sigma_0^2)$ and $\mathrm{L}(0, \sigma_0)$, with $\sigma_0 = 0.2, 0.4, 0.6, 0.8, 1.0$. Furthermore, $\hat{f}_{\mathrm{P}}(x)$ is the parametric estimate of the density of $\mathrm{N}(\mu, \sigma^2)$ and $\lambda\ \mathrm{N}(\mu_1, \sigma_1^2) + (1-\lambda)\mathrm{N}(\mu_2, \sigma_2^2)$, with known variances σ^2, σ_1^2, and σ_2^2, but unknown μ and (λ, μ_1, μ_2). The rate of convergence in the MISE of such a parametric estimator is $\mathcal{O}(n^{-1})$.

In order to compare the proposed estimator $\hat{f}_{\mathrm{B}}$ with $\hat{f}_{\mathrm{F}}$, $\hat{f}_{\mathrm{P}}$, and $\tilde{f}_{\mathrm{K}}$, we plot the point-wise mean squared error in Figure 1, from which we see that $\hat{f}_{\mathrm{B}}$ almost uniformly outperforms $\hat{f}_{\mathrm{F}}$ for both unimodal and bimodal f. We also see that if f

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0173

is unimodal and smooth enough that k is large, as in Theorems 2 and 3, then $\hat{f}_{\mathrm{B}}$ almost uniformly outperforms $\tilde{f}_{\mathrm{K}}$, which is based on the clean data.

The (mixture) normal density f has continuous kth derivative $f^{(k)}$, for all k. In practice, a random variable may have an approximate normal distribution, supported by the central limit theorem and some goodness-of-fit test.

In the second simulation study, presented in Table 2, we generated the sample $x_1, \ldots, x_n$ from a "nearly normal" distribution NN(d), for $d = 4$, with the distribution of the sample mean $\bar{u}_d$ of $u_1, \ldots, u_d$ taken from uniform(0, 1). The errors $\varepsilon_1, \ldots, \varepsilon_n$ were generated from $\mathrm{N}(0, \sigma_0^2)$ and $\mathrm{L}(0, \sigma_0)$, where $12d\sigma_0^2 = 0.2^2$, 0.4^2, 0.6^2, 0.8^2, 1.0^2 and $d = 4$. Thus, $\sigma_0 = (0.05, 0.10, 0.15, 0.20, 0.25)/\sqrt{3}$.

The central limit theorem shows that $\mathrm{NN}(d) \approx \mathrm{N}(1/2, 1/(12d))$ for large d. Let $\rho(n)$ denote the probability that the Shapiro-test based on a sample of size n rejects the normality of NN(4) with significance level 0.05. Based on 5,000 Monte Carlo runs, $\rho(n)$ is estimated to be 0.0398, 0.0504, and 0.0966 for $n = 100, 200$, and 400, respectively. Let φ_d be the density of NN(d). If $d \geq 2$, then $\varphi_d^{(d-2)}$ is 1-Hölder continuous, but $\varphi_d \notin C^{(d-1)}[0, 1]$. We also have $\varphi_d(x) = [x(1-x)]^{d-1} h_d(x)$, where $h_d(x) \geq d^d/(d-1)!$ on [0,1], and $h_d^{(d-2)}$ is 1-Hölder continuous, but $h_d \notin C^{(d-1)}[0, 1]$. Therefore, the conditions of Theorem 1 are satisfied with $a = b = d - 1$, $f_0 = h_d$, $b_0 = d^d/(d-1)!$, $\alpha = 1$, and $r = d - 2$. The parametric estimate $\hat{f}_{\mathrm{P}}$ of φ_d in this simulation is based on the normal model $\mathrm{N}(\mu, \sigma^2)$, with known $\sigma^2 = 1/(12d)$ and a known error density. For φ_4, we have $k = r + \alpha = 3$. Although in this case the conditions $a = b = 0$ and $k > 4$ of Theorem 3 are not fulfilled, the proposed

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0173

estimator $\hat{f}_{\mathrm{B}}$ still performs better than $\hat{f}_{\mathrm{F}}$ in such a bad scenario. In this case, $\hat{f}_{\mathrm{B}}$ performs worse than $\tilde{f}_{\mathrm{K}}$, which is based on the *uncontaminated data* and has a rate of $\mathcal{O}(n^{-4/5})$.

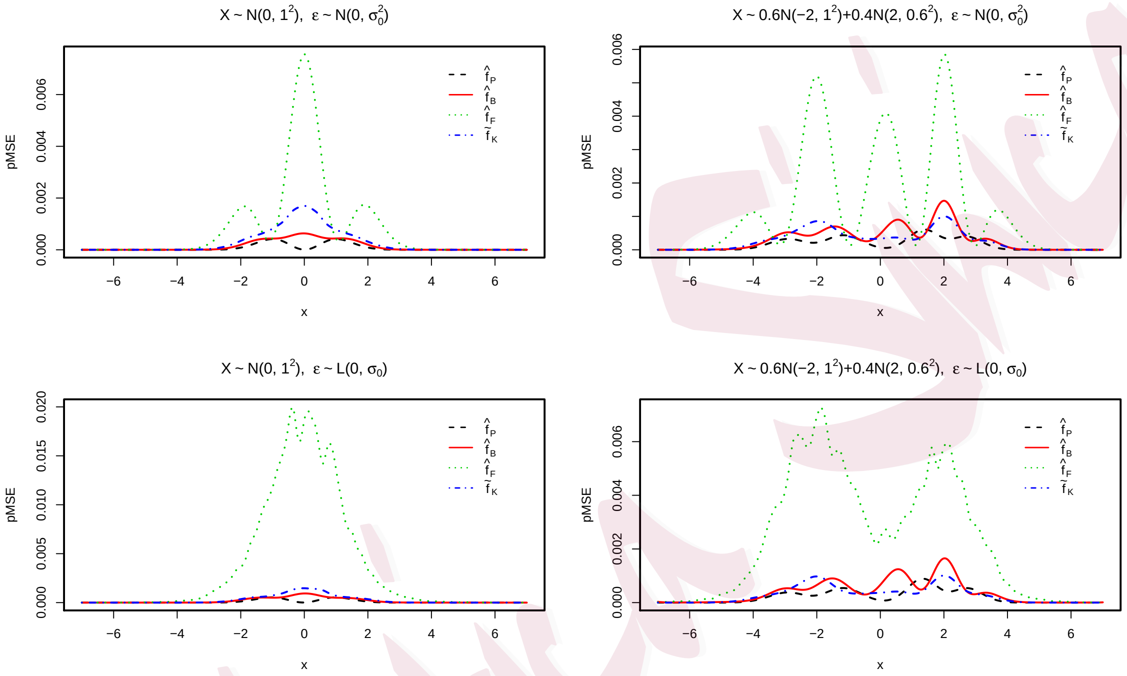

Figure 1: The simulated point-wise mean squared errors of the parametric estimator $\hat{f}_{\mathrm{P}}$(dashed), proposed estimator $\hat{f}_{\mathrm{B}}$(solid), inverse Fourier estimator $\hat{f}_{\mathrm{F}}$(dotted), and kernel density estimator based on the simulated uncontaminated data $\tilde{f}_{\mathrm{K}}$(dotdash), at $t_i = a + i(b-a)/N$ $(i = 0, 1, \ldots, N,\ N = 512)$. The sample size is $n = 200$. The truncation interval is $[a, b] = [-7, 7]$. In the parametric models, all variances are assumed to be known. $\mathcal{M} = \{10, 11, \ldots, 100\}$. Upper panels: the error distribution is $\mathrm{N}(0, \sigma_0^2)$, with $\sigma_0 = 0.6$; Lower panels: the error distribution is Laplace $\mathrm{L}(0, \sigma_0)$, with $\sigma_0 = 0.6$.

Tables 1 and 2 (the case $n = 200$ is provided in the Supplementary Material) show

Table 1: The square root multiplied by 100 of the mean integrated squared error. $\hat{f}_{\mathrm{P}}$, the parametric estimator; $\hat{f}_{\mathrm{B}}$, the proposed estimator; $\hat{f}_{\mathrm{F}}$, the inverse Fourier estimator; $\tilde{f}_{\mathrm{K}}$, the kernel density estimator based on uncontaminated data and 1000 Monte Carlo runs, with $x_1, \ldots, x_n$ generated from normal and mixture normal distributions, and the errors $\varepsilon_1, \ldots, \varepsilon_n$ from $\mathrm{N}(0, \sigma_0^2)$ and $\mathrm{L}(0, \sigma_0)$. In the parametric models, the variances are assumed to be known. $\mathcal{M} = \{10, 11, \ldots, 100\}$.

	$X \sim \mathrm{N}(0,1)$					$X \sim 0.6\mathrm{N}(-2,1) + 0.4\mathrm{N}(2, 0.8^2)$				
σ_0	0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0
					$\varepsilon \sim \mathrm{N}(0, \sigma_0^2)$					
					$n = 100$					
$\hat{f}_{\mathrm{P}}$	3.96	4.08	4.44	4.93	5.35	5.97	6.42	6.83	7.64	8.97
$\hat{f}_{\mathrm{B}}$	5.26	5.54	5.79	6.24	6.94	7.36	7.82	8.43	9.34	10.93
$\hat{f}_{\mathrm{F}}$	9.40	11.03	13.13	15.90	19.23	9.10	14.74	16.42	17.98	19.37
$\tilde{f}_{\mathrm{K}}$	8.14	8.37	8.29	8.21	8.19	8.26	8.35	8.27	8.11	8.26
					$n = 400$					
$\hat{f}_{\mathrm{P}}$	1.87	2.04	2.12	2.47	2.64	2.96	2.75	3.46	3.81	4.45
$\hat{f}_{\mathrm{B}}$	2.65	2.90	3.14	3.53	3.83	4.74	4.70	5.46	6.05	6.88
$\hat{f}_{\mathrm{F}}$	5.90	7.25	9.36	11.91	14.69	5.56	8.84	11.17	13.36	15.42
$\tilde{f}_{\mathrm{K}}$	4.77	4.72	4.66	4.70	4.71	4.79	4.63	4.84	4.74	4.77
					$\varepsilon \sim \mathrm{L}(0, \sigma_0)$					
					$n = 100$					
$\hat{f}_{\mathrm{P}}$	4.24	4.36	5.01	5.68	5.97	6.02	6.55	7.59	8.92	10.49
$\hat{f}_{\mathrm{B}}$	5.47	5.71	6.58	7.39	8.27	7.41	7.99	9.42	10.61	12.24
$\hat{f}_{\mathrm{F}}$	14.17	19.17	24.17	27.66	30.59	12.92	16.76	20.42	23.56	26.91
$\tilde{f}_{\mathrm{K}}$	8.48	8.21	8.36	8.37	7.98	8.15	8.22	8.25	8.18	8.17
					$n = 400$					
$\hat{f}_{\mathrm{P}}$	1.97	2.07	2.35	2.94	2.98	2.97	3.26	3.80	4.45	5.36
$\hat{f}_{\mathrm{B}}$	2.73	3.06	3.72	4.21	4.69	4.87	5.26	5.91	6.78	7.77
$\hat{f}_{\mathrm{F}}$	9.48	15.10	20.28	24.85	27.71	8.47	12.26	15.76	19.92	22.62
$\tilde{f}_{\mathrm{K}}$	4.79	4.64	4.70	4.75	4.58	4.89	4.81	4.84	4.79	4.82

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0173

Table 2: The square root multiplied by 100 of the mean integrated squared error. $\hat{f}_{\mathrm{P}}$, the parametric estimator; $\hat{f}_{\mathrm{B}}$, the proposed estimator; $\hat{f}_{\mathrm{F}}$, the inverse Fourier estimator; $\tilde{f}_{\mathrm{K}}$, the kernel density estimator based on uncontaminated data 1000 Monte Carlo runs, with $x_1, \ldots, x_n$ generated from the nearly normal distribution NN(4), and the errors $\varepsilon_1, \ldots, \varepsilon_n$ from the normal $\mathrm{N}(0, \sigma_0^2)$ and Laplace $\mathrm{L}(0, \sigma_0)$. We assume the normal distribution $\mathrm{N}(\mu, \sigma^2)$ with known variance $\sigma^2 = 1/48$ as the parametric model. $\mathcal{M} = \{2, 3, \ldots, 100\}$.

	$X \sim \mathrm{NN}(4)$, $\varepsilon \sim \mathrm{N}(0, \sigma_0^2)$					$X \sim \mathrm{NN}(4)$, $\varepsilon \sim \mathrm{L}(0, \sigma_0)$				
$\sqrt{3}\sigma_0$	0.05	0.10	0.15	0.20	0.25	0.05	0.10	0.15	0.20	0.25
					$n = 100$					
$\hat{f}_{\mathrm{P}}$	10.80	11.22	12.07	12.98	14.11	10.85	11.89	13.89	15.26	16.55
$\hat{f}_{\mathrm{B}}$	23.07	26.24	30.81	36.48	42.65	23.42	28.28	35.98	41.29	47.36
$\hat{f}_{\mathrm{F}}$	24.21	54.71	87.24	95.17	95.71	28.31	34.25	39.92	44.97	49.48
$\tilde{f}_{\mathrm{K}}$	21.70	20.79	21.28	21.21	20.92	21.02	21.51	21.00	21.64	21.06
					$n = 400$					
$\hat{f}_{\mathrm{P}}$	6.83	7.06	7.42	8.01	8.26	6.86	6.95	7.74	8.43	9.18
$\hat{f}_{\mathrm{B}}$	12.22	14.34	18.20	22.45	38.59	12.82	16.40	21.72	25.71	30.33
$\hat{f}_{\mathrm{F}}$	16.91	45.52	75.44	93.35	95.80	16.08	21.04	26.35	30.40	34.48
$\tilde{f}_{\mathrm{K}}$	12.23	12.13	12.16	12.26	12.27	11.98	12.08	12.13	12.01	12.16

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0173

that the proposed $\hat{f}_{\mathrm{B}}$ outperforms the Fourier transform method $\hat{f}_{\mathrm{F}}$. In some cases, especially when σ_0 is much smaller than σ, $\hat{f}_{\mathrm{B}}$ is three times as efficient as $\tilde{f}_{\mathrm{K}}$ in terms of the square root of the MISE. Although the simulation setup prefers the parametric methods, the results show that in most cases, the proposed approach has an MISE that leans toward the parametric one. Moreover, the proposed method outperforms the kernel estimate based on the uncontaminated data for the unimodal model or if the magnitude of the error variance is not too large. Because of the involvement of $\tilde{f}_{\mathrm{K}}$ in the comparison, it is not necessary to include any other kernel methods that improve upon $\hat{f}_{\mathrm{F}}$ in the simulation. Table 3 in the Supplementary Material presents simulation results for samples were from beta(3.5, 5.5), which belongs the Hölder class, as in Juditsky and Lambert-Lacroix (2004), and violates the assumptions of Theorem 1. This simulation shows that the proposed estimation could outperform the kernel method, even when the assumptions of Theorem 1 are not fulfilled.

4 Discussion

As shown in the Theorems in Section 2.4 and the simulation results, the performance of the proposed method leans toward that of the parametric approach when the correct parametric model is specified. The classical *exact* parametric method is subject to model misspecification. Our approach is an approximate parametric solution to a nonparametric problem, and speeds up the density deconvolution sig-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0173

nificantly, with a computation cost paid for searching an optimal model degree m, of course, under the assumption that the underlying unknown density has a positive lower bound on a known compact support. The condition imposed on the error distribution is satisfied by the family of generalized normal distributions, which include the supersmooth normal distribution and the ordinary smooth Laplace distribution. From our real-data example, we see that when replicated observations are available for estimating the error density g, a better nonparametric estimator of g is desirable. It is our intention to apply the Bernstein model to solve this problem.

As commented in Remark 1, the special case of Theorem 2 with $g = \delta$ is an enhancement of Theorem 4.1 of Guan (2016) and Theorem 3.3 of Guan (2017), which require a positive lower bound for the underlying density f. The simulation studies indicate there is possible room to improve the result of Theorem 4.

We have assumed that the underlying density f is continuously differentiable on [0,1] and has bounded support. Therefore, caution should be exercised, because this restriction might limit the application of the proposed method. Thus, further generalization and a sensitivity analysis of the proposed model is required.

5 Supplementary Material

The online Supplementary Material contains a real-data application, some additional simulation results, and technical proofs.

Acknowledgements

The author would like to thank Professor Jianqing Fan who brought the density deconvolution to his attention. He is also grateful to the Editor, Associate Editor, and anonymous reviewers for their careful reading and valuable comments and suggestions.

References

Bernstein, S. N. (1912). Démonstration du théorème de Weierstrass fondée sur le calcul des probabilitiés. *Communications of the Kharkov Mathematical Society* **13**, 1-2.

Bernstein, S. N. (1932). Complétement à l'article de E. Voronowskaja. *C. R. Acad. Sci. U.R.S.S.* 86-92.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1998). *Efficient and adaptive estimation for semiparametric models*. Springer-Verlag, New York.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association* **71**, 791–799.

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

Csörgő, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. John Wiley & Sons Inc., New York, 1st edition.

Delaigle, A. and Gijbels, I. (2004). Bootstrap bandwidth selection in kernel density

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0173

estimation from a contaminated sample. *Ann. Inst. Statist. Math.* **56**, 19–47.

Delaigle, A. and Hall, P. (2014). Parametrically assisted nonparametric estimation of a density in the deconvolution problem. *Journal of the American Statistical Association* **109**, 717-729.

Delaigle, A. and Meister, A. (2008). Density estimation with heteroscedastic error. *Bernoulli* **14**, 562–579.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**, 1–38.

Devroye, L. (1989). Consistent deconvolution in density estimation. *Canad. J. Statist.* **17**, 235–239.

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

Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory, and Applications*. Springer Series in Statistics, Springer, New York.

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

Fan, J. (1992). Deconvolution with supersmooth distributions. *Canad. J. Statist.* **20**, 155–169.

Grenander, U. (1981). *Abstract inference*. John Wiley & Sons Inc., New York. Wiley Series in Probability and Mathematical Statistics.

Guan, Z. (2016). Efficient and robust density estimation using Bernstein type polynomials. *Journal of Nonparametric Statistics* **28**, 250-271.

Guan, Z. (2017). Bernstein polynomial model for grouped continuous data. *Journal of Nonparametric Statistics* **29**, 831-848.

Horowitz, J. L. and Markatou, M. (1996). Semiparametric estimation of regression models for panel data. *The Review of Economic Studies* **63**, pp. 145-168.

Ibragimov, I. and Khasminskii, R. (1981). *Statistical estimation*, volume 16 of *Applications of Mathematics*. Springer-Verlag, New York-Berlin.

Ibragimov, I. and Khasminskii, R. (1983). Estimation of distribution density belonging to a class of entire functions. *Theory of Probability & Its Applications* **27**, 551-562.

Juditsky, A. and Lambert-Lacroix, S. (2004). On minimax density estimation on $\mathbb{R}$. *Bernoulli* **10**, 187–220.

Lorentz, G. G. (1963). The degree of approximation by polynomials with positive coefficients. *Mathematische Annalen* **151**, 239–251.

Neumann, M. H. (1997). On the effect of estimating the error density in nonparametric deconvolution. *Journal of Nonparametric Statistics* **7**, 307–330.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* **26**, 195–239.

Schipper, M. (1996). Optimal rates and constants in L_2-minimax estimation of probability density functions. *Mathematical Methods of Statistics* **5**, 253–274.

Shen, X. (1997). On methods of sieves and penalization. *The Annals of Statistics* **25**, 2555–2591.

Stefanski, L. and Carroll, R. J. (1990). Deconvoluting kernel density estimators.

Statistics **21**, 169–184.

Stepanova, N. (2013). On estimation of analytic density functions in L_p. *Mathematical Methods of Statistics* **22**, 114–136.

Tenbusch, A. (1994). Two-dimensional Bernstein polynomial density estimators. *Metrika* **41**, 233–253.

Vitale, R. A. (1975). Bernstein polynomial approach to density function estimation. In *Statistical Inference and Related Topics (Proc. Summer Res. Inst. Statist. Inference for Stochastic Processes, Indiana Univ., Bloomington, Ind., 1974, Vol. 2; dedicated to Z. W. Birnbaum)*, 87–99. Academic Press, New York.

Wang, X.-F. and Wang, B. (2011). Deconvolution estimation in measurement error models: The R package decon. *Journal of Statistical Software* **39**, 1–24.

Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics* **23**, 339–362.

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

Zhang, C.-H. (1990). Fourier methods for estimating mixing densities and distributions. *The Annals of Statistics* **18**, 806–831.

Department of Mathematical Sciences, Indiana University South Bend, South Bend, Indiana 46634, USA

E-mail: zguan@iusb.edu