Yicheng Kang and Peihua Qiu

Nonparametric Deconvolution by Fourier Transformation and Other Related Approaches

1

Nonparametric Deconvolution by Fourier Transformation and Other Related Approaches

CONTENTS

Nonparametric deconvolution in density estimation and its companion in regression (i.e., nonparametric regression with measurement errors) have broad applications. Many nonparametric deconvolution methods in the literature are based on kernel estimation. There are also some nonparametric deconvolution methods constructed based on the Fourier transformation, splines, wavelet, and other function expansions in specific basis function spaces. In this paper, some representative methods in the latter type are introduced. Some recent methods in the related image deblurring area are also described.

1.1 Introduction

Assume that we are interested in estimating the distribution of a random variable X, but it cannot be observed directly. What is observed is its contaminated version X^* defined by

$$X^* = X + U, \tag{1.1}$$

where U is the measurement error, and X and U are assumed to be independent with probability density functions f_X and f_U, respectively. From (1.1),

it is obvious that

$$f_{X^*}(x^*) = \int f_U(x^* - x) f_X(x)\, dx, \tag{1.2}$$

where f_{X^*} denotes the density of X^*. Namely, the density of X^* is a convolution of the densities of X and U. In the literature, it is often assumed that f_U is known and we will make that assumption throughout this chapter; for the case where f_U is unknown, see Chapter **??** (Delaigle and Van Keilegom, 2021). Then, our major goal is to estimate f_X from a set of observations of X^* through (1.2), which is a *deconvolution* problem.

Let $\{X_1^*, X_2^*, \ldots, X_n^*\}$ be a set of independent and identically distributed (i.i.d.) observations of X^*. To obtain a nonparametric estimator of f_X, denoted as $\hat{f}_X$, we wish to "invert" (1.2), since f_{X^*} can be estimated from the observations of X^*. Thus, the deconvolution problem is an *inverse* problem. More specifically, let $\mathcal{F}\{f\}(\omega)$ or $\tilde{f}(\omega)$ denote the Fourier transformation $\int_{-\infty}^{\infty} \exp(i\omega x) f(x) dx$ of a function $f(x)$, and $\mathcal{F}^{-1}\{\tilde{f}\}(x)$ denote the inverse Fourier transformation of $\tilde{f}(\omega)$. Then, by the property of the Fourier transformation that the Fourier transformation of the convolution of two functions equals the product of the Fourier transformations of the two individual functions, we have

$$\mathcal{F}\{f_{X^*}\}(\omega) = \mathcal{F}\{f_U\}(\omega)\mathcal{F}\{f_X\}(\omega).$$

Therefore, a reasonable estimator of $\mathcal{F}\{f_X\}(\omega)$, denoted as $\widehat{\mathcal{F}}\{f_X\}(\omega)$, can be defined as

$$\widehat{\mathcal{F}}\{f_X\}(\omega) = \frac{\widehat{\mathcal{F}}\{f_{X^*}\}(\omega)}{\mathcal{F}\{f_U\}(\omega)},$$

where $\widehat{\mathcal{F}}\{f_{X^*}\}(\omega)$ is an estimator of $\mathcal{F}\{f_{X^*}\}(\omega)$ that can be obtained from the observations of X^* and $\mathcal{F}\{f_U\}(\omega)$ can be computed from the known function f_U. In the above expression, $\mathcal{F}\{f_U\}(\omega)$ generally approaches zero rapidly when $|\omega|$ increases, but for $|\omega|$ large, $\widehat{\mathcal{F}}\{f_{X^*}\}(\omega)$ is a poor estimator that oscillates. Therefore, the ratio $\widehat{\mathcal{F}}\{f_{X^*}\}(\omega)/\mathcal{F}\{f_U\}(\omega)$ is numerically unstable, which makes the deconvolution problem challenging to solve. Therefore, the deconvolution problem is a so-called "ill-posed" problem in the literature (Tikhonov and Arsenin, 1977).

The deconvolution problem has a vigorous history, including considerable work on kernel-based methods that involve estimating f_{X^*} by a kernel estimator and then solving equation (1.2) using a Fourier transformation. See for example the contribution of Carroll and Hall (1988), Devroye (1989), Diggle and Hall (1993), Efromovich (1997), Fan (1991a), Fan (1991b), Liu and Taylor (1989), Masry (1991), Masry (1993), Stefanski (1990), Stefanski and Carroll (1990), Taylor and Zhang (1990), and Zhang (1990). In addition to the kernel-based methods, several alternative approaches using series expansions have been suggested over the years. For instance, a Fourier series (cosine-sine) method was proposed by Hall and Qiu (2005) and Delaigle et al. (2006).

Spline-based methods were discussed by Mendelsohn and Rice (1982) and Koo and Park (1996). Wavelet methods were addressed by Pensky and Vidakovic (1999), Walter (1999), Fan and Koo (2002) and Pensky (2002). Taylor series low order approximation methods were explored by Carroll and Hall (2004).

Kernel-based methods for density and regression estimation are introduced in Chapters **??** (Delaigle, 2021) and **??** (Apanasovich and Liang, 2021). This chapter aims to introduce some of the alternative methods. Our introduction will focus on the main ideas of some fundamental methodologies, their major strengths and limitations, their potential applications in other areas (e.g., image analysis), and certain important open problems for future research. The remaining part of the chapter is organized as follows. Methods based on the Fourier series expansion are discussed in Section 1.2. Spline-based methods are described in Section 1.3. Wavelet methods are introduced in Section 1.4. Methods using low order approximations are discussed in Section 1.5. Then, in Section 1.6, we discuss the image deblurring problem which is essentially a deconvolution problem in image analysis. Some concluding remarks are given in Section 1.7.

1.2 Deconvolution by the Fourier Series Expansion

The deconvolution method suggested by Hall and Qiu (2005) is based on the Fourier series expansion. Its major idea is to use the property of trigonometric-series expansions that the effect of the random error U in (1.1) can be factorized out and becomes separate from the effect of X. More specifically, assume that the support of the distribution of X is contained in a known compact interval $\mathcal{I}$. This requires only knowledge of an upper bound for the support of f_X, although methodologies have been developed for estimating the actual support (e.g., Delaigle and Gijbels 2006) when $\mathcal{I}$ is unknown. Without loss of generality, let $\mathcal{I} = [0, \pi]$. Then, we consider the cosine-series expansion of f_X on $\mathcal{I}$ (the reason for choosing the cosine series, rather than the full cosine and sine series, will be explained later):

$$f_X(x) = a + \sum_{j \geq 1} a_{1j} \cos(jx),$$

where $a = \pi^{-1}$, $a_{kj} = (2/\pi) \int_{\mathcal{I}} f_X(x) cs_{kj}(x)\, dx$, and

$$cs_{kj}(x) = \begin{cases} \cos(jx), & k = 1, \\ \sin(jx), & k = 2. \end{cases}$$

The case where $k = 2$ is considered above because the related quantities will be used below. Then, f_X can be estimated by $\overline{f}_X(x) = \pi^{-1} + \sum_{j \geq 1} \overline{a}_{1j} \cos(jx)$,

where $\overline{a}_{kj}$ denotes an estimator of a_{kj}, for $k = 1, 2$ and $j \geq 1$. To obtain $\overline{a}_{kj}$, recall that $\alpha_{kj} = E[cs_{kj}(U)]$ is known for all j and k. By the basic properties of trigonometric functions, we have

$$b_{kj} \equiv \frac{2}{\pi} E[cs_{kj}(X^*)] = \frac{2}{\pi} E[cs_{kj}(X+U)] = \begin{cases} a_{1j}\alpha_{1j} - a_{2j}\alpha_{2j}, & \text{if } k = 1, \\ a_{2j}\alpha_{1j} + a_{1j}\alpha_{2j}, & \text{if } k = 2. \end{cases} \tag{1.3}$$

Therefore, $b_j = (b_{1j}, b_{2j})^T$ can be expressed as $b_j = M_j a_j$, where $a_j = (a_{1j}, a_{2j})^T$ and M_j is the 2×2 matrix with $(M_j)_{11} = (M_j)_{22} = \alpha_{1j}$ and $(M_j)_{21} = -(M_j)_{12} = \alpha_{2j}$. If

$$|\alpha_{1j}| + |\alpha_{2j}| \neq 0 \quad (j \geq 1), \tag{1.4}$$

then a_j can be expressed as

$$\begin{pmatrix} a_{1j} \\ a_{2j} \end{pmatrix} = \frac{1}{\alpha_{1j}^2 + \alpha_{2j}^2} \begin{pmatrix} \alpha_{1j} & \alpha_{2j} \\ -\alpha_{2j} & \alpha_{1j} \end{pmatrix} \begin{pmatrix} b_{1j} \\ b_{2j} \end{pmatrix}. \tag{1.5}$$

So, a_{1j} and a_{2j} can be estimated by replacing b_{1j} and b_{2j} by their estimators on the right-hand side of (1.5). Now, $\hat{b}_{kj} = (2/\pi)n^{-1}\sum_i cs_{kj}(X_i^*)$ are the moment estimators of b_{kj}, for all k and j, which are unbiased. Therefore, a_{kj} can be estimated by

$$\begin{pmatrix} \hat{a}_{1j} \\ \hat{a}_{2j} \end{pmatrix} = \frac{1}{\alpha_{1j}^2 + \alpha_{2j}^2} \begin{pmatrix} \alpha_{1j} & \alpha_{2j} \\ -\alpha_{2j} & \alpha_{1j} \end{pmatrix} \begin{pmatrix} \hat{b}_{1j} \\ \hat{b}_{2j} \end{pmatrix}. \tag{1.6}$$

Thus, the following estimator of a_{1j} is obtained:

$$\hat{a}_{1j} = \frac{\alpha_{1j}\hat{b}_{1j} + \alpha_{2j}\hat{b}_{2j}}{\alpha_{1j}^2 + \alpha_{2j}^2}. \tag{1.7}$$

In the literature, it is often assumed that the distribution of U is known and symmetric about its mean. In such cases, without loss of generality, we can further assume that the mean of U equals zero. Then, $\alpha_{2j} = 0$ and the assumption (1.4) reduces to $\alpha_{1j} \neq 0$, for each j. Then, (1.7) becomes $\hat{a}_{1j} = \alpha_{1j}^{-1}\hat{b}_{1j}$, and the corresponding estimator of $f_X(x)$ can be defined as

$$\overline{f}_X(x) = \pi^{-1} + \sum_{j=1}^{\infty} \alpha_{1j}\overline{b}_{1j}\cos(jx), \tag{1.8}$$

where $\overline{b}_{1j} = \hat{b}_{1j}$ for $j \leq m$ and 0 otherwise, and m is a parameter. Here, m works as a smoothing parameter and it can be chosen by a smoothing parameter selection procedure, such as the cross-validation procedure.

In (1.8), the cosine-series is used, instead of the sine-series or the full cosine/sine-series, because the corresponding estimator of $f_X(x)$ is robust

against edge effects on $[0, \pi]$, as explained below. If $f_1(x)$ and $f_2(x)$ have a smooth derivative on $[0, \pi]$ and $[-\pi, \pi]$, respectively, then (by integration by parts)

$$\begin{aligned}\int_0^\pi f_1(x)\cos(jx)\,dx &= -\frac{1}{j}\int_0^\pi f_1'(x)\sin(jx)\,dx \\ &= \frac{1}{j^2}\left[(-1)^j f_1'(\pi) - f_1'(0)\right] - \frac{1}{j^2}\int_0^\pi f_1''(x)\cos(jx)\,dx \\ &= j^{-2}\left[(-1)^j f_1'(\pi) - f_1'(0)\right] + o\left(j^{-2}\right), \end{aligned} \tag{1.9}$$

Similarly, we have

$$\int_{-\pi}^\pi f_2(x)\sin(jx)\,dx = j^{-1}(-1)^{j+1}\left[f_2(\pi) - f_2(-\pi)\right] + o\left(j^{-1}\right). \tag{1.10}$$

Formula (1.10) implies that, unless $f_X(x)$ satisfies the periodic-continuity condition $f_X(\pi) = f_X(-\pi)$, coefficients in the full cosine/sine-series on $[-\pi, \pi]$ converge only at rate of j^{-1}. On the other hand, (1.9) implies that, without such a condition, the cosine-series on $[0, \pi]$ converges at rate of j^{-2}. This faster rate entails a smaller order of bias, and hence a faster rate of convergence in mean squares, of the estimator $\overline{f}_X$ in (1.10).

A counterpart of the deconvolution problem in regression is the errors-in-variables problem described below. Let Y be a response variable, and X be a predictor. We are interested in the functional relationship between X and Y. However, X cannot be observed directly. Instead, we can observe X^*. Then, a statistical model for describing this errors-in-variables problem can be the following one:

$$\begin{cases} Y = g(X) + V, \\ X^* = X + U, \end{cases} \tag{1.11}$$

where U and V are random errors, and X, U and V are independent of each other. Let (X_1^*, Y_1), (X_2^*, Y_2), ..., (X_n^*, Y_n) be i.i.d. observations of (X^*, Y). Then, the major goal of the errors-in-variables problem is to estimate the regression function g from the observations $\{(X_i^*, Y_i), i = 1, 2, \ldots, n\}$. To this end, the method based on the cosine-series expansion discussed above can be modified for solving the errors-in-variables problem, as described below. Let $\psi(x) = g(x)f_X(x)$. We again assume that the support of f_X is included in $[0, \pi]$. Then the support of ψ is included in $[0, \pi]$ as well. In the case where we do not have knowledge of an upper bound for the support of f_X, we can estimate the support using the methodology proposed in Delaigle and Gijbels (2006).

Next, the cosine-series expansion of ψ is

$$\psi(x) = \beta_0 + \sum_{j\geq 1}\beta_j\cos(jx), \qquad \text{for } x \in [0, \pi],$$

where $\beta_0 = \pi^{-1}$, and $\beta_j = (2/\pi)\int_0^\pi \psi(x)\cos(jx)\ dx$, for all j. Note that β_j can be written as

$$\beta_j = \frac{2}{\pi} E\left[\cos(jX)g(X)\right].$$

By the elementary properties of the trigonometric functions, we have

$$\begin{aligned}\cos(jX^*)g(X) &= \cos(jX + jU)g(X) \\ &= \cos(jX)g(X)\cos(jU) - \sin(jX)g(X)\sin(jU). \qquad (1.12)\end{aligned}$$

By the assumptions that the distribution of U is symmetric about zero and $\alpha_{1j} \neq 0$, we have

$$E\left[\cos(jX)g(X)\right] = \alpha_{1j}^{-1} E\left[\cos(jX^*)g(X)\right].$$

Now, $E\left[\cos(jX^*)g(X)\right] = E\left[\cos(jX^*)Y\right]$ and $n^{-1}\sum_i \cos(jX_i^*)Y_i$ is an unbiased estimator of $E\left[\cos(jX^*)Y\right]$. So, $\psi(x)$ can be estimated by

$$\hat{\psi}(x) = \beta_0 + \sum_{j\geq 1} \hat{\beta}_j \cos(jx),$$

where $\hat{\beta}_j = (2/\pi)\alpha_{1j}^{-1} n^{-1}\sum_i \cos(jX_i^*)Y_i$ for $j \leq m$, and 0 otherwise. As below (1.8), m is a smoothing parameter in the above expression. Next, one can construct the estimator of $g(x)$ by $\hat{g}(x) = \hat{\psi}(x)/\hat{f}_X(x)$, where $\hat{f}_X(x)$ can be $\overline{f}_X(x)$ defined in (1.8). A similar estimator of $g(x)$ in the special case when X is uniformly distributed on an interval and U has a normal distribution was considered in Efromovich (1994) and Efromovich (1999).

One advantage to use the Fourier series expansions as the basis for inference in the deconvolution problems is that the effect of the random errors can be factorized out in a way that is easy to handle empirically (i.e., (1.3) and (1.12)). This property is from the elementary addition formulae for sine and cosine functions, and it is not readily available when one is using methods based on the continuous Fourier transformation. It allows to construct relatively simple estimators, which are based on additions of finite series, rather than integrations. These methods are particularly effective when edge effects are involved, and they are easy to code too. However, there is room for improvement. For instance, the estimator defined by (1.8) can produce density estimates that take negative values when the true density is close to zero at certain places. Thus, when the method is used to estimate smooth densities it will suffer from the same vulnerability to negativity as kernel-based methods.

In comparison with kernel-based methods theoretically, the convergence rate achieved by the estimator (1.8) is governed by the rate of convergence of the cosine-series expansion, which is not a natural property for kernel estimators. In other words, it is possible to construct densities where the Fourier series approach gives much faster convergence rates than the kernel methods.

We refer readers to the numerical examples provided in Section 3 in Hall and Qiu (2005) for more details.

Delaigle et al. (2006) modified the Fourier series approach described above for solving the so-called Berkson errors-in-variables problem which is the same as the problem described by (1.11), except that the positions of X and X^* are switched and X^*, U and V are assumed independent in this case. Interested readers can read the paper Delaigle et al. (2006) for more details.

1.3 Deconvolution by Splines

Deconvolution by B-splines and least squares estimation. Motivated by the analysis of DNA-content data obtained by microfluorimetry, Mendelsohn and Rice (1982) proposed to approximate f_X by a function in a finite dimensional family of densities $\mathscr{L}_p$ of dimension p, where p is a tuning parameter. Because of computational convenience of the B-splines, $\mathscr{L}_p$ can be chosen to be the space spanned by the B-spline basis functions with fixed knot locations. Then, the estimate of f_X, $\hat{f}_X$, can be defined as a linear combination of the B-spline basis functions

$$\hat{f}_X(x) = \sum_{j=1}^{p} \hat{\gamma}_j B_j(x), \tag{1.13}$$

where $B_j(x)$ is the jth B-spline basis function of degree $k-1$ with knots t_1, t_2, …, t_{p+k}, the coefficients $\{\hat{\gamma}_j\}$ are the solution of the following least squares problem:

$$\min_{\gamma_j, 1\le j\le p} \sum_{i=1}^{m} \left(\hat{f}_{X^*}(s_i) - \sum_{j=1}^{p} \gamma_j f_U * B_j(s_i) \right)^2, \tag{1.14}$$

$\{\hat{f}_{X^*}(s_i), i = 1, 2, \ldots, m\}$ is a histogram estimate of f_{X^*}, and $f_U * B_j$ denotes the convolution product and can be numerically evaluated using the Simpson's rule.

One advantage to use the estimator (1.13) is that it is convenient to make $\hat{f}_X$ a legitimate probability density (i.e., a nonnegative function with a unit integration) as follows. By the properties of the B-splines that $\int B_j(x)\, dx = (t_{j+k} - t_k)/k$ and $B_j(x) \ge 0$, we can solve the least squares problem (1.14) by restricting the coefficients $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$ to be $\boldsymbol{\gamma} \ge 0$ and $\mathbf{c}^T\boldsymbol{\gamma} = 1$, where $\mathbf{c} = (c_1, \ldots, c_p)^T$ and $c_j = (t_{j+k} - t_k)/k$, for $j = 1, 2, \ldots, p$. Then, it follows immediately that $\hat{f}_X$ is a probability density. On the other hand, there

are several potential limitations with the above procedure. First, Mendelsohn and Rice (1982) suggested that the knots were chosen equally spaced and p was determined by increasing its value until the occurrence of high frequency oscillations. This is similar to the eyeball method when choosing a bandwidth parameter for kernel smoothing methods. Thus, it is subjective and could be labor intensive as well. Second, in the application to the DNA-content data in Mendelsohn and Rice (1982), the data were used in the form of histograms, which might be less efficient compared to the likelihood-based approach to estimate the coefficients in (1.14). This latter approach was studied in Koo and Park (1996) and will be discussed in more details in the next part.

Logspline deconvolution with the EM algorithm. Koo and Park (1996) considered the deconvolution problem in the following complete-incomplete data specification. The complete unobserved data were (X_1, X_1^*), (X_2, X_2^*), ..., (X_n, X_n^*), which was a random sample of size n from the joint distribution of (X, X^*) whose probability density was given by $f_{(X,X^*)}(x, x^*) = f_U(x^* - x)f_X(x)$. The incomplete observed data were X_1^*, X_2^*, ..., X_n^*. They considered how to estimate f_X from the observed data alone. Let L and R be numbers such that $-\infty \le L \le R \le \infty$, $L < t_1 < t_2 < \cdots < t_K < R$ be K knots. Define S to be the space of all twice continuously differentiable functions s on (L, R) such that s is a cubic polynomial on each of the intervals $[t_1, t_2]$, $[t_2, t_3]$, ..., $[t_{K-1}, t_K]$, and is a linear function on each of the intervals $(L, t_1])$ and $[t_K, R)$. Then, all functions in S are called natural cubic splines, and S has a basis B_1, B_2, ..., B_K that can be generated from the conventional B-splines (Stone and Koo, 1986). Now, we assume that the support of f_X is $[L, R]$ and consider the following logspline model:

$$f_{X|\boldsymbol{\theta}}(x) = \exp(s(x; \boldsymbol{\theta}) - C(\boldsymbol{\theta})), \quad L < x < R,$$

where

$$s(x; \boldsymbol{\theta}) = \sum_{j=1}^{K-1} \theta_j B_j(x) \quad \text{and} \quad C(\boldsymbol{\theta}) = \log \left(\int_L^R \exp(s(x; \boldsymbol{\theta}))\, dx \right) < \infty.$$

Then $f_{X|\boldsymbol{\theta}}(\cdot)$ is a positive density function on (L, R). Clearly, the logspline model assumes that f_X is twice continuously differentiable. Also, by the property of B-splines that $\sum_{j=1}^{K} B_j(x) = 1$ for all $x \in (L, R)$, the last basis function will not be used in the above expression to make the logspline model identifiable (Stone and Koo, 1986). Now, the log likelihood function of the incomplete observed data X_1^*, X_2^*, ..., X_n^* is

$$\begin{aligned} l_{X^*}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log \left(\int_L^R f_U(X_i^* - x) f_{X|\boldsymbol{\theta}}(x)\, dx \right) \\ &= \sum_{i=1}^{n} \log \left(\int_L^R f_U(X_i^* - x) \exp\left(s(x; \boldsymbol{\theta}) - C(\boldsymbol{\theta})\right)\, dx \right) \end{aligned}$$

$$= \sum_{i=1}^{n} \log \left(\int_L^R f_U(X_i^* - x) \exp(s(x;\boldsymbol{\theta}))\, dx \right) - nC(\boldsymbol{\theta}).$$

The usual maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ is the maximizer of $l_{X^*}(\boldsymbol{\theta})$. Direct maximization of $l_{X^*}(\boldsymbol{\theta})$ is difficult numerically, because of the integrals inside the logarithms. The EM algorithm, which is briefly described below, provides a simpler approach. Let $f_{X|X^*,\boldsymbol{\theta}}(x|x^*)$ denote the conditional density of X given X^*. We have

$$f_{X|X^*,\boldsymbol{\theta}}(x|x^*) = \frac{f_{(X,X^*)|\boldsymbol{\theta}}(x,x^*)}{f_{X^*|\boldsymbol{\theta}}(x^*)}.$$

In terms of log-likelihoods, we have

$$l_{X^*}(\boldsymbol{\theta}) = l_{(X,X^*)}(\boldsymbol{\theta}) - l_{X|X^*}(\boldsymbol{\theta}).$$

Taking conditional expectations with respect to the distribution of $X|X^*$ governed by the parameter $\boldsymbol{\theta_0}$ gives

$$\begin{aligned} l_{X^*}(\boldsymbol{\theta}) &= \mathrm{E}\left[l_{(X,X^*)}(\boldsymbol{\theta})|X^*,\boldsymbol{\theta_0}\right] - \mathrm{E}\left[l_{X|X^*}(\boldsymbol{\theta})|X^*,\boldsymbol{\theta_0}\right] \\ &\equiv G(\boldsymbol{\theta}|\boldsymbol{\theta_0}) - R(\boldsymbol{\theta}|\boldsymbol{\theta_0}). \end{aligned}$$

In each iteration of the EM algorithm, we first compute $G(\boldsymbol{\theta}|\boldsymbol{\theta_0})$ (i.e., the expectation step) and then maximize $G(\boldsymbol{\theta}|\boldsymbol{\theta_0})$ over $\boldsymbol{\theta}$ (i.e., the maximization step). Thus, a key quantity in the EM algorithm is the following conditional expectation:

$$\begin{aligned} G(\boldsymbol{\theta}|\boldsymbol{\theta_0}) =& E_{X|\boldsymbol{\theta_0},X^*}\left[\log\left(f_{(X,X^*)|\boldsymbol{\theta}}(X_i,X_i^*)\right)\right] \\ =& \int_L^R f_{X|X^*,\boldsymbol{\theta_0}}(x|X_i^*) \log\left(f_U(X_i^*-x) f_{X|\boldsymbol{\theta}}(x)\right)\, dx \\ =& \int_L^R f_{X|X^*,\boldsymbol{\theta_0}}(x|X_i^*) s(x;\boldsymbol{\theta})\, dx - C(\boldsymbol{\theta}) + \text{terms not involving } \boldsymbol{\theta} \\ =& \sum_{j=1}^{K-1} \theta_j \int_L^R B_j(x) f_{X|X^*,\boldsymbol{\theta_0}}(x|X_i^*)\, dx - C(\boldsymbol{\theta}) + \text{terms not involving } \boldsymbol{\theta}, \end{aligned}$$

where

$$\begin{aligned} f_{X|X^*,\boldsymbol{\theta}}(x|x^*) &= f_U(x^*-x)\exp(s(x;\boldsymbol{\theta}) - C(\boldsymbol{\theta}|x^*)), \\ C(\boldsymbol{\theta}|x^*) &= \log\left(\int_L^R f_U(x^*-x)\exp(s(x;\boldsymbol{\theta}))\, dx\right). \end{aligned}$$

In the maximization step, if $\boldsymbol{\theta_1}$ maximizes $G(\boldsymbol{\theta}|\boldsymbol{\theta_0})$, then we have $l_{X^*}(\boldsymbol{\theta_1}) \geq l_{X^*}(\boldsymbol{\theta_0})$. To see this, note that $R(\boldsymbol{\theta}|\boldsymbol{\theta_0})$ is the expectation of a log-likelihood

of a density (indexed by $\boldsymbol{\theta}$), with respect to the same density indexed by $\boldsymbol{\theta_0}$. Hence, by Jensen's inequality, $R(\boldsymbol{\theta}|\boldsymbol{\theta_0}) \leq R(\boldsymbol{\theta_0}|\boldsymbol{\theta_0})$ and it follows that

$$l_{X^*}(\boldsymbol{\theta_1}) - l_{X^*}(\boldsymbol{\theta_0}) = [G(\boldsymbol{\theta_1}|\boldsymbol{\theta_0}) - G(\boldsymbol{\theta_0}|\boldsymbol{\theta_0})] - [R(\boldsymbol{\theta_1}|\boldsymbol{\theta_0}) - R(\boldsymbol{\theta_0}|\boldsymbol{\theta_0})] \geq 0.$$

We now state the EM algorithm for computing $\hat{\boldsymbol{\theta}}$ in the complete-incomplete data specification. It starts with an initial estimate $\hat{\boldsymbol{\theta}}^0$ and iteratively updates the estimate as follows.

Expectation-Step: Given the current estimate $\hat{\boldsymbol{\theta}}^{(k)}$ of $\boldsymbol{\theta}$, calculate

$$b_j\left(\hat{\boldsymbol{\theta}}^{(k)}\right) = \sum_{i=1}^{n} \int_L^R B_j(x) f_{X|X^*,\hat{\boldsymbol{\theta}}^{(k)}}(x|X_i^*)\ dx, \quad 1 \leq j \leq K-1.$$

Maximization-Step: Determine the updated estimate $\hat{\boldsymbol{\theta}}^{(k+1)}$ by maximizing

$$Q\left(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k)}\right) = \sum_{j=1}^{K-1} \theta_j b_j\left(\hat{\boldsymbol{\theta}}^{(k)}\right) - nC(\boldsymbol{\theta}).$$

The EM algorithm stops when $l_{X^*}(\hat{\boldsymbol{\theta}}^{(k+1)}) - l_{X^*}(\hat{\boldsymbol{\theta}}^{(k)}) < 10^{-6}$. In the Maximization-step, the Newton-Raphson algorithm with step-halving can be employed for the maximization. More specifically, let $\mathbf{S}(\boldsymbol{\theta})$ be the score function at $\boldsymbol{\theta}$ with elements $b_j(\hat{\boldsymbol{\theta}}^{(k)}) - \partial C(\boldsymbol{\theta})/\partial\theta_j$ and let $\mathbf{H}(\boldsymbol{\theta})$ be the Hessian matrix of $C(\boldsymbol{\theta})$ at $\boldsymbol{\theta}$ with elements

$$\int_L^R B_{j_1}(x) B_{j_2}(x) f_{\boldsymbol{X}|\boldsymbol{\theta}}(x)\ dx - \int_L^R B_{j_1}(x) f_{\boldsymbol{X}|\boldsymbol{\theta}}(x)\ dx \int_L^R B_{j_2} f_{\boldsymbol{X}|\boldsymbol{\theta}}(x)\ dx.$$

Now, the computation of $\hat{\boldsymbol{\theta}}^{(k+1)}$ starts with $\tilde{\boldsymbol{\theta}}^{(0)} = \hat{\boldsymbol{\theta}}^{(k)}$ and iteratively determines $\tilde{\boldsymbol{\theta}}^{(m+1)}$ according to the expression

$$\tilde{\boldsymbol{\theta}}^{(m+1)} = \tilde{\boldsymbol{\theta}}^{(m)} + 2^{-q}\left[\mathbf{H}\left(\tilde{\boldsymbol{\theta}}^{(m)}\right)\right]^{-1} \mathbf{S}\left(\tilde{\boldsymbol{\theta}}^{(m)}\right),$$

where q is the smallest nonnegative integer such that

$$l_{X^*}\left(\tilde{\boldsymbol{\theta}}^{(m)} + 2^{-q}\left[\mathbf{H}\left(\tilde{\boldsymbol{\theta}}^{(m)}\right)\right]^{-1} \mathbf{S}\left(\tilde{\boldsymbol{\theta}}^{(m)}\right)\right) \geq l_{X^*}\left(\tilde{\boldsymbol{\theta}}^{(m)}\right).$$

The Newton-Raphson algorithm stops when $l_{X^*}(\tilde{\boldsymbol{\theta}}^{(m+1)}) - l_{X^*}(\tilde{\boldsymbol{\theta}}^{(m)}) < 10^{-6}$.

To make the entire estimation procedure fully automatic, a data-driven rule for determining the final number and locations of knots is needed. Choosing the number of knots is similar to choosing a bandwidth in kernel smoothing. Too many knots would lead to a noisy estimate, and too few knots would give an estimate that is overly smoothed. The knot locations are also important, since more knots are needed in a region where the curvature of f_X is

larger. To this end, Koo and Park (1996) suggested the following stepwise knot deletion strategy. Let K^* be an initial number of knots, and the knot locations are determined by an initial knot placement rule. Koo and Park (1996) suggested that K^* can be chosen to be the integer closest to $3(\log n)^{1/2}$, and the knots can initially be placed at equally spaced percentiles of X_i^* once the initial number of knots is given. Then, at each stage of the stepwise knot deletion process, the EM algorithm is used to obtain the estimate $\hat{\boldsymbol{\theta}}_K$ with K being the number of remaining knots at that stage. The Bayesian information criterion (BIC) is calculated as:

$$\text{BIC} = -2l_{X^*}\left(\hat{\boldsymbol{\theta}}_K\right) + (K-1)\log n.$$

Delete the knote that results in the biggest drop in BIC and repeat this deletion process until BIC stops to decrease.

The logspline deconvolution method has advantages similar to those of logspline density estimates in the usual density estimation problem (cf., Kooperberg and Stone, 1991). It gives density estimates that are positive with unit integrations. However, that method also has some limitations. First, it is known that under fairly general conditions the EM algorithm could converge to a local maximum of $l_{X^*}(\boldsymbol{\theta})$. When this function is not concave, there is no guarantee that such a local maximum has unique maximizer, or that it is a global maximum. Second, the numerical studies in Koo and Park (1996) showed that the logspline deconvolution estimates may require a large sample size to perform well when the true density function has a complex (e.g., bimodal) structure. Third, the above procedure for determining K^* is ad hoc in nature.

1.4 Deconvolution by Wavelets

Wavelet methods are well received in the literature of density estimation (e.g., Walter 1981, Penskaya 1985, Kerkyacharian and Picard 1992, Masry 1994, Hall and Patil 1995) and curve estimation (e.g., Antoniadis et al. 1994, Abramovich and Silverman 1998, Donoho and Johnstone 1995, Hall et al. 1998, Hall et al. 1997, Walter 1994). Pensky and Vidakovic (1999) proposed several wavelet estimators of the density function f_X in the deconvolution problem (1.1). The underlying idea is to represent f_X via a wavelet expansion and estimate the coefficients using a deconvolution algorithm. Their estimators are based on the Meyer-type wavelets, which are band-limited and their Fourier transforms would have bounded supports (Hernández and Weiss, 1996). These wavelet estimators are constructed as follows.

Assume that f_X is square integrable and the Fourier transform of f_{X^*},

denoted as $\tilde{f}_{X^*}(\omega)$, does not vanish for real ω. Let $\varphi(x)$ and $\psi(x)$ be a scaling function and a wavelet for an orthonormal multi-resolution decomposition of $L^2(-\infty,\infty)$, respectively. Then, for any integer m, f_X has the following expansion:

$$f_X(x) = \sum_{k\in\mathbb{Z}} a_{m,k}\varphi_{m,k}(x) + \sum_{k\in\mathbb{Z}}\sum_{j=m}^{\infty} b_{j,k}\psi_{j,k}(x), \tag{1.15}$$

where $\varphi_{m,k}(x) = 2^{m/2}\varphi(2^m x - k)$ and $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$, and the coefficients $a_{m,k}$ and $b_{j,k}$ have the forms

$$a_{m,k} = \int_{-\infty}^{\infty} \varphi_{m,k}(x) f_X(x)\, dx, \quad b_{j,k} = \int_{-\infty}^{\infty} \psi_{j,k}(x) f_X(x)\, dx. \tag{1.16}$$

The scaling function $\varphi(x)$ and the wavelet function $\psi(x)$ can be defined as the functions whose Fourier transforms are (Walter 1994; Zayed and Walter 1996)

$$\tilde{\varphi}(\omega) = \left[\int_{\omega-\pi}^{\omega+\pi} dP\right]^{1/2}, \quad \tilde{\psi}(\omega) = \exp(-i\omega/2)\left[\int_{|\omega|/2-\pi}^{|\omega|-\pi} dP\right]^{1/2},$$

where P is a probability measure that satisfies the following conditions: (i) support of P is included in $[-\pi/3, \pi/3]$, and (ii) $\tilde{\varphi}(\omega)$ and $\tilde{\psi}(\omega)$ are $s \geq 2$ times continuously differentiable on $(-\infty,\infty)$. This ensures that $\varphi(x)$ and $\psi(x)$ have sufficient rates of descent as $|x| \to \infty$. Then, both $\tilde{\varphi}(\omega)$ and $\tilde{\psi}(\omega)$ would have bounded supports: supp $\tilde{\varphi} \subset [-4\pi/3, 4\pi/3]$ and supp $\tilde{\psi} \subset [-8\pi/3, -2\pi/3] \cup [2\pi/2, 8\pi/3]$. Moreover, we have

$$C_\varphi = \sup_x [|\varphi(x)|(|x|^s + 1)] < \infty, \quad C_\psi = \sup_x [|\psi(x)|(|x|^s + 1)] < \infty.$$

Let $u_{m,k}$ and $v_{j,k}$ be the solutions of the following equations:

$$\int_{-\infty}^{\infty} f_U(z-x)u_{m,k}(z)\, dz = \varphi_{m,k}(x), \quad \int_{-\infty}^{\infty} f_U(z-x)v_{j,k}(z)\, dz = \psi_{j,k}(x). \tag{1.17}$$

Then, the coefficients $a_{m,k}$ and $b_{j,k}$ can be viewed as mathematical expectations of the functions $u_{m,k}$ and $v_{j,k}$ (by (1.16)):

$$\begin{aligned} a_{m,k} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_U(z-x) f_X(x) u_{m,k}(z)\, dxdz \\ &= \int_{-\infty}^{\infty} f_{X^*}(z) u_{m,k}(z)\, dz = E\left[u_{m,k}(X^*)\right], \\ b_{j,k} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_U(z-x) f_X(x) v_{j,k}(z)\, dxdz \end{aligned} \tag{1.18}$$

$$= \int_{-\infty}^{\infty} f_{X^*}(z) v_{j,k}(z)\, dz. = E\left[v_{j,k}(X^*)\right]. \tag{1.19}$$

It is worth mentioning that (1.18) and (1.19) are much more convenient to work with, compared to (1.16), since the X_i^*'s are observable whereas the X_i's are not. Before the wavelet estimators of f_X can be formally defined, we still need to obtain the expressions for $u_{m,k}$ and $v_{j,k}$. To this end, after taking the Fourier transform on both sides of (1.17), we have $u_{m,k}(x) = 2^{m/2} U_m(2^m x - k)$, and $v_{j,k}(x) = 2^{j/2} V_j(2^j x - k)$, where $U_m(\cdot)$ and $V_j(\cdot)$ are the inverse Fourier transforms of the functions

$$\tilde{U}_m(\omega) = \frac{\tilde{\varphi}(\omega)}{\tilde{f}_U(-2^m\omega)}, \quad \tilde{V}_j(\omega) = \frac{\tilde{\psi}(\omega)}{\tilde{f}_U(-2^j\omega)},$$

respectively. Thus, we can estimate $a_{m,k}$ and $b_{j,k}$ by

$$\hat{a}_{m,k} = \frac{1}{n}\sum_{i=1}^{n} 2^{m/2} U_m(2^m X_i^* - k), \quad \hat{b}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} 2^{j/2} V_j(2^j X_i^* - k).$$

Then, from (1.15), we can define a linear wavelet estimator of f_X as

$$\hat{f}_{X,n}^{(L)}(x) = \sum_{k\in\mathbb{Z}} \hat{a}_{m,k}\varphi_{m,k}(x),$$

and a nonlinear wavelet estimator of $f_X(x)$ as

$$\hat{f}_{X,n}^{(N)} = \sum_{k\in\mathbb{Z}} \hat{a}_{m,k}\varphi_{m,k}(x) + \sum_{j=m}^{m+r}\left[\sum_{k\in\mathbb{Z}} \hat{b}_{j,k}\psi_{j,k}(x)\right] I\left(\sum_{k\in\mathbb{Z}} \hat{b}_{j,k}^2 > \delta_{j,n}^2\right),$$

where $\delta_{j,n}$ are thresholding parameters. Note that $\hat{f}_{X,n}^{(L)}(x)$ and $\hat{f}_{X,n}^{(N)}(x)$ both seem computationally intractable since their definitions involve the calculation of infinite series. Under some minor conditions, we can modify the infinite series estimators $\hat{f}_{X,n}^{(L)}(x)$ and $\hat{f}_{X,n}^{(N)}(x)$ to be finite series estimators as follows

$$\hat{f}_{X,n}^{(LF)}(x) = \sum_{|k|\le K_n} \hat{a}_{m,k}\varphi_{m,k}(x),$$

$$\hat{f}_{X,n}^{(NF)} = \sum_{|k|\le M_n} \hat{a}_{m,k}\varphi_{m,k}(x) + \sum_{j=m}^{m+r}\left[\sum_{|k|\le L_n} \hat{b}_{j,k}\psi_{j,k}(x)\right] I\left(\sum_{|k|\le L_n} \hat{b}_{j,k}^2 > \delta_{j,n}^2\right),$$

without any loss in the convergence rates. Here, K_n, M_n, and L_n are smoothing parameters.

Pensky and Vidakovic (1999) showed that $\hat{f}_{X,n}^{(L)}(x)$, $\hat{f}_{X,n}^{(LF)}(x)$, $\hat{f}_{X,n}^{(N)}(x)$,

and $\hat{f}_{X,n}^{(NF)}(x)$ all achieve the optimal rates of convergence established in Fan (1993). Furthermore, in the case when f_U is super smooth (i.e., $\tilde{f}_U(\omega)$ exponentially decreases as ω increases), the linear wavelet estimators $\hat{f}_{X,n}^{(L)}(x)$ and $\hat{f}_{X,n}^{(LF)}(x)$ are adaptive in the sense that the choice of m, which yields the optimal rate of convergence, does not depend on the unknown smoothness of the density f_X. Similarly, the nonlinear wavelet estimators $\hat{f}_{X,n}^{(N)}(x)$ and $\hat{f}_{X,n}^{(NF)}(x)$ are adaptive in the case when $\tilde{f}_U$ has a polynomial descent. Therefore, all these wavelet estimators have nice theoretical properties. However, the above construction of these estimators cannot guarantee that they are nonnegative with unit integrations. More recent work (e.g., Bigot and Van Bellegem 2009) has lifted this limitation by proposing new wavelet-based methods that guarantee the related estimators to be legitimate density functions.

1.5 Deconvolution by Low Order Approximations

The deconvolution methods discussed in the previous sections all assume that the distribution of U is known. This is a common assumption in the deconvolution literature because the related problem is difficult to solve otherwise. However, little information is generally available in practice about the distribution of U. Even under the assumption of a known f_U, Carroll and Hall (1988) have showed that the fastest possible convergence rate of a density estimator in the deconvolution problem is in the reciprocal of the logarithmic of the sample size when U has a normal distribution. Fan (1991b) discussed settings in which the optimal convergence rate could be in the reciprocal of a polynomial of the sample size. Even in such settings, the optimal convergence rate is often slow, unless f_U is very unsmooth (e.g., it has a discontinuity). The implication of these results is that consistent estimation of f_X is difficult in practical terms. An alternative approach would be to estimate a function that approximates f_X and is relatively easy to estimate. Carroll and Hall (2004) suggested an approach based on this idea. It requires knowledge of only low order moments of f_U. This information is often available either from a sample drawn from the distribution of U or from a small number of repeated observations of X^* for the same X. This approach is introduced below in two parts.

Density estimation by low order approximations. Conventional kernel estimators of f_{X^*} and f_X are given by

$$\hat{f}_{X^*}(x^*) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x^* - X_i^*}{h}\right),$$

$$\hat{f}_X(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where $K(\cdot)$ is a kernel function and h is a bandwidth. Because we do not observe the X_i's in practice, $\hat{f}_X$ cannot be computed directly from the observed data. Nevertheless we can try to find a good approximation to $\hat{f}_X$ or its mean. To this end, let $K^{(j)}$ be the jth derivative of K. Define

$$\begin{aligned}
\lambda_j(x) &= E\left[K^{(j)}\left(\frac{x - X}{h}\right)\right], \\
\kappa_j(x) &= E\left[K^{(j)}\left(\frac{x - X^*}{h}\right)\right], \\
u_j &= E\left(U^j\right).
\end{aligned}$$

Assume that U has finite moments and K is an analytic function in the sense that all its derivatives are well defined on the entire real line. Obviously, the analyticity condition is satisfied for the commonly used Gaussian kernel function. Now, by the Taylor expansion of $K^{(j)}(\cdot)$ at $(x - X)/h$, we have

$$\begin{aligned}
\lambda_j(x) &= \kappa_j(x) - \sum_{k_1 \geq 1} \frac{(-1)^{k_1} u_{k_1}}{k_1! h^{k_1}} \lambda_{j+k_1}(x) \\
&= \kappa_j(x) - \sum_{k_1 \geq 1} \frac{(-1)^{k_1} u_{k_1}}{k_1! h^{k_1}} \left[\kappa_{j+k_1}(x) - \sum_{k_2 \geq 1} \frac{(-1)^{k_2} u_{k_2}}{k_2! h^{k_2}} \lambda_{j+k_1+k_2}(x)\right] \\
&= \kappa_j(x) - \sum_{k_1 \geq 1} \frac{(-1)^{k_1} u_{k_1}}{k_1! h^{k_1}} \kappa_{j+k_1}(x) + \sum_{k_1 \geq 1} \sum_{k_2 \geq 1} \frac{(-1)^{k_1+k_2} u_{k_1} u_{k_2}}{k_1! k_2! h^{k_1+k_2}} \lambda_{j+k_1+k_2}(x) \\
&\vdots \\
&= \kappa_j(x) + \sum_{r=1}^{\infty} \sum_{k_1=1}^{\infty} \cdots \sum_{k_r=1}^{\infty} \frac{(-1)^{k_1+\ldots+k_r+r}}{k_1! \ldots k_r! h^{k_1+\ldots+k_r}} u_{k_1} \cdots u_{k_r} \kappa_{k_1+\ldots+k_r+j}(x).
\end{aligned} \tag{1.20}$$

From (1.20), we have

$$\begin{aligned}
E[\hat{f}_X^{(j)}(x)] &= E[\hat{f}_{X^*}^{(j)}(x)] \\
&\quad + \sum_{r=1}^{\infty} \sum_{k_1=1}^{\infty} \cdots \sum_{k_r=1}^{\infty} \frac{(-1)^{k_1+\ldots+k_r+r}}{k_1! \ldots k_r!} u_{k_1} \cdots u_{k_r} E[\hat{f}_{X^*}^{(k_1+\ldots+k_r+j)}(x)].
\end{aligned}$$

Therefore, if $\hat{u}_j$ is a consistent estimator of u_j and $\mathrm{Var}(U) \to 0$, then a reasonable estimator of the νth-order approximation to $E[\hat{f}_X^{(j)}(x)]$ is

$$\begin{aligned}
\hat{f}_{X,\nu}^{(j)}(x) &= \hat{f}_{X^*}^{(j)}(x) \\
&\quad + \sum_{r \geq 1, k_1 \geq 1, \cdots, k_r \geq 1 : k_1+\ldots+k_r \leq \nu} \cdots \sum \frac{(-1)^{k_1+\ldots+k_r+r}}{k_1! \ldots k_r!} \hat{u}_{k_1} \cdots \hat{u}_{k_r} \hat{f}_{X^*}^{(k_1+\ldots+k_r+j)}(x).
\end{aligned}$$

In particular, the estimators of the second-, fourth- and sixth-order approximations are given by

$$\hat{f}^{(j)}_{X,2}(x) = \hat{f}^{(j)}_{X^*}(x) + \hat{u}_1 \hat{f}^{(j+1)}_{Y}(x) + \frac{1}{2}(2\hat{u}_1^2 - \hat{u}_2)\hat{f}^{(j+2)}_{X^*}(x), \tag{1.21}$$

$$\begin{aligned}\hat{f}^{(j)}_{X,4}(x) = {} & \hat{f}^{(j)}_{X,2}(x) + \frac{1}{6}(6\hat{u}_1^3 - 6\hat{u}_1\hat{u}_2 + \hat{u}_3)\hat{f}^{(j+3)}_{X^*}(x) \\ & + \frac{1}{24}(24\hat{u_1}^4 - 36\hat{u}_1^2\hat{u}_2 + 8\hat{u}_1\hat{u}_3 + 6\hat{u}_2^2 - \hat{u}_4)\hat{f}^{(j+4)}_{X^*}(x),\end{aligned} \tag{1.22}$$

$$\begin{aligned}\hat{f}^{(j)}_{X,6}(x) = {} & \hat{f}^{(j)}_{X,4}(x) + \frac{1}{120}(120\hat{u}_1^5 - 240\hat{u}_1^3\hat{u}_2 + 60\hat{u}_1^2\hat{u}_3 \\ & + 180\hat{u}_1\hat{u}_2^2 + 10\hat{u}_1\hat{u}_4 - 20\hat{u}_2\hat{u}_3 + \hat{u}_5)\hat{f}^{(j+5)}_{X^*}(x) \\ & + \frac{1}{720}(720\hat{u}_1^6 - 1800\hat{u}_1^4\hat{u}_2 + 480\hat{u}_1^3\hat{u}_3 + 1080\hat{u}_1^2\hat{u}_2^2 - 90\hat{u}_1^2\hat{u}_4 \\ & + 12\hat{u}_1\hat{u}_5 - 360\hat{u}_1\hat{u}_2\hat{u}_3 - 90\hat{u}_2^3 + 20\hat{u}_3^2 + 30\hat{u}_2\hat{u}_4 - \hat{u}_6)\hat{f}^{(j+6)}_{X^*}(x).\end{aligned} \tag{1.23}$$

In the literature, it is common to assume that the distribution of U is symmetric. In such cases, we can let $\hat{u}_k = 0$, for odd k, and only use approximations of even orders. Then, the estimator of the 2νth order approximation to $E[\hat{f}^{(j)}_X(x)]$ can be simplified to

$$\begin{aligned}\hat{f}^{(j)}_{X,2\nu}(x) = {} & \hat{f}^{(j)}_{X^*}(x) \\ & + \sum_{r\geq 1, k_1 \geq} \sum_{1,\cdots,k_r \geq 1 : k_1+\ldots+k_r \leq \nu} \frac{(-1)^r}{(2k_1)!\ldots(2k_r)!}\hat{u}_{2k_1}\cdots\hat{u}_{2k_r}\hat{f}^{(2k_1+\ldots+2k_r+j)}_{X^*}(x),\end{aligned}$$

and the equations (1.21) – (1.23) would have simpler forms as well.

Errors-in-variables regression by low order approximations. It is worth mentioning that the above approximation procedure can also be applied to the errors-in-variables regression (cf., (1.11)). Let us consider the Nadaraya-Watson (NW) estimator of the regression function $g(x) = E(Y|X = x)$, defined by

$$\bar{g}_{NW}(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)},$$

where h is a bandwidth and K is a kernel function. Again, because the X_i's are unobservable, $\bar{g}_{NW}$ cannot be obtained from the observed data. Nonetheless we can obtain estimators of the low order approximations to its mean using results similar to (1.20). To this end, we have

$$\begin{aligned}& E\left[YK\left(\frac{x-X}{h}\right)\right] = E\left[YK\left(\frac{x-X^*}{h}\right)\right] + \\ & \sum_{r\geq 1}\sum_{k_1\geq 1}\cdots\sum_{k_r\geq 1}\frac{(-1)^{k_1+\ldots+k_r+r}}{k_1!\ldots k_r! h^{k_1+\ldots+k_r}}u_{k_1}\ldots u_{k_2}E\left[YK^{(k_1+\ldots+k_r)}\left(\frac{x-X^*}{h}\right)\right].\end{aligned} \tag{1.24}$$

Then, by (1.20) and (1.24), an estimator of the second order (i.e., $\nu = 2$) approximation to $E(\bar{g}_{NW})$ is

$$\hat{g}_{NW}(x) = \frac{\sum_{i=1}^{n} Y_i L\left(\frac{x-X_i^*}{h}\right)}{\sum_{i=1}^{n} L\left(\frac{x-X_i^*}{h}\right)},$$

where $L(v) = K(v) - (\hat{u}_2/(2h^2))K^{(2)}(v)$. Of course, estimators of higher orders approximations can be constructed similarly.

1.6 Deconvolution Problems in Image Analysis

As discussed in Section 1, the deconvolution problem is an ill-posed inverse problem. There are similar ill-posed inverse problems in different disciplines and areas. In image processing, the image deblurring problem is a such inverse problem, which will be discussed below. In image deblurring, we aim to recover a true image f from its blurred and noisy version Z, where the relationship between Z and f can be described by the following model:

$$Z(x, y) = (h * f)(x, y) + \varepsilon(x, y), \text{ for } (x, y) \in \Omega, \tag{1.25}$$

where $\varepsilon(x, y)$ denotes pointwise noise at (x, y), Ω is the design space of the image, and $(h*f)(x, y) = \int\int f(x-u, y-v)h(u, v)\, dudv$ is a spatially blurred version of f. Obviously, the blurred image $(h * f)$ is the convolution between a *point spread function* (psf) h and the true image f, and the blurring mechanism is described by the psf h. Similar to the density deconvolution problem (1.1), the image deblurring problem (1.25) is ill-posed in the following sense. In the case when h is unknown, there could be many different sets of h and f that correspond to the same observed image. Even in cases when h is completely known, an estimated true image is often numerically unstable, as discussed in Section 1 about the inverse problem (1.1). Thus, image deblurring is a challenging problem and has attracted much attention in the image processing community (cf., Campisi and Egiazarian 2016, Hansen et al. 2006, Qiu 2007).

Image deblurring methods that assume a known h are often called *non-blind image deblurring* methods. The assumption that h is known might be reasonable in certain applications. For instance, the linear psf h is appropriate for describing the image blur caused by relative location move between the image acquisition device and the object. The Gaussian blur is often used for describing image blur caused by atmospheric turbulence in remote sensing and aerial imaging. After h is specified, f can be estimated based on the relationship that

$$\mathcal{F}\{Z\}(u, v) = \mathcal{F}\{h\}(u, v)\mathcal{F}\{f\}(u, v) + \mathcal{F}\{\varepsilon\}(u, v), \text{ for } (u, v) \in \mathbb{R}^2,$$

where $\mathcal{F}\{g\}$ denotes the Fourier transform of function g. Many estimators of f, called inverse filters in the image processing literature, have been proposed. A popular one is the Wiener filter (Gonzalez and Woods, 2008), defined by

$$\hat{f}(x,y) = \frac{1}{(2\pi)^2}\mathcal{R}\left\{\int\int \frac{\overline{\mathcal{F}\{h\}(u,v)}\mathcal{F}\{Z\}(u,v)}{|\mathcal{F}\{h\}(u,v)|^2 + \alpha(u^2+v^2)^{\beta/2}} \exp\{-i(ux+vy)\}\, dudv\right\}, \tag{1.26}$$

where $\overline{\mathcal{F}\{h\}(u,v)}$ denotes the complex conjugate of $\mathcal{F}\{h\}(u,v)$, $\mathcal{R}\{C\}$ denotes the real part of the complex number C, and $\alpha, \beta > 0$ are two parameters. It can be seen that $\hat{f}(x,y)$ in (1.26) bears some resemblance with the deconvolution kernel density estimator in Stefanski and Carroll (1990). In (1.26), inclusion of the term $\alpha(u^2+v^2)^{\beta/2}$ is mainly for handling the noise effect. If the pointwise noise in the observed image can be ignored, then we actually do not need this term. In cases where the pointwise noise is substantial, the noise effect would dominate the image estimator, because $\mathcal{F}\{f\}(u,v)$ and $\mathcal{F}\{h\}(u,v)$ usually converge to zero rapidly, as u^2+v^2 approaches infinity, but $\mathcal{F}\{\varepsilon\}(u,v)$ goes to zero much more slowly.

In many applications, however, it is difficult to specify the psf h completely based on our prior knowledge about the image acquisition device. Image deblurring when h is unknown is often referred to as *blind image deblurring*. A popular blind image deblurring approach in the literature is based on the total variation (TV) minimization. Based on the work by You and Kaveh (1996), Chan and Wong (1998) introduced the TV blind deblurring procedure as

$$\min_{f,h\in BV}\left\{||h*f - Z||_{L^2}^2 + \lambda_f\int_\Omega |\nabla h(x,y)|dxdy + \lambda_h\int_\Omega |\nabla f(x,y)|dxdy\right\}, \tag{1.27}$$

where λ_f and λ_h are two positive parameters, ∇ is the gradient operator, and the space of all functions with bounded variation (BV) is defined as $BV = \{s \in L^1(\Omega) : \int_\Omega |\nabla s(x,y)|dxdy < \infty\}$. Clearly, in (1.27), the first term measures the goodness-of-fit of the estimators, and the second and third terms regularize their total variations. Chan and Wong (1998) solved the minimization problem by an iterative algorithm, after λ_f and λ_h are properly selected. More recently, Hall and Qiu (2007b) suggested estimating the psf h from an observed test image of an imaging device. The true test image considered in the paper has a square block in the middle with a uniform background. The idea in the paper, however, can also be applied to cases when the test image has step edges at known locations with uniform background. In a follow-up research, Hall and Qiu (2007a) suggested a two-step image deblurring procedure. In the first step, the psf h was estimated from an observed test image. Then, in the second step, any observed image $Z(x,y)$ taken by the same camera could be deblurred using the estimated psf obtained in the first step. The psf h in this method was assumed to have a parametric form. This assumption was later removed in the paper Qiu (2008). A more flexible blind image deblurring

method without restrictive assumptions on both h and f was proposed in Qiu and Kang (2015), based on the hierarchical nature of image blurring that the image structure was altered most significantly around step edges, less significantly around roof/valley edges, and least significantly at places where the true image intensity function was straight. For related discussions, see papers Kang (2020), Kang et al. (2018) and Kang and Qiu (2014).

1.7 Some Concluding Remarks

In the previous sections, we introduced some nonparametric deconvolution methods based on the discrete Fourier transformation and other series expansion approaches. These methods have their own strengths and limitations. For instance, the methods based on the Fourier series expansion are computationally convenient, but they cannot guarantee the estimators to be legitimate probability densities. The methods based on low order approximations are flexible to use in practice, because they only require information about the low order moments of the error distribution. However, they may not be statistically consistent. Therefore, the deconvolution problem is far away from being solved satisfactorily and much more future research is required. In addition, the deconvolution problem and its variants have broad applications in econometrics, signal processing, image restoration and many other areas. The related inverse problems in these applications often have similar structures (e.g., they are usually ill-posed), but with their own features and special properties. It should be helpful if the related research in different areas are better connected.

Acknowledgments: The authors thank Professor Aurore Delaigle for her invitation and many helpful suggestions. This research is supported in part by an NSF grant.

Bibliography

Abramovich, F. u. and B. Silverman (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika 85*(1), 115–129.

Antoniadis, A., G. Gregoire, and I. McKeague (1994). Wavelet methods for curve estimation. *Journal of the American Statistical Association 89*(428), 1340–1353.

Apanasovich, T. and H. Liang (2021). Nonparametric measurement errors models for regression. In G. Yi, A. Delaigle, and P. Gustafson (Eds.), *Handbook of Measurement Error Models*, Chapter X, pp. XXX–XXX. CRC.

Bigot, J. and S. Van Bellegem (2009). Log-density deconvolution by wavelet thresholding. *Scandinavian Journal of Statistics 36*(4), 749–763.

Campisi, P. and K. Egiazarian (2016). *Blind image deconvolution: theory and applications*. CRC press.

Carroll, R. J. and P. Hall (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association 83*(404), 1184–1186.

Carroll, R. J. and P. Hall (2004). Low order approximations in deconvolution and regression with errors in variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(1), 31–46.

Chan, T. F. and C.-K. Wong (1998). Total variation blind deconvolution. *IEEE transactions on Image Processing 7*(3), 370–375.

Delaigle, A. (2021). Deconvolution kernel density estimation. In G. Yi, A. Delaigle, and P. Gustafson (Eds.), *Handbook of Measurement Error Models*, Chapter X, pp. XXX–XXX. CRC.

Delaigle, A. and I. Gijbels (2006). Estimation of boundary and discontinuity points in deconvolution problems. *Statistica Sinica*, 773–788.

Delaigle, A., P. Hall, and P. Qiu (2006). Nonparametric methods for solving the berkson error-in-variables problem. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(2), 201–220.

Delaigle, A. and I. Van Keilegom (2021). Deconvolution with unknown error distribution. In G. Yi, A. Delaigle, and P. Gustafson (Eds.), *Handbook of Measurement Error Models*, Chapter X, pp. XXX–XXX. CRC.

Devroye, L. (1989). Consistent deconvolution in density estimation. *Canadian Journal of Statistics 17*(2), 235–239.

Diggle, P. J. and P. Hall (1993). A fourier approach to nonparametric deconvolution of a density estimate. *Journal of the Royal Statistical Society. Series B (Methodological)*, 523–531.

Donoho, D. L. and I. M. Johnstone (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association 90*(432), 1200–1224.

Efromovich, S. (1994). Nonparametric curve estimation from indirect observations. *Computing Science and Statistics 26*, 196–200.

Efromovich, S. (1997). Density estimation for the case of supersmooth measurement error. *Journal of the American Statistical Association 92*(438), 526–535.

Efromovich, S. (1999). *Nonparametric curve estimation: Methods, Theory, and Applications*. New York: Springer.

Fan, J. (1991a). Global behavior of deconvolution kernel estimates. *Statistica Sinica*, 541–551.

Fan, J. (1991b). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 1257–1272.

Fan, J. (1993). Adaptively local one-dimensional subproblems with application to a deconvolution problem. *The Annals of Statistics*, 600–610.

Fan, J. and J.-Y. Koo (2002). Wavelet deconvolution. *IEEE transactions on information theory 48*(3), 734–747.

Gonzalez, R. C. and R. E. Woods (2008). *Digital Image Processing*. Pearson Prentice Hall.

Hall, P., G. Kerkyacharian, D. Picard, et al. (1998). Block threshold rules for curve estimation using kernel and wavelet methods. *The Annals of Statistics 26*(3), 922–942.

Hall, P. and P. Patil (1995). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *The Annals of Statistics*, 905–928.

Hall, P., S. Penev, G. Kerkyacharian, and D. Picard (1997). Numerical performance of block thresholded wavelet estimators. *Statistics and Computing 7*(2), 115–124.

Hall, P. and P. Qiu (2005). Discrete-transform approach to deconvolution problems. *Biometrika 92*(1), 135–148.

Hall, P. and P. Qiu (2007a). Blind deconvolution and deblurring in image analysis. *Statistica Sinica*, 1483–1509.

Hall, P. and P. Qiu (2007b). Nonparametric estimation of a point-spread function in multivariate problems. *The Annals of Statistics*, 1512–1534.

Hansen, P. C., J. G. Nagy, and D. P. O'leary (2006). *Deblurring images: matrices, spectra, and filtering*. SIAM.

Hernández, E. and G. Weiss (1996). *A first course on wavelets*. CRC press.

Kang, Y. (2020). Consistent blind image deblurring using jump-preserving extrapolation. *Journal of Computational and Graphical Statistics 29*(2), 372–382.

Kang, Y., P. Mukherjee, and P. Qiu (2018). Efficient blind image deblurring using nonparametric regression and local pixel clustering. *Technometrics 60*, in press.

Kang, Y. and P. Qiu (2014). Jump detection in blurred regression surfaces. *Technometrics 56*, 539–550.

Kerkyacharian, G. and D. Picard (1992). Density estimation in besov spaces. *Statistics & Probability Letters 13*(1), 15–24.

Koo, J.-Y. and B. U. Park (1996). B-spline deconvolution based on the em algorithm. *Journal of Statistical Computation and Simulation 54*(4), 275–288.

Kooperberg, C. and C. J. Stone (1991). A study of logspline density estimation. *Computational Statistics & Data Analysis 12*(3), 327–347.

Liu, M. C. and R. L. Taylor (1989). A consistent nonparametric density estimator for the deconvolution problem. *Canadian Journal of Statistics 17*(4), 427–438.

Masry, E. (1991). Multivariate probability density deconvolution for stationary random processes. *IEEE Transactions on Information Theory 37*(4), 1105–1115.

Masry, E. (1993). Strong consistency and rates for deconvolution of multivariate densities of stationary processes. *Stochastic processes and their applications 47*(1), 53–74.

Masry, E. (1994). Probability density estimation from dependent observations using wavelets orthonormal bases. *Statistics & Probability Letters 21*(3), 181–194.

Mendelsohn, J. and J. Rice (1982). Deconvolution of microfluorometric histograms with b splines. *Journal of the American Statistical Association 77*(380), 748–753.

Penskaya, M. (1985). Projection estimators of the density of an a prior distribution and of functionals of it. *Theor. Probab. and Math. Statist 31*, 113–124.

Pensky, M. (2002). Density deconvolution based on wavelets with bounded supports. *Statistics & probability letters 56*(3), 261–269.

Pensky, M. and B. Vidakovic (1999). Adaptive wavelet estimator for nonparametric density deconvolution. *The Annals of Statistics 27*(6), 2033–2053.

Qiu, P. (2007). Jump surface estimation, edge detection, and image restoration. *Journal of the American Statistical Association 102*, 745–756.

Qiu, P. (2008). A nonparametric procedure for blind image deblurring. *Computational Statistics and Data Analysis 52*, 4828–4841.

Qiu, P. and Y. Kang (2015). Blind image deblurring using jump regression analysis. *Statistica Sinica 25*, 879–899.

Stefanski, L. A. (1990). Rates of convergence of some estimators in a class of deconvolution problems. *Statistics & Probability Letters 9*(3), 229–235.

Stefanski, L. A. and R. J. Carroll (1990). Deconvolving kernel density estimators. *Statistics 21*(2), 169–184.

Stone, C. J. and C.-Y. Koo (1986). Logspline density estimation. *Contemp. Math 59*(1), 15.

Taylor, R. and H. Zhang (1990). On a strongly consistent nonparametric density estimator for the deconvolution problem. *Communications in Statistics-Theory and Methods 19*(9), 3325–3342.

Tikhonov, A. N. and V. Y. Arsenin (1977). *Solutions of ill-posed problems*. New York: John Wiley.

Walter, G. (1981). Orthogonal series estimators of the prior distribution. *Sankhyā: The Indian Journal of Statistics, Series A*, 228–245.

Walter, G. G. (1994). *Wavelets and Other Orthogonal Systems with Applications*. CRC Press, Boca Raton, FL.

Walter, G. G. (1999). Density estimation in the presence of noise. *Statistics & probability letters 41*(3), 237–246.

You, Y.-L. and M. Kaveh (1996). A regularization approach to joint blur identification and image restoration. *IEEE Transactions on Image Processing 5*(3), 416–428.

Zayed, A. I. and G. G. Walter (1996). Characterization of analytic functions in terms of their wavelet coefficients. *Complex Variables and Elliptic Equations 29*(3), 265–276.

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