Aurore Delaigle

Deconvolution kernel density estimation

1

Deconvolution kernel density estimation

CONTENTS

1.1 Introduction 1
1.2 Model and data 2
1.3 Deconvolution kernel density estimator 3
1.4 Overview of some theoretical properties 5
1.4.1 Error type 6
1.4.2 Mean integrated squared error 6
1.4.3 Asymptotic mean integrated squared error 7
1.4.4 Fourier domain and supersmooth distributions of X 8
1.4.5 Further reading 10
1.5 Computing the estimator in practice 10
1.5.1 Importance of the bandwidth 10
1.5.2 Importance of the kernel 11
1.5.3 Computing the estimator in practice 13
1.6 Bandwidth selection in practice 14
1.6.1 Cross-validation bandwidth 15
1.6.2 Plug-in and normal reference bandwidths 16
1.6.3 Bootstrap bandwidth 18
1.6.4 SIMEX bandwidth 19
1.6.5 Further reading 21
1.7 Generalisations 22
1.7.1 Settings with analytic inversion formulae 22
1.7.2 When the Fourier transform vanishes 24
1.7.3 Heteroscedatic errors 26
1.7.4 Multivariate and dependent cases 26
1.7.5 Further reading 28

“nobreak

1.1 Introduction

In this chapter, we consider nonparametric estimation of the density f_X of a variable X which is observed with an independent additive noise U of known distribution. This problem has received a lot of attention in the literature, and the most popular estimator is the deconvolution kernel density estimator. In this chapter, we introduce this estimator and study some of its theoretical and practical properties. We also discuss some extensions and related problems. Alternative nonparametric procedures of density deconvolution are discussed in Chapter 11 (Kang and Qiu, 2021) and the case where the error distribution is unknown is studied in Chapter 12 (Delaigle and Van Keilegom, 2021). The regression case is discussed in Chapter 14 (Apanasovich and Liang, 2021).

This chapter is organised as follows. We introduce the classical measurement error model in Section 1.2 and the deconvolution kernel density estimator in Section 1.3. In Section 1.4 we discuss some of its L_2 theoretical properties. These depend heavily on the smoothness of the error distribution, and we introduce two types of errors usually considered in the literature: ordinary smooth and supersmooth errors. We compute the mean integrated squared error of the estimator for the two types of errors and study its rates of convergence for various levels of smoothness of the density f_X. Computing the deconvolution kernel density estimator in practice requires to choose a kernel function and a smoothing parameter called bandwidth. We discuss the impact of those choices in Section 1.5, where we also present some of the numerical issues that can be encountered when computing the deconvolution kernel density estimator. The choice of the bandwidth is particularly important: in order for the estimator to work well, the bandwidth must be selected in a carefully designed data-driven way. We dedicate Section 1.6 to several data-driven procedures of bandwidth selection. Finally, we discuss generalisations of the classical problem in Section 1.7 to the cases where the data are dependent or multivariate, where the measurement errors are not identically distributed and the characteristic function of the errors vanishes at some isolated points. We also show that in some cases, it is possible to compute the estimator using a simple analytic formula.

1.2 Model and data

In the classical measurement error problem, we are interested in a variable X but can only observe independent and identically distribution (i.i.d.) data $X_1^*, \ldots, X_n^*$ from a noisy version X^* of X. The variable X^* comes from the classical measurement error model

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

where U represents a measurement error independent of X, and which has density f_U. There are many applications where it is reasonable to assume that the noisy data come from (1.1), but often the model holds after transformation. For example, (1.1) is often employed for measured food consumption in nutrition studies. In those studies, food intake is often obtained from 24 hour recalls, where individuals report their food intake of the last 24 hours. It is common to assume that the model at (1.1) holds for $X^* = \log(\text{reported intake from 24 hour recalls})$. There, X is the logarithm of the usual intake, which can be roughly described as the average intake of an individual over a long period of time. In this chapter, we follow the literature on nonparametric deconvolution and ignore such transformations; that is, we assume that the data come directly from (1.1). For comprehensive reviews of the measurement error problem and techniques in the parametric context, see Carroll et al. (2006) and Buonaccorsi (2010).

Throughout this chapter we assume that the error density f_U is known and is an even function. It is possible to deal with the case where f_U is unknown, as long as it can be estimated nonparametrically, either from a sample from f_U (Diggle and Hall, 1993; Neumann, 1997), from replicated noisy measurements of all or some of the individuals (Delaigle et al., 2008), or even without additional data but under some conditions that permit its identifiability (Delaigle and Hall, 2016); it will be the topic of Chapter 12 (Delaigle and Van Keilegom, 2021), where we will also discuss parametric and semiparametric procedures for estimating f_U.

1.3 Deconvolution kernel density estimator

In measurement error problems, it is often of interest to estimate the density f_X of X from data $X_1^*, \ldots, X_n^*$ coming from (1.1). For example, if X represents the usual intake of a nutrient, knowing its density can help understand the food consumption of individuals in a population. The most popular nonparametric estimator of f_X in this context is the deconvolution kernel density estimator of Stefanski and Carroll (1990) and Carroll and Hall (1988). It is constructed by noting that under (1.1), we have

$$\varphi_{X^*} = \varphi_X \, \varphi_U \,,$$

where throughout this chapter we use $\varphi_T(t) = E(e^{iTt})$ to denote the characteristic function of a variable T or the Fourier transform $\int e^{itx} T(x)\, dx$ of an absolutely integrable function T, and i denotes the complex number such that

$i^2 = -1$. Therefore, if $\varphi_U(t) \neq 0$ for all $t \in \mathbb{R}$, then $\varphi_X = \varphi_X^*/\varphi_U$, and by the Fourier inversion theorem, we deduce that

$$f_X(x) = \frac{1}{2\pi}\int e^{-itx}\varphi_X(t)\,dt = \frac{1}{2\pi}\int e^{-itx}\frac{\varphi_X^*(t)}{\varphi_U(t)}\,dt\,. \tag{1.2}$$

Since f_U is known, then φ_U is known and the only unknown in the second integral at (1.2) is $\varphi_X^*(t)$, which can be easily estimated by the empirical characteristic function $\hat{\varphi}_{X^*}(t) = n^{-1}\sum_{j=1}^n e^{itX_j^*}$. It is tempting to construct an estimator of $f_X(x)$ by plugging $\hat{\varphi}_{X^*}$ into that integral. However, $\hat{\varphi}_{X^*}(t)$ is very unreliable in the tails, i.e., for large $|t|$, whereas $\varphi_U(t) \to 0$ as $|t| \to \infty$; as a result, $\hat{\varphi}_X^*(t)/\varphi_U(t)$ is not integrable. To overcome these unreliable tail fluctuations, Stefanski and Carroll (1990) and Carroll and Hall (1988) introduced a weight function $w(t)$, which is such that $w(t)$ is close to one when $\hat{\varphi}_X^*(t)$ is reliable, and close to zero elsewhere. Specifically, they took

$$\hat{f}_X(x) = \frac{1}{2\pi}\int e^{-itx}\frac{\hat{\varphi}_X^*(t)w(t)}{\varphi_U(t)}\,dt\,,$$

where $w(t) = \varphi_K(ht)$, with a smoothing parameter $h > 0$, called bandwidth, and a univariate smooth function K that integrates to 1, called kernel.

Thus, the deconvolution kernel density estimator of Stefanski and Carroll (1990) and Carroll and Hall (1988) is defined by

$$\begin{aligned}\hat{f}_X(x;h) =& \frac{1}{2\pi}\int e^{-itx}\hat{\varphi}_X^*(t)\frac{\varphi_K(ht)}{\varphi_U(t)}\,dt && (1.3)\\ =& \frac{1}{nh}\sum_{j=1}^n K_U\Big(\frac{x-X_j^*}{h}\Big)\,, && (1.4)\end{aligned}$$

where the deconvolution kernel K_U is defined by

$$K_U(x) = \frac{1}{2\pi}\int e^{-itx}\frac{\varphi_K(t)}{\varphi_U(t/h)}\,dt\,. \tag{1.5}$$

The following conditions guarantee that this estimator is well defined :

$$\varphi_U(t) \neq 0 \text{ for all } t\,; \tag{1.6}$$

$$\int |\varphi_X(t)|\,dt < \infty; \tag{1.7}$$

$$\sup_{t\in\mathbb{R}} |\varphi_K(t)/\varphi_U(t/h)| < \infty \quad\text{and}\quad \int |\varphi_K(t)/\varphi_U(t/h)|\,dt < \infty\,. \tag{1.8}$$

Liu and Taylor (1989) discussed a variant of this estimator where they truncated the domain of the integral at (1.5) to an interval $[-M_n, M_n]$, where $M_n \to \infty$ as $n \to \infty$. However, such truncation is not needed when φ_K is compactly supported (something very common in the deconvolution problem, see Section 1.5.2), in which case their estimator reduces to the deconvolution kernel density estimator. Even when φ_K is not compactly supported, the advantage of using two parameters (h and M_n) is unclear, as reflected by the numerical results in Liu and Taylor (1990), who found that better results were often obtained by taking $h = 0$ when M_n was positive. Another variant of the deconvolution kernel estimator that is consistent under the L_1 norm was introduced by Devroye (1989), and Zhang (1990) studied a closely related problem of estimation of mixing densities.

Although in principle the weight function $w(t)$ above could take various forms, taking $w(t) = \varphi_K(ht)$ has several useful interpretations. First, when $U \equiv 0$, that is when there are no errors and $\varphi_U(t) = 1$ for all t, this estimator reduces to the standard kernel density estimator; indeed, in that case (1.3) reduces to $\hat{f}_X(x;h) = \hat{f}_X^*(x;h)$, where

$$\hat{f}_X^*(x;h) = \frac{1}{2\pi}\int e^{-itx}\hat{\varphi}_X^*(t)\varphi_K(ht)\,dt = \frac{1}{nh}\sum_{j=1}^{n} K\Big(\frac{x - X_j^*}{h}\Big) \tag{1.9}$$

is the standard kernel density estimator of $f_X^*(x)$. Second, by Fourier inversion of (1.9), we have $\varphi_{\hat{f}_X^*(\cdot;h)}(t) = \hat{\varphi}_X^*(t)\varphi_K(ht)$. Comparing with (1.3), we see that the deconvolution kernel estimator at (1.3) is nothing but the estimator of $f_X(x)$ obtained by replacing φ_{X^*} in the second integral at (1.2) by the Fourier transform of the kernel density estimator of f_{X^*}. See also Delaigle (2014) for a discussion of main principles of deconvolution, including a description of the general unbiased score technique for deconvolution that also leads to the estimator at (1.4).

1.4 Overview of some theoretical properties

In this section we review some of the most important theoretical properties of the deconvolution kernel density estimator. As we shall see, they depend heavily on the smoothness of the error distribution. In Section 1.4.1 we introduce two types of error distributions (ordinary smooth and supersmooth) typically encountered in the deconvolution literature. We describe the mean integrated squared error (MISE) of the estimator in Section 1.4.2 and its asymptotic expression in Section 1.4.3. As we shall see, standard convergence based on the this asymptotic expression is often slow. However, such rates are developed under the assumption that f_X only has a finite number of derivatives. In Section 1.4.4 we will see that by studying the MISE of the estimator in the Fourier domain, it is possible to show that if f_X is infinitely differentiable then the convergence rates are considerably faster. Finally, in Section 1.4.5 we mention some other useful theoretical properties that have been established in the literature.

While we will mention some of the important conditions, in this section, to keep the text readable, we will not list all the technical conditions required to write the results we present. We refer to the original papers and to Meister (2009a) for a deep and rigourous account of theoretical properties.

1.4.1 Error type

The rate of convergence of $\hat{f}_X$ to f_X depends on the smoothness of the error distribution, which is characterised by the rate of decay of its characteristic function in the tails. Following the terminology in Fan (1991a,b,c), one typically distinguishes between two classes of errors called supersmooth and ordinary smooth. An error U is supersmooth of order β if, for some constants $\beta_0 \le \beta_1$, $0 < d_0 \le d_1$, $\beta > 0$ and $\gamma > 0$,

$$d_0|t|^{\beta_0} \exp(-|t|^\beta/\gamma) \le |\varphi_U(t)| \le d_1|t|^{\beta_1} \exp(-|t|^\beta/\gamma) \quad \text{for large } |t|\,; \quad (1.10)$$

for example, normal and Cauchy distributions are supersmooth. As noted by Butucea and Tsybakov (2008a,b), most densities in this class that are well known and can be expressed in a closed form are such that $\beta \le 2$.

An error U is ordinary smooth of order β if, for some constants $0 < d_0 \le d_1$ and $\beta > 0$,

$$d_0|t|^{-\beta} \le |\varphi_U(t)| \le d_1|t|^{-\beta} \quad \text{for large } |t|\,; \quad (1.11)$$

for example, a Laplace distribution is ordinary smooth.

We will see that supersmooth errors make the deconvolution problem much more difficult than ordinary smooth errors.

1.4.2 Mean integrated squared error

Theoretical properties of $\hat{f}_X$ are usually assessed via the mean integrated square error defined by

$$\text{MISE}(h) = \int \text{Bias}^2\{\hat{f}_X(x;h)\}\,dx + \int \text{var}\{\hat{f}_X(x;h)\}\,dx\,,$$

where we follow the usual approach in the literature and omit the dependence on the kernel K in the notation.

It follows from the conditional unbiased score property

$$E\big\{K_U(x - X_j^*)\big|X_j\big\} = E\big\{K(x - X_j)\big\}$$

which is at the heart of the deconvolution kernel technique (Stefanski and Carroll, 1990; Delaigle, 2014) that the bias of the deconvolution kernel density estimator is the same as that of the error-free kernel density estimator:

$$\text{Bias}\{\hat{f}_X(x;h)\} = E\{\hat{f}_X(x;h)\} - f_X(x) = K_h * f_X(x) - f_X(x)\,,$$

where $K_h(x) = K(x/h)/h$ and $f * g(x) = \int f(x-u)g(u)\,du$ denotes the convolution product of two functions f and g. In particular, the MISE of the deconvolution kernel estimator computed from the contaminated X_j^*'s differs from that of the standard kernel density estimator computed from the error-free X_j's only through its variance.

The integrated variance of the deconvolution kernel density estimator is equal to (Stefanski and Carroll, 1990; Stefanski, 1990)

$$\int \operatorname{var}\{\hat{f}_X(x;h)\}\,dx = \frac{1}{2\pi nh}\int \frac{|\varphi_K(t)|^2}{|\varphi_U(t/h)|^2}\,dt - n^{-1}\int (K_h * f_X)^2(x)\,dx\,.$$

More formally, Stefanski and Carroll (1990) showed that under Conditions (1.6) to (1.8) and if K is integrable, we have

$$\begin{aligned}\mathrm{MISE}(h) =& \frac{1}{2\pi nh}\int \frac{|\varphi_K(t)|^2}{|\varphi_U(t/h)|^2}\,dt + (1-n^{-1})\int (K_h * f_X)^2(x)\,dx \\ &+ \int f_X^2(x)\,dx - 2\int K_h * f_X(x) f_X(x)\,dx\,.\end{aligned}$$

1.4.3 Asymptotic mean integrated squared error

As in the error-free case, the MISE is difficult to interpret, and it is standard to analyse instead its asymptotically dominating part. For this, it is useful to recall that a kth order kernel K is a kernel whose moments satisfy

$$\mu_{K,j} = \int x^j K(x)\,dx = \begin{cases} 1 & \text{for } j = 0 \\ 0 & \text{for } j = 1,\dots,k-1 \\ c & \text{for } j = k\,, \end{cases}$$

where $c \neq 0$ is a finite constant. As usual with kernel density estimation, K is almost always chosen to be symmetric around zero, so that k is even (indeed, for k odd we could not have $\mu_{K,k} = c \neq 0$). Using a Taylor expansion, Stefanski and Carroll (1990) showed that under (1.6) to (1.8), if K is a kth order kernel such that $\int |x^{k+1}K(x)|\,dx < \infty$, f_X has $k+1$ continuous and bounded derivatives, $f_X^{(k)}$ is square integrable and $h \to 0$ and $nh \to \infty$ as $n \to \infty$, then $\mathrm{MISE}(h) = \mathrm{AMISE}(h) + O(n^{-1}) + o(h^{2k})$, where

$$\mathrm{AMISE}(h) = \frac{h^{2k}}{(k!)^2}\mu_{K,k}^2 \int \{f_X^{(k)}(x)\}^2\,dx + \frac{1}{2\pi nh}\int \frac{|\varphi_K(t)|^2}{|\varphi_U(t/h)|^2}\,dt\,. \tag{1.12}$$

We learn from this expression that the rate of convergence of $\hat{f}_X$ to f_X depends on the smoothness of f_X and the smoothness of the error distribution; see Carroll and Hall (1988), Stefanski and Carroll (1990) and Fan (1991b,c) for detailed calculations and asymptotic results. In particular, those authors have shown that, under sufficient conditions, in the ordinary smooth error case at (1.11), $\int |\varphi_K(t)|^2 |\varphi_U(t/h)|^{-2}\,dt \sim h^{-2\beta}$, so that $\mathrm{AMISE}(h) \sim c_1 h^{2k} +$

$c_2/(nh^{2\beta+1})$, where c_1 and c_2 are positive constants. Thus, in the ordinary smooth error case, if we use a kth order kernel and f_X is smooth enough, the estimator $\hat{f}_X$ converges the fastest by taking $h \sim n^{-1/(2\beta+2k+1)}$, which results in $\text{MISE}(h) \sim \text{AMISE}(h) \sim n^{-2k/(2\beta+2k+1)}$. Fan (1991c) showed that, under sufficient smoothness conditions, these rates are optimal. Fan (1991a) also established asymptotic normality of the estimator.

Things are more involved in the supersmooth error case at (1.10), where the fast decay of φ_U in its tails makes it difficult to find kernels for which the integral $\int |\varphi_K(t)|^2 |\varphi_U(t/h)|^{-2}\, dt$ exists, except if we take kernels for which φ_K is compactly supported; the choice of the kernel will be discussed in Section 1.5.2. Furthermore, it is difficult to obtain an explicit expression for the integrated variance term, for which we typically only have upper bounds. Despite these complications, following Fan (1991a,b,c), it is possible to find bandwidths for which the rate of the $\text{AMISE}(h)$ is the smallest possible. Specifically, if we choose K such that φ_K is supported on $[-B, B]$ for some $0 < B < \infty$, then we have $\int |\varphi_K(t)|^2 |\varphi_U(t/h)|^{-2}\, dt = O\{h^{2\beta_0} \exp(2|B/h|^{\beta}/\gamma)\}$, which is of order $O\big\{n^{d^{-\beta}-1}(\log n)^{(-2\beta_0+1)/\beta}\big\}$, if we take $h = dB(2/\gamma)^{1/\beta}(\log n)^{-1/\beta}$, with $d > 1$. With the same choice of h, the integrated squared bias term is of exact order $(\log n)^{-2k/\beta}$, so that $\text{MISE}(h) \sim \text{AMISE}(h) \sim (\log n)^{-2k/\beta}$, since the integrated variance term is negligible compared to the integrated bias term.

Despite the pessimistic very slow convergence rates in the supersmooth case, deconvolution in practice works reasonably well even if the error is supersmooth. This has partly to do with the fact that these traditional asymptotic results do not take the variance of the errors into account, whereas these play an important role in the success of a deconvolution procedure. To take the magnitude of the error variance into account, several authors have developed a double asymptotic procedure where the error variance is assumed to tend to zero as $n \to \infty$. See for example Fan (1992), Carroll and Hall (2004), Delaigle (2008), Van Es and Gugushvili (2010) and Chapter 12 (Delaigle and Van Keilegom, 2021). Fan (1991c) showed that if f_X has a finite number of continuous derivatives, these rates are optimal. However, if f_X is very smooth, then using a finite order kernel can give rise to considerably suboptimal results, as we show in Section 1.4.4.

1.4.4 Fourier domain and supersmooth distributions of X

The asymptotic analysis in Section 1.4.3 and the rates of convergence discussed there only exploit the fact that f_X has a finite number, k, of derivatives. However, many densities have an infinite number of derivatives; for those, the convergence rate of the estimator, and that of alternative estimators (Pensky et al., 1999) is considerably faster if we use a so-called infinite order kernel. Specifically, convergence rates can be much faster if, as in Butucea (2004b) and Butucea and Tsybakov (2008a,b), f_X is such that

$$(2\pi)^{-1} \int |\varphi_X(t)|^2 \exp(2\alpha |t|^{\beta_X})\, dt \le L\,, \tag{1.13}$$

with $L > 0$, $\alpha > 0$ and where $\beta_X > 0$ are finite constants.

To understand why, using Parseval's identity, we express the MISE in the Fourier domain as (Stefanski, 1990)

$$\begin{aligned}\text{MISE}(h) =&\frac{1}{2\pi nh}\int \frac{|\varphi_K(t)|^2}{|\varphi_U(t/h)|^2}\,dt + \frac{(1-n^{-1})}{2\pi}\int |\varphi_X(t)|^2|\varphi_K(ht)|^2\,dt \\ &+\frac{1}{2\pi}\int |\varphi_X(t)|^2\,dt - \frac{1}{\pi}\int |\varphi_X(t)|^2\varphi_K(ht)\,dt\,. \end{aligned} \tag{1.14}$$

In the particular case of the sinc kernel K defined by $\varphi_K(t) = I\{|t| \leq 1\}$, where $I\{\cdot\}$ is the indicator function (see Section 1.5.2 for a detailed discussion of kernels), this reduces to

$$\text{MISE}(h) = \frac{1}{2\pi nh}\int_{-1}^{1} \frac{1}{|\varphi_U(t/h)|^2}\,dt + \frac{1}{2\pi}\int_{|t|>1/h} |\varphi_X(t)|^2\,dt + O(n^{-1})\,.$$

Now, using (1.13), we have

$$\begin{aligned}\frac{1}{2\pi}\int_{|t|>1/h} |\varphi_X(t)|^2\,dt =&\frac{1}{2\pi}\int_{|t|>1/h} |\varphi_X(t)|^2\exp(2\alpha|t|^{\beta_X})\exp(-2\alpha|t|^{\beta_X})\,dt \\ \leq&\frac{e^{-2\alpha h^{-\beta_X}}}{2\pi}\int_{|t|>1/h} |\varphi_X(t)|^2\exp(2\alpha|t|^{\beta_X})\,dt \\ \leq& L\exp(-2\alpha h^{-\beta_X})\,.\end{aligned}$$

On the other hand, we know from Section 1.4.3 that $\int |\varphi_U(t/h)|^{-2}\,dt = O\{h^{2\beta_0}\exp(2h^{-\beta}/\gamma)\}$ if U is supersmooth of order β, and $\int |\varphi_U(t/h)|^{-2}\,dt \sim h^{-2\beta}$ if U is ordinary smooth of order β. Thus, if U is ordinary smooth of order β then

$$\text{MISE}(h) \leq L\exp(-2\alpha h^{-\beta_X}) + c_1 h^{-2\beta-1}n^{-1} + O(n^{-1})\,,$$

with $c_1 > 0$ a constant; taking $h = \{\log n/(2\alpha)\}^{-1/\beta_X}$ as in Butucea (2004b), we deduce that $\text{MISE}(h) = O(n^{-1}) + O\big\{n^{-1}(\log n)^{(2\beta+1)/\beta_X}\big\} = O\big\{n^{-1}(\log n)^{(2\beta+1)/\beta_X}\big\}$, which is only slightly slower than the parametric n^{-1} rate, and faster than the rate $n^{-2k/(2\beta+2k+1)}$ obtained when assuming only that f_X has k smooth derivatives.

If U is supersmooth of order β, then we have

$$\text{MISE}(h) \leq L\exp(-2\alpha h^{-\beta_X}) + c_1 n^{-1}h^{2\beta_0-1}\exp(2h^{-\beta}/\gamma) + O(n^{-1})\,. \tag{1.15}$$

In that case too, Butucea and Tsybakov (2008a,b) showed that it is possible to choose a more precise bandwidth than the one suggested in Section 1.4.3, such that the rates are faster than the slow logarithmic rates obtained when assuming only that f_X has k smooth derivatives. Indeed, if we take $h \leq d(\log n)^{-1/\beta_X}$ for some constant $d > 0$, then the first term of (1.15) is of order n^{-a} for some $a > 0$, and if we take $h = d(2/\gamma)^{1/\beta}(\log n)^{-1/\beta}$ with $d > 1$, then we have seen in the previous section that the second term of (1.15) is of order $O\big\{n^{d^{-\beta}-1}(\log n)^{(-2\beta_0+1)/\beta}\big\}$. Thus, if $\beta \leq \beta_X$, then by taking $h = d(2/\gamma)^{1/\beta}(\log n)^{-1/\beta}$ with $d > 1$, the estimator converges at polynomial rates. The case where $\beta_X < \beta$ was studied by Butucea and Tsybakov (2008a,b), who proposed a more precise choice of h. They showed that, using that bandwidth, the bias contribution dominates the variance contribution, and that again the rate is faster than the one obtained when assuming that f_X has only a finite number k of derivatives.

1.4.5 Further reading

Stefanski (1990) established strong uniform consistency of the estimator, and Liu and Taylor (1989) established strong uniform consistency for their truncated estimator. Devroye (1989) established L_1 consistency of his modified kernel estimator and Song (2010) studied moderate deviations in the ordinary smooth error case. Asymptotic normality for supersmooth errors of order β was studied by Van Es and Uh (2004, 2005), who noted that, when considering pointwise properties of the estimator, the case where $\beta < 1$ behaves differently from that where $\beta > 1$; see also Butucea and Tsybakov (2008a) for a similar remark and Holzmann and Boysen (2006) for the asymptotic distribution of the integrated squared error of the estimator in the supersmooth error case, which corrects a result from Butucea (2004a). Finally, Zu (2015) established asymptotic normality with a logarithmic chi-square error.

1.5 Computing the estimator in practice

In practice, some care needs to be taken when computing the deconvolution kernel density estimator. First, computing the estimator requires the choice of a bandwidth and a kernel. We discuss these issues in Sections 1.5.1 and 1.5.2. Second, often the estimator has no analytic expression and needs to be computed numerically; we discuss issues related to this in Section 1.5.3.

1.5.1 Importance of the bandwidth

As in the error-free case, the choice of the bandwidth h is crucial for the empirical success of $\hat{f}_X$: a too small h will result in a too variable, wiggly, estimator, and a too large h will result in a biased, oversmoothed, estimator. In theory, the best choice of h is the one that minimises the distance between f_X and $\hat{f}_X$. There are many ways to choose such a distance. For example, a distance can be global (distance between the whole curves f_X and $\hat{f}_X$), or local (distance between $f_X(x)$ and $\hat{f}_X(x)$ at each x of interest). The two most popular global distances are the MISE or its asymptotic approximation AMISE, and the most popular local distance is the mean squared error

$$\mathrm{MSE}(x;h) = E\big[\{\hat{f}_X(x;h) - f_X(x)\}^2\big] = \mathrm{Bias}^2\{\hat{f}_X(x;h)\} + \mathrm{var}\{\hat{f}_X(x;h)\}\,.$$

or its asymptotic version $\mathrm{AMSE}(x;h)$.

We choose h by minimising a global distance when we intend to use the same bandwidth h at each x where we compute the estimator $\hat{f}_X$; such a bandwidth is called a global bandwidth. For example, the MISE and the AMISE bandwidths are defined by, respectively,

$$h_{\mathrm{MISE}} = \mathrm{argmin}_h \mathrm{MISE}(h)\,, \quad h_{\mathrm{AMISE}} = \mathrm{argmin}_h \mathrm{AMISE}(h)\,.$$

We choose h by minimising a local distance if we wish to use a different bandwidth $h(x)$ at each x; such a bandwidth is called a local bandwidth. For example, the MSE and the AMSE bandwidths are defined by, respectively,

$$h_{\mathrm{MSE}}(x) = \mathrm{argmin}_h \mathrm{MSE}(x;h)\,, \quad h_{\mathrm{AMISE}}(x) = \mathrm{argmin}_h \mathrm{AMSE}(x;h)\,.$$

A local bandwidth is preferable when f_X has sharp features (e.g., is very wiggly) in some parts of its domain and is very smooth (e.g., non wiggly) elsewhere. There, ideally h should be smaller in wiggly areas and larger elsewhere.

In practice, we cannot compute any of those theoretically optimal bandwidths since they all depend on the unknown f_X that we are trying to estimate. In Section 1.6 we will discuss various methods that have been developed in the literature for approximating them in practice.

1.5.2 Importance of the kernel

As in standard kernel density estimation problems without measurement errors, the choice of the kernel function K is less important than the choice of the bandwidth h. However, in order for the estimator at (1.3) to work well in practice, the kernel K needs to be a smooth and unimodal function that integrates to 1 (so that $\hat{f}_X$ integrates to 1 too). Moreover, in order for the estimator at (1.3) to be well defined, the integral at (1.5) needs to exist. Therefore, K has to be such that $\varphi_K(t)$ tends to zero faster than $\varphi_U(t/h)$ as $|t| \to \infty$, which often makes it impossible to use kernels that are frequently used in the error-free case such as the Epanechnikov kernel, defined by $K(x) = 3/4\,(1-x^2)\cdot I\{|x| \le 1\}$, or the standard normal kernel, defined by $K = \phi$, the standard normal density.

To understand why standard kernels such as the Epanechnikov kernel can often not be used in the error case, note that the Fourier transform of the Epanechnikov kernel is given by $\varphi_K(t) = 3(\sin t - t\cos t)/t^3$, which, for large $|t|$, behaves like $|t|^{-2}$. Thus, for ordinary smooth errors satisfying (1.11), we can only guarantee that the integral at (1.5) exists if $\beta \le 1$, which is not even satisfied by Laplace errors, for which β in (1.11) is equal to 2. It is easy to see that this integral cannot exist in the supersmooth error case.

On the other hand, the standard normal kernel can be used with ordinary smooths at (1.11). Indeed, in this case we have $\varphi_K(t) = \exp(-t^2/2)$, which tends to zero much faster than $|t|^{-\beta}$, so that the integral at (1.5) exists. An advantage of this kernel in this case is that its associated K_U has an explicit (analytic) formula, and so do various quantities required to compute a data-driven bandwidth. For supersmooth errors at (1.10) however, with this kernel, the integral at (1.5) exists only if $\beta < 2$, or if $\beta = 2$ and $h > \gamma^{-1/2}$. This is only possible for n small, since the estimator at (1.3) is only consistent if $h \to 0$ as $n \to \infty$. Therefore, this kernel is not used for supersmooth errors.

To guarantee that the integral at (1.5) exists for both supersmooth and ordinary smooth errors, it is common to use kernels whose Fourier transform is compactly supported. Two kernels are commonly employed: the sinc kernel $K_1(x) = \sin(x)/(\pi x)$, whose Fourier transform is equal to $\varphi_K(t) = I\{|t| \leq 1\}$ and the kernel K_2, defined through its characteristic function by $\varphi_{K_2}(t) = (1-t^2)^3 I\{|t| \leq 1\}$. A drawback of these kernels is that often they do not produce an analytic formula for K_U, which has to be computed by numerical integration; see Section 1.5.3.

While the sinc kernel K_1 is not absolutely integrable, it has the advantage that it is an infinite order kernel: it automatically adapts to the smoothness of f_X, which, in simple terms, is the number of smooth and bounded derivatives that f_X has. As in the error-free case, this means that, with this kernel, the deconvolution kernel density estimator of f_X tends to have a smaller bias and is able to better capture sharp features of f_X, sometimes much better than finite order kernels; see Section 1.4.4. However, this typically comes at the cost of pronounced unattractive negative wiggles in the tail of the estimator.

Lütkenöner (2015) proposed a family of kernels whose Fourier transform is of the form

$$\varphi_K(t) = \{\cos(\pi t/2)\}^\kappa I\{|t| \leq 1\}\,,$$

where $\kappa \geq 0$. For example, when $\kappa = 0$, $\varphi_K(t) = I\{|t| \leq 1\}$ corresponds to the sinc kernel, which does not have a single finite moment. When $\kappa \geq 2$ (resp., $\kappa \geq 4$), the corresponding kernels have at least 2 (resp., at least 4) finite moments. Lütkenöner (2015) showed that the advantage of using these kernels is that in the normal error case where $U \sim N(0, \sigma^2)$, for $\kappa > 0$, we can write K_U explicitly as

$$K_U(x) = \frac{1}{2^k} \sum_{k=0}^{\kappa} \binom{\kappa}{k} K_0\big\{x - (\kappa/2 - k)\pi\big\}\,,$$

where K_0 corresponds to K_U for the sinc kernel:

$$K_0(x) = \frac{-\lambda}{\sqrt{2\pi}} e^{1/(2\lambda^2)} \mathbb{I}\big[\exp(-i|x|) w\{(i\lambda|x| - 1/\lambda)/\sqrt{2}\}\big]\,,$$

with $\mathbb{I}$ denoting the imaginary part, $i^2 = -1$, $\lambda = h/\sigma$ and $w(z) = e^{-z^2}\text{erfc}(-iz)$, with erfc the complementary error function. As pointed by

Lütkenöner (2015), w is known as the Faddeeva function and there exist efficient algorithms to evaluate it numerically.

In the case without measurement errors, an optimal kernel of order k is sometimes defined to be a kernel of order k that minimises the MISE of the standard kernel estimator. Since that problem is degenerate (it does not have a unique solution), Granovsky et al. (1995) and Granovsky and Müller (1991, 1989) proposed to choose the kth order kernel that minimises the MISE and satisfies a number of additional side constraints, such as having $k - 2$ sign changes. Delaigle and Hall (2006) argued that these side constraints are not appropriate in the case with measurement errors, where kernels chosen in that way result in much poorer practical performance, compared to the kernels discussed above. They also showed that the choice of the kernel is much more important in the error case than in the error-free case, in that it influences much more the value of the MISE and the practical performance of the estimator.

1.5.3 Computing the estimator in practice

In general, computing the deconvolution kernel density estimator is not straightforward as it requires computing an inverse Fourier transform through K_U, which often does not have an analytic formula.

In some cases, it is possible to express the deconvolution kernel density estimator analytically without going through the Fourier domain. For example, in the case where U is a Laplace(σ) random variable, we have $\varphi_U(t) = 1/(1 + \sigma^2 t^2)$, so that

$$K_U(x) = \frac{1}{2\pi}\int e^{-itx}\varphi_K(t)\,dt + \frac{\sigma^2}{h^2}\cdot\frac{1}{2\pi}\int e^{-itx}t^2\varphi_K(t)\,dt = K(x) - \frac{\sigma^2}{h^2}K''(x)\,.$$

Thus, the deconvolution kernel density estimator can be expressed as

$$\hat{f}_X(x;h) = \hat{f}_{X^*}(x;h) - \sigma^2 \hat{f}''_{X^*}(x;h)\,. \tag{1.16}$$

where $\hat{f}_{X^*}(x;h)$ is the standard kernel density estimator at (1.9) and $\hat{f}''_{X^*}(x;h)$ is its second derivative with respect to x; see Section 1.7.1 for other examples.

In those simple cases, we can use a standard kernel of the same type as those used in the error-free case, for example the standard normal kernel, and the deconvolution kernel density estimator can be computed using this simple analytic formula. It is important to note that if we choose kernels with a compactly supported Fourier transform equal to a polynomial on its domain (e.g., K_1 or K_2 introduced in Section 1.5.2), then even if K_U has an analytic expression such as the one derived above, that expression often cannot be used directly as it can cause dramatic cancellation. For a detailed discussion of this issue, see Delaigle and Gijbels (2007), who also proposed a solution to this problem. However, if we bypass completely the computation of K_U and compute $\hat{f}_X$ directly, then we can avoid such problems; for example in the Laplace case we can avoid this problem if we compute $\hat{f}_{X^*}(x;h)$ and $\hat{f}''_{X^*}(x;h)$, and then take $\hat{f}_X(x;h) = \hat{f}_{X^*}(x;h) - \sigma^2 \hat{f}''_{X^*}(x;h)$.

More generally, and especially in the supersmooth error case, there is often no analytic formula for K_U and $\hat{f}_X$ has to be computed numerically. If we use (1.4) to compute $\hat{f}_X$, at each x we need to evaluate n integrals (one for each version of K_U in (1.4)). Delaigle and Gijbels (2007) have shown that computing K_U numerically is also quite difficult. Indeed, the periodic oscillations inside the integral at (1.5) can cause the integral to be poorly approximated if standard fast iterative integration algorithms are employed, such as the Romberg method. While a more accurate approximation can be obtained by the fast Fourier transform or the trapezoidal rule on a fine grid, the calculation of K_U can be bypassed entirely if we compute $\hat{f}_X$ numerically through the formula at (1.3). The latter only requires to evaluate one integral numerically for each x at which we compute $\hat{f}_X$.

As noted in Section 1.5.2 for the sinc kernel, the deconvolution kernel density estimator typically has some negative wiggles in the tail, although these vanish as $n \to \infty$ (recall that the estimator is consistent). More generally, even if we use a kernel K that is a density, the deconvolution kernel K_U is typically not a density (it integrates to 1 but has negative wiggles, whose magnitude increases to as h decreases and/or σ^2 increases). In the error-free case, it is sometimes advocated that a density estimator $\hat{f}$ that takes negative values should be truncated to zero and rescaled to integrate to 1 (Hall and Murison, 1993), that is, take

$$\tilde{f}(x) = \max\{\hat{f}(x), 0\} \Big/ \int \max\{\hat{f}(y), 0\}\, dy\,.$$

However, since K_U often has big negative wiggles, the negative wiggles in the tails of $\hat{f}_X$ in the error case can be much larger than those in the error-free case. Therefore, rescaling the estimator as above is not necessarily a good idea: as $\hat{f}_X$ (with its negative wiggles) integrates to 1, rescaled its truncated version will often introduce a large bias. Instead, in the error case, it is often more appropriate to take

$$\tilde{f}_X(x;h) = \max\{\hat{f}_X(x;h), 0\}\,,$$

without rescaling, even though this often implies that $\tilde{f}_X$ is not a density since it does not integrate exactly to 1.

1.6 Bandwidth selection in practice

In Section 1.5.1 we highlighted the importance of the choice of h for the empirical success of the estimator $\hat{f}_X$. There we defined some theoretical bandwidths that result in good practical performance, but none of them can be computed in practice since they all depend on the unknown f_X. In this section, we discuss several strategies for approximating those bandwidths from the data. As we shall see, some (e.g. cross-validation, bootstrap and SIMEX) are based on the MISE, and others (e.g. plug-in) are based on asymptotic expressions such as the AMISE at (1.12).

It is important to note that the AMISE at (1.12) relies on the kernel to have a finite number k of moments, and f_X to be at least k times differentiable. Therefore, techniques based on the AMISE cannot be employed for the sinc kernel. We will see that the cross-validation, the bootstrap and the SIMEX bandwidths can all be used with the sinc kernel. Most of the techniques discussed in this section are available in the R package `deconvolve` (Delaigle et al., 2020), while Matlab codes are available on Delaigle's webpage.

1.6.1 Cross-validation bandwidth

The first data-driven procedure for selecting h in practice was proposed by Stefanski and Carroll (1990). It was originally developed for normal errors but it can be applied more generally (Hesse, 1999). As in the standard error-free case, it is designed to provide an estimator of the bandwidth that minimises the integrated squared error $\text{ISE}(h) = \int \{\hat{f}_X(x) - f_X(x)\}^2\, dx$, or equivalently, which minimises

$$\text{ISE}(h) - \int \{f_X(x)\}^2\, dx = \int \{\hat{f}_X(x)\}^2\, dx - 2\int \hat{f}_X(x) f_X(x)\, dx\,. \tag{1.17}$$

In the Fourier domain, the first term on the right hand side of (1.17) can be expressed as $(2\pi)^{-1}\int |\varphi_K(ht)|^2|\hat{\varphi}_{X^*}(t)|^2|\varphi_U(t)|^{-2}\, dt$. Stefanski and Carroll (1990) showed that the second term has the same expectation as a cross-validation quantity that can be rewritten as $\{\pi(n-1)\}^{-1}\int \varphi_K(-ht)\{n|\hat{\varphi}_{X^*}(t)|^2 - 1\}|\varphi_U(t)|^{-2}\, dt$. Motivated by this, they defined the cross-validation bandwidth by

$$\hat{h}_{\text{CV}} = \text{argmin}_h \text{CV}(h)\,, \tag{1.18}$$

where CV(h) is the cross-validation criterion defined by

$$\text{CV}(h) = \frac{1}{2\pi}\int \frac{|\varphi_K(ht)|^2|\hat{\varphi}_{X^*}(t)|^2 - 2(n-1)^{-1}\varphi_K(-ht)\{n|\hat{\varphi}_{X^*}(t)|^2 - 1\}}{|\varphi_U(t)|^2}\, dt\,.$$

Theoretical properties of this bandwidth were studied by Hesse (1999) and Youndjé and Wells (2002) in the ordinary smooth error case; the authors showed that an advantage of the cross-validation bandwidth is that it relies on very few smoothness assumptions, unlike the plug-in technique introduced in the next section. However, in practice the CV bandwidth has a tendency to select too small bandwidths, which is useful for capturing sharp features such as sharp peaks, but it often comes at the price of an estimator that is too variable. Moreover, as in the error-free case, the cross-validation bandwidth is not always uniquely defined, and in that case it is not clear which local solution of (1.18) should be chosen; Delaigle and Gijbels (2004b) recommended to choose among them the smallest bandwidth for which the estimator appears visually smooth enough, or if this is not possible, to choose the largest solution.

In the particular case of the sinc kernel, CV simplifies into

$$\mathrm{CV}(h) = \frac{1}{2\pi(n-1)} \int_{-1/h}^{1/h} \frac{2-(n+1)|\hat{\varphi}_{X^*}(t)|^2}{|\varphi_U(t)|^2}\, dt \tag{1.19}$$

so that

$$\mathrm{CV}'(h) = \frac{1}{\pi(n-1)h^2} \frac{(n+1)|\hat{\varphi}_{X^*}(1/h)|^2 - 2}{|\varphi_U(1/h)|^2}\, .$$

Since $\hat{h}_{\mathrm{CV}}$ is a solution of $\mathrm{CV}'(h)=0$, in this case it can be found by solving $|\hat{\varphi}_{X^*}(1/h)|^2 = 2/(n+1)$. Interestingly, as pointed by Stefanski and Carroll (1990), since the bandwidth chosen by cross-validation does not depend on the error distribution, we see that in this case, it is the same for estimating f_X and for estimating f_{X^*}. As we will see later, this is generally not true for the approaches based on a finite order kernel.

As shown by Stefanski and Carroll (1990) the CV bandwidth can be viewed as an estimator of the bandwidth h_{MISE} that minimises the MISE. For example, in the case of the sinc kernel, using (1.14) and recalling that $\varphi_X = \varphi_{X^*}/\varphi_U$, we have

$$\mathrm{MISE}(h) = \frac{1}{2\pi n} \int_{-1/h}^{1/h} \frac{1-(n+1)|\varphi_{X^*}(t)|^2}{|\varphi_U(t)|^2}\, dt + \frac{1}{2\pi}\int |\varphi_X(t)|^2\, dt\, . \tag{1.20}$$

(Note that the last term does not depend on h). Therefore,

$$\mathrm{MISE}'(h) = \frac{1}{\pi n h^2} \frac{(n+1)|\varphi_{X^*}(1/h)|^2 - 1}{|\varphi_U(1/h)|^2}\, .$$

Since h_{MISE} satisfies $\mathrm{MISE}'(h)=0$, in the particular case of the sinc kernel, h_{MISE} is a solution of $|\varphi_{X^*}(1/h)|^2 = 1/(n+1)$. This equation uniquely identifies h under some conditions (Stefanski and Carroll, 1990) and is the same as the equation for finding the CV bandwidth given above, except that the unknown φ_{X^*} is replaced by its estimator $\hat{\varphi}_{X^*}$, and -2 is replaced here by -1 (this is to remove some of the bias of the estimator $|\hat{\varphi}_{X^*}|^2$ of $|\varphi_{X^*}|^2$).

1.6.2 Plug-in and normal reference bandwidths

When we can reasonably assume that the density f_X has at least k smooth derivatives, and we use a kth order kernel, with k even, instead of using the cross-validation bandwidth we can use instead the plug-in (PI) bandwidth (Delaigle and Gijbels, 2004b). As in the error-free case, this bandwidth is quite popular because it often works well in practice and is much faster to compute than the cross-validation bandwidth. We also discuss another bandwidth called the normal reference bandwidth, which is even simpler to compute but usually gives rather poor practical performance.

The plug-in bandwidth is obtained by plugging, in the AMISE expression, a kernel estimator of the unknown quantity, and then minimise the resulting estimator of the AMISE with respect to h. Specifically, recalling the AMISE expression from (1.12), the only unknown is $\theta_k \equiv \int\{f_X^{(k)}(x)\}^2\,dx$, which is estimated by $\hat\theta_k \equiv \int\{\hat f_X^{(k)}(x;h_k)\}^2\,dx$, where

$$\hat f_X^{(k)}(x;h_k) = \frac{1}{nh_k^{k+1}}\sum_{j=1}^n K_U^{(k)}\Big(\frac{x-X_j^*}{h_k}\Big) = \frac{1}{2\pi}\int(-it)^k e^{-itx}\,\frac{\hat\varphi_X^*(t)\varphi_K(h_kt)}{\varphi_U(t)}\,dt$$

is the kth derivative of $\hat f_X(x;h_k)$ at (1.4) computed with a bandwidth h_k (Delaigle and Gijbels, 2002) and $K_U^{(k)}(x) = (2\pi)^{-1}\int(-it)^k e^{-itx}\varphi_K(t)/\varphi_U(t/h)\,dt$ is the kth derivative of K_U. As mentioned earlier, the kernel K is almost always taken to be symmetric so that its order k is even; in that case, $(-i)^k = (-1)^{k/2}$.

Using Parseval's identity, $\hat\theta_k$ can be expressed as

$$\hat\theta_k = (2\pi)^{-1}\int t^{2k}|\hat\varphi_X^*(t)|^2|\varphi_K(h_kt)|^2|\varphi_U(t)|^{-2}\,dt\,,$$

which leads to the following estimator $\widehat{\text{AMISE}}(h)$ of AMISE(h):

$$\frac{\mu_{K,k}^2}{(k!)^2}\frac{h^{2k}}{2\pi h_k^{2k+1}}\int t^{2k}\,\frac{|\hat\varphi_X^*(t/h_k)|^2|\varphi_K(t)|^2}{|\varphi_U(t/h_k)|^2}\,dt + \frac{1}{2\pi nh}\int\frac{|\varphi_K(t)|^2}{|\varphi_U(t/h)|^2}\,dt\,.$$

Then, the bandwidth is chosen by minimising $\widehat{\text{AMISE}}(h)$ with respect to h. If the error density is unknown and estimated from data as in Chapter 12 (Delaigle and Van Keilegom, 2021), what often works well in practice is to replace φ_U in this expression by its estimator $\hat\varphi_U$.

To compute $\widehat{\text{AMISE}}(h)$, we need to choose the bandwidth h_k used to compute $\hat\theta_k$. Here, h_k should be chosen to estimate θ_k the best way possible. Delaigle and Gijbels (2002, 2004b) suggest choosing h_k to minimise (an estimator of) the mean squared error (MSE) of $\hat\theta_k$, and doing this typically results in h_k different from h. For example, it leads to $h_k \sim n^{-1/(2\beta+3k+1)}$ if U is ordinary smooth of order β, whereas we saw in Section 1.4.3 that the bandwidth h that minimises the MISE of $\hat f_X$ is of order $h \sim n^{-1/(2\beta+2k+1)}$. Thus, in this case case, h_k is an order of magnitude larger than h.

If the error is supersmooth of order β, Delaigle and Gijbels (2002) showed that if φ_K is supported on $[-B,B]$, the optimal convergence rate for estimating θ_k is obtained by taking $h = dB(2/\gamma)^{1/\beta}(\log n)^{-1/\beta}$ with $d>1$. This is also the bandwidth found in Section 1.4.3 for computing f_X at the best possible rate. Thus in theory, in the supermsooth error case we could take

$h_k = h = dB(2/\gamma)^{1/\beta}(\log n)^{-1/\beta}$ with $d > 1$. However, the value of d influences the practical success of the estimator and it is not clear how to choose d. Instead, as in the ordinary smooth error case, taking h and h_k to minimise an estimator of the AMISE of $\hat{f}_X$ and of the MSE of $\hat{\theta}_k$, respectively, often gives good results; see Delaigle and Gijbels (2004b).

The MSE of $\hat{\theta}_k$, used to choose h_k, also contains some unknowns that need to be estimated from the data. This is done via an iterative process; see Delaigle and Gijbels (2002, 2004b). The plug-in bandwidth often performs significantly better than the CV bandwidth, which tends to be too small. An exception to this is when the f_X has very sharp features, in which case its deconvolution kernel density estimator with the PI bandwidth tends to oversmooth those features, whereas the estimator computed with the CV bandwidth is often able to capture them better.

As in the error-free case, we can also compute a quick and dirty bandwidth called the normal reference bandwidth (Delaigle and Gijbels, 2004b). It consists in estimating θ_k by pretending that f_X is a normal density, that is, in estimating θ_k by $\hat{\theta}_k = (2k)!/\{(2\hat{\sigma}_X)^{2k+1}k!\sqrt{\pi}\}$, where $\hat{\sigma}_X^2 = \max\{1/n, \hat{\sigma}_{X^*}^2 - \text{var}(U)\}$ is an estimator of the variance of X, with $\hat{\sigma}_{X^*}^2$ the empirical variance of the X_i^*'s. The normal reference bandwidth is obtained by minimising the resulting estimator of AMISE(h). Since it is obtained under the assumption that f_X is normal, it is often too large and makes the estimator of f_X oversmoothed. Delaigle and Gijbels (2004b) also developed an alternative plug-in bandwidth called the solve-the-equation bandwidth but found it had little practical advantage compared to their main plug-in bandwidth introduced above.

In practice, since we usually do not know how smooth f_X is, it is common to take $k = 2$ and use a second order kernel. As in the error-free case, even if we had the information that f_X had more than 2 derivatives using a kernel of order $k > 2$ would produce more wiggly estimators, although it would capture sharp features better (taking k larger means taking h smaller and thus having a smaller bias but a larger variance). See also our discussion about the sinc kernel in Section 1.6.1. Note that the sinc kernel does not have a finite number k of moments and cannot be used with the plug-in or the normal reference bandwidths.

1.6.3 Bootstrap bandwidth

Delaigle and Gijbels (2004a) proposed a bootstrap bandwidth that directly attempts to minimise an estimator of the MISE. Recalling the expression at (1.14), and noting that $\int |\varphi_X(t)|^2\,dt$ does not depend on h, they propose to choose h by minimising

$$\text{MISE}_2^*(h) = \frac{1}{2\pi nh}\int \frac{|\varphi_K(t)|^2}{|\varphi_U(t/h)|^2}\,dt + \frac{(1-n^{-1})}{2\pi}\int |\hat{\varphi}_{X,\tilde{h}}(t)|^2|\varphi_K(ht)|^2\,dt$$

$$- \frac{1}{\pi} \int |\hat{\varphi}_{X,\tilde{h}}(t)|^2 \varphi_K(ht)\, dt\,,$$

where $\hat{\varphi}_{X,\tilde{h}}(t) = \hat{\varphi}_{X^*}(t)\varphi_K(\tilde{h}t)/\varphi_U(t)$ is the Fourier transform of the deconvolution kernel density estimator at (1.3), computed with a bandwidth $\tilde{h}$. Although Delaigle and Gijbels (2004a) originally motivated this approach by bootstrap ideas, it does not require to generate any bootstrap sample. The bandwidth $\tilde{h}$ is generally not taken to be equal to h, and in the case of a kernel of a finite order k, Delaigle and Gijbels (2004a) suggested choosing $\tilde{h}$ equal to the bandwidth h_k used to compute the plug-in bandwidth introduced in Section 1.6.2. In their numerical investigation, Delaigle and Gijbels (2004b) found that the bootstrap bandwidth was often outperformed by the plug-in bandwidth.

In the particular case of the sinc kernel, if we take $\tilde{h} \geq h$, we have

$$\text{MISE}_2^*(h) = \frac{1}{2\pi n} \int_{-1/h}^{1/h} \frac{1 - (n+1)|\hat{\varphi}_{X^*}(t)|^2}{|\varphi_U(t)|^2}\, dt\,,$$

which is an estimator of the first term of (1.20) (recall that the second term of (1.20) does not depend on h), where $|\varphi_{X^*}(t)|^2$ is estimated by $|\hat{\varphi}_{X^*}(t)|^2$. This is almost identical to the cross-validation criterion at (1.19), with 2 there replaced by 1 here; see our discussion about this in the last paragraph of Section 1.6.1.

1.6.4 SIMEX bandwidth

The procedures introduced above are all targeted at the estimation of the density f_X. Delaigle and Hall (2008) introduced a more general SIMEX (simulation extrapolation) procedure that can be applied in a wide class of deconvolution problems. Their procedure applies to the bandwidth selection context, the SIMEX procedure developed by Cook and Stefanski (1994) and Stefanski and Cook (1995). A preliminary version of the SIMEX bandwidth, that used only one level of simulated data, was proposed by Delaigle and Meister (2007) in a heteroscedastic errors-in-variables regression context. The two-level SIMEX described here was proposed by Delaigle and Hall (2008). In this section we assume that the error density f_U is known; see Delaigle and Hall (2008) for a variant of the SIMEX bandwidth in the case where f_U is unknown but replicated data are available.

The main idea of SIMEX is as follows: suppose we want to estimate a curve, g say, that depends on a variable X that we cannot observe directly. For example, in the density estimation problem, $g = f_X$ and in the regression estimation problem, $g(x) = E(Y|X = x)$, where Y is a dependent variable. Instead of observing a sample $X_1, \ldots, X_n \sim f_X$, we observe $X_1^*, \ldots, X_n^*$ coming from the classical measurement error model at (1.1). Let $\hat{g}$ denote a nonparametric estimator of g based on the contaminated X_i^*, which requires the choice of one or several smoothing parameters, and let $\hat{g}_{EF}$ denote the a nonparametric estimator of g that we would use if we could observe the X_i's (here EF stands for error-free). For example, in the density estimation case, $\hat{g}$ could be the deconvolution kernel density estimator $\hat{f}_X$ introduced earlier, which requires the choice of a bandwidth h, and $\hat{g}_{EF}$ could be the standard error-free kernel density estimator computed from the X_i's. In the errors-in-variables regression case, $\hat{g}$ could be the local constant errors-in-variables estimator of Fan and Truong (1993) computed from a sample of (X_i^*, Y_i)'s distributed as (X^*, Y), or the more general local polynomial errors-in-variables estimator of Delaigle et al. (2009), both of which require the choice of a bandwidth h and, optionally, of a ridge parameter ρ; $\hat{g}_{EF}$ could be the standard error-free local polynomial estimator of g computed from a sample of (X_i, Y_i)'s distributed as (X, Y).

Let H denote the smoothing parameters required to compute $\hat{g} \equiv \hat{g}(\cdot; H)$. For example, in the density estimation problem, $H = h$, the bandwidth. In the regression estimation problem, $H = (h, \rho)$. If we were able to compute $\hat{g}$ and $\hat{g}_{EF}$, an approach we could take to choose H would be to choose

$$\hat{H} = \text{argmin}_H D\{\hat{g}(\cdot; , H), \hat{g}_{EF}\}\,, \tag{1.21}$$

where D denotes a distance, for example, a weighted integrated squared error

$$D(\hat{g}, \hat{g}_{EF}) = \int \{\hat{g}(x; H) - \hat{g}_{EF}(x)\}^2 w(x)\,,$$

with w a weight function of our choice. In general, such a procedure is justified because the convergence rate of $\hat{g}_{EF}$ to g is faster than that of $\hat{g}$ to g, so that choosing H that minimises the distance above is asymptotically equivalent to choosing H that minimises $D\{\hat{g}(\cdot; , H), g\}$. In the density case, we can take $w(x) = 1$; in the regression case, since nonparametric regression estimators are not able to estimate a regression curve well at the tails of the distribution of the explanatory variable, we can take $w(x) = 1_{[a,b]}(x)$ with a and b some finite values, for example, a lower and an upper quantile of X.

Of course, we cannot compute (1.21) in practice since we do not observe the X_i's and so we cannot compute $\hat{g}_{EF}$. The idea of SIMEX is to simulate artificial data which mimic the data we cannot observe, and use them to approximate the bandwidth $\hat{H}$ at (1.21). To do this, first we simulate data $X_1^{**}, \ldots, X_n^{**}$ and $X_1^{***}, \ldots, X_n^{***}$ from the following two error models:

$$X_i^{**} = X_i^* + U_i^*\ ,\ X_i^{***} = X_i^{**} + U_i^{**}\,, \tag{1.22}$$

where $U_i^* \sim f_U$ and $U_i^{**} \sim f_U$ are independent. Let g^* and g^{**} denote the versions of g with X replaced by X^* and X_i^{**}, respectively. For example, in the density case, $g^* = f_{X^*}$ and $g^{**} = f_{X^{**}}$ and in the regression case, $g^*(x) = E(Y|X^* = x)$ and $g^{**}(x) = E(Y|X^{**} = x)$. Since we observe the X_i^*'s and we generate the U_i^*'s and the U_i^{**}'s, then in the two error models at (1.22), we observe the contaminated variables X_i^{**} and X_i^{***}, as well as their non contaminated version X_i^* and X_i^{**}, respectively. Therefore, we can compute the deconvolution estimators $\hat{g}^*$ and $\hat{g}^{**}$ of g^* and g^{**}, based on the X_i^{**}'s and the X_i^{***}'s, respectively, but we can also compute the standard estimators $\hat{g}^*_{EF}$ and $\hat{g}^{**}_{EF}$ of g^* and g^{**}, based on the X_i^*'s and X_i^{**}'s, respectively. Therefore, we are able to compute the versions of (1.21) that correspond to those more contaminated data. That is, we can compute

$$\hat{H}^* = \text{argmin}_H D\{\hat{g}^*(\cdot; , H), \hat{g}^*_{EF}\} \ , \ \hat{H}^{**} = \text{argmin}_H D\{\hat{g}^{**}(\cdot; , H), \hat{g}^{**}_{EF}\} \, .$$

Of course, such $\hat{H}^*$ and $\hat{H}^{**}$ depend heavily on the particular X_i^{**}'s and X_i^{***}'s generated at (1.22). To reduce this dependence, we simulate repeated, say B, samples $X^{**}_{b,1}, \ldots, X^{**}_{b,n}$ and $X^{***}_{b,1}, \ldots, X^{***}_{b,n}$ in the same way as at (1.22), for $b = 1, \ldots, B$, and take

$$\hat{H}^* = \text{argmin}_H B^{-1} \sum_{b=1}^{B} D\{\hat{g}^*_b(\cdot; , H), \hat{g}^*_{EF}\} \, ,$$

$$\hat{H}^{**} = \text{argmin}_H B^{-1} \sum_{b=1}^{B} D\{\hat{g}^{**}_b(\cdot; , H), \hat{g}^{**}_{b,EF}\} \, ,$$

where we used the index b to indicate that an estimator was computed from the bth sample; $\hat{g}^*_{EF}$ does not have an index b since only one sample from X^* is available.

Next, since X^{***} (resp., X^{**}) measures X^{**} (resp., X^*) in the same way as X^* measures X, it is reasonable to think that when H is a single parameter (typically a bandwidth), the relationship between $\hat{H}^{**}$ and $\hat{H}^*$ approximates reasonably well the relationship between $\hat{H}^*$ and $\hat{H}$. This suggests that $\hat{H}/\hat{H}^* \approx \hat{H}^*/\hat{H}^{**}$, which suggests approximating $\hat{H}$ by

$$\hat{H} \approx (\hat{H}^*)^2/\hat{H}^{**} \, .$$

In the case where H is multivariate, the approximation needs to be done with more thought as the relationship between the various components of H may be important. For example, in the regression case where $H = (h, \rho)$, the values of h and ρ influence each other; see Delaigle and Hall (2008).

In general, since the errors used in the models at (1.22) have the same distribution as the U_i's, then when $H = h$ is a bandwidth, the SIMEX bandwidth $\hat{h}$ converges to zero at the same speed as the optimal bandwidth corresponding to the distance used at (1.21). For example, for the deconvolution kernel density estimator, in the ordinary smooth error case, if $h_{\text{MISE}} \sim c_1 n^{-1/(2\beta+2k+1)}$, then $\hat{h} \sim c_2 n^{-1/(2\beta+2k+1)}$, but in general $c_2 \neq c_1$. That is, the SIMEX bandwidth is not a consistent estimator of the optimal bandwidth. However, in general too, c_2 is good approximation of c_1. This is contrast with the bandwidths introduced in the previous sections, which were consistent estimators of an optimal bandwidth. In cases where it is possible to compute a consistent estimator of the optimal bandwidth, it is often preferable to use such a consistent estimator, and dedicate SIMEX to more complex problems where such a bandwidth cannot be computed easily enough.

1.6.5 Further reading

In the case where the error density f_U is unknown and estimated from direct data generated from f_U, and K is the sinc kernel, Diggle and Hall (1993) and Barry and Diggle (1995) proposed regression-based practical bandwidths.

Achilleos and Delaigle (2012) developed a local Empirical Bias Bandwidth Selector (EBBS) and a local integrated plug-in approach, both based on estimating the asymptotically dominating part of the MSE of $\hat{f}_X$ (the AMSE). They also proposed a local SIMEX approach based on the MSE. They showed that their plug-in and EBBS methods give good practical results and perform considerably better than global bandwidths in cases where f_X has sharp features, without degrading much the quality of estimators (compared to using a global bandwidth) when f_X is very smooth. R codes for those bandwidths are available on Delaigle's webpage.

1.7 Generalisations

Over the last three decades, a lot of effort has been dedicated to various problems related to nonparametric kernel density deconvolution, so much that it is not possible to present them all here, and we discuss only a few extensions. In Section 1.7.1, we show that in some cases, it is possible to perform the deconvolution explicitly without having to go through the Fourier domain. In Section 1.7.2, we discuss the case where the characteristic function of the errors vanishes and in Section 1.7.3 we discuss that where the errors are not identically distributed. In Section 1.7.4 we discuss multivariate extensions. We conclude with Section 1.7.5, where we briefly summarise some other related work developed in the literature.

1.7.1 Settings with analytic inversion formulae

One of the difficulties of estimating a density from data measured with classical errors is that it involves a complex inversion process. Often, this deconvolution process can only be written analytically in the Fourier domain, as reflected by the Fourier approach taken by the deconvolution kernel density estimator. However, in some cases, it is possible to express f_X analytically in terms of f_{X^*} and thus to deconvolve without going through Fourier transforms.

Van Es and Kok (1998) considered two such particular cases. The first is when $U = \lambda_1 E_1 + \ldots + \lambda_m E_m$, where the E_j's are independent standard exponential random variables, m is a known positive integer and the λ_j's are known constants; the second is when $U = \lambda_1 L_1 + \ldots + \lambda_m L_m$, where the L_j's are independent standard Laplace random variables and the λ_j's are known constants. They showed that in the exponential case we can write

$F_X(x) = F_{X^*}(x) + \sum_{j=1}^m e_j f_{X^*}^{(j-1)}(x)$ and $f_X(x) = f_{X^*}(x) + \sum_{j=1}^m e_j f_{X^*}^{(j)}(x)$, where $e_j = \sum_{1 \le i_1 < \cdots < i_j \le m} \lambda_{i_1} \cdots \lambda_{i_j}$; in the Laplace case we have $F_X(x) = F_{X^*}(x) + \sum_{j=1}^m (-1)^j \ell_j f_{X^*}^{(2j-1)}(x)$ and $f_X(x) = f_{X^*}(x) + \sum_{j=1}^m (-1)^j \ell_j f_{X^*}^{(2j)}(x)$, where $\ell_j = \sum_{1 \le i_1 < \cdots < i_j \le m} \lambda_{i_1}^2 \cdots \lambda_{i_j}^2$. Therefore, in the exponential we can estimate $F_X(x)$ and $f_X(x)$ by

$$\hat{F}_X(x) = \hat{F}_{X^*}(x) + \sum_{j=1}^m e_j \hat{f}_{X^*}^{(j-1)}(x) \ , \ \hat{f}_X(x) = \hat{f}_{X^*}(x) + \sum_{j=1}^m e_j \hat{f}_{X^*}^{(j)}(x) \, ,$$

and in the Laplace case we can take

$$\hat{F}_X(x) = \hat{F}_{X^*}(x) + \sum_{j=1}^m (-1)^j \ell_j \hat{f}_{X^*}^{(2j-1)}(x), \hat{f}_X(x) = \hat{f}_{X^*}(x) + \sum_{j=1}^m (-1)^j \ell_j \hat{f}_{X^*}^{(2j)}(x),$$

where $\hat{F}_{X^*}$ is a standard error-free kernel estimator of F_{X^*} and $\hat{f}_{X^*}^{(j)}$ is the jth derivative of the standard kernel density estimator of f_{X^*}, all constructed from the X_i^*'s. In the Laplace case, the estimator is identical to the deconvolution kernel density estimator (see our example at (1.16) in Section 1.5.3). However, in the exponential case, the deconvolution kernel density estimator does not exist because φ_U has some zeros; see Section 1.7.2.

Another particular case was considered by Groeneboom and Jongbloed (2003), who studied the case where U follows a uniform distribution on $[0, 1]$ and f_X is supported on $[0, \infty)$ (the extension to a uniform $[0, b]$ error distribution is straightforward and obtained by simple rescaling). In that case,

$$f_{X^*}(x) = F_X(x) - F_X(x-1) \, , \tag{1.23}$$

a relationship that can be exploited to construct a nonparametric maximum likelihood estimator (NPMLE) $\hat{F}_X$ of F_X. This estimator is uniquely defined; it can only assign masses to points in the set $\{X_i^*, \ldots, X_n^*\}$. From there, a continuous estimator of $f_X(x)$ can be obtained by taking $\hat{f}_X(x) = \int K_h(x - y)\, d\hat{F}_X(y)\, dy$. As usual with kernel estimators, if f_X is not continuous at 0, then $\hat{f}_X$ suffers from boundary problems which can be corrected by standard boundary correction techniques.

These authors discussed an alternative kernel smoothing approach for estimating f_X. Its construction is based on a recursive application of the relationship $F_X(x) = F_X(x-1) + f_{X^*}(x)$, which follows from (1.23). Together with the fact that for all $x \in \mathbb{R}$, $F_X(x - j) = 0$ when $j > x$, it leads to

$$F_X(x) = \sum_{j=0}^\infty f_{X^*}(x-j) \text{ and } f_X(x) = \sum_{j=0}^\infty f'_{X^*}(x-j) \, , \tag{1.24}$$

where the second equality is obtained by differentiation of the first. This suggests estimating $f_X(x)$ by $\hat{f}_X(x) = \sum_{j=0}^\infty \hat{f}'_{X^*}(x-j)$, where $\hat{f}'_{X^*}$ denotes the first derivative of the standard kernel density estimator of f_{X^*} (or a boundary corrected version if f_X is not continuous as 0). Here too we cannot apply the deconvolution kernel density estimator since we have $\varphi_U(t) = (e^{it} - 1)/(it)$, and thus $\varphi_U(t) = 0$ for $t = 2k\pi$, for all $k \in \mathbb{Z}_0$; see Section 1.7.2.

Delaigle and Meister (2011) considered a more general setting than the uniform case. Assuming that the distribution of X has a finite left endpoint a, they considered error distributions whose characteristic function φ_U has periodic isolated zeros, namely, where zeros of φ_U can only be of the form $t = k\lambda$ for a fixed λ, where $k \in \mathbb{Z}_0$. For example, the ordinary smooth error case can be generalised by allowing φ_U to satisfy

$$|\varphi_U(t)| \geq d_0 |\sin(t\pi/\lambda)|^{\nu} |t|^{-\beta} \quad \text{for large } |t| \tag{1.25}$$

and the supersmooth error case can be generalised by allowing φ_U to satisfy

$$|\varphi_U(t)| \geq d_0 |\sin(t\pi/\lambda)|^{\nu} |t|^{\beta_0} \exp(-|t|^{\beta}/\gamma) \quad \text{for large } |t| \,. \tag{1.26}$$

This includes symmetric uniform distributions, their self convolutions, and their convolution with ordinary smooth or supersmooth distributions.

For such error distributions, we cannot estimate $\varphi_X(t)$ by $\hat{\varphi}_{X^*}(t)/\varphi_U(t)$ for all t since φ_U has zeros. To overcome this difficulty, instead of directly estimating $\varphi_X(t)$, Delaigle and Meister (2011) considered estimating first

$$\varphi_p(t) \equiv \big\{ \exp(2it\pi/\lambda) - 1 \big\}^{\nu} \varphi_X(t) \,.$$

The term $\{\exp(2it\pi/\lambda) - 1\}^{\nu}$ is introduced to compensate for the term $|\sin(t\pi/\lambda)|^{\nu}$ term at (1.25) and (1.26). It is such that $\lim_{t\to k\lambda} |\exp(2it\pi/\lambda) - 1|^{\nu} |\sin(t\pi/\lambda)|^{-\nu}$ is finite, so that dividing by φ_U no longer causes difficulties. In particular we can estimate $\varphi_p(t)$ by

$$\hat{\varphi}_p(t) = \begin{cases} \big\{ \exp(2it\pi/\lambda) - 1 \big\}^{\nu} \hat{\varphi}_{X^*}(t)/\varphi_U(t) & \text{if } t \neq k\lambda, k \in \mathbb{Z}_0 \\ c_k & \text{if } t = k\lambda, k \in \mathbb{Z}_0 \,, \end{cases}$$

where $c_k = \lim_{t\to k\lambda} \hat{\varphi}_p(t)$ if it exists and 0 otherwise.

To deduce an estimator of f_X, let $p(x) = (2\pi)^{-1} \int e^{itx} \varphi_p(t)\, dt$. If f_X is supported on $[a, \infty)$ with a finite, then for all $J \geq \lambda(x-a)/(2\pi)$ we can write $f_X(x) = \sum_{k=0}^{J} \eta_k \, p(x - 2k\pi/\lambda)$, where $(\eta_0, \ldots, \eta_J) = (1, 0, \ldots, 0)\mathbf{\Gamma}^{-1}$ and $\mathbf{\Gamma}$ is an upper triangular matrix whose component (j, k), above the diagonal, is equal to $\binom{\nu}{k-j}(-1)^{\nu-k+j}$. Letting $\hat{p}(x) = (2\pi)^{-1} \int e^{itx} \varphi_K(ht) \hat{\varphi}_p(t)\, dt$ with K a kernel and h a bandwidth, and taking $J = J(x) = \lceil \lambda(x-a)/(2\pi) \rceil$, or taking $J = \lceil \lambda(b-a)/(2\pi) \rceil$ if f_X is supported on $[a, b]$ with b finite, we can estimate $f_X(x)$ by

$$\hat{f}_X(x) = \sum_{k=0}^{J} \eta_k \, \hat{p}(x - 2k\pi/\lambda) \cdot I\{a \leq x < \infty\} \,.$$

Delaigle and Meister (2011) showed that this estimator can be expressed in a form very similar to the deconvolution kernel density estimator.

1.7.2 When the Fourier transform vanishes

As we discussed in Section 1.7.1, since it involves dividing by $\varphi_U(t)$, the deconvolution kernel density estimator cannot be used when $\varphi_U(t)$ vanishes at some t. Since compactly supported distributions have a characteristic function that vanishes at some points, there are many important cases where this estimator cannot be used. In Section 1.7.1 we discussed simple alternative procedures that can be used in particular cases where f_X can be expressed analytically in terms of f_{X^*} without the need to work in the Fourier domain. In this section we discuss other procedures that can be used even when no such analytic expressions are available.

Devroye (1989) was one of the first to consider this problem. He showed that if φ_U vanishes on a set of measure zero, we can still construct a consistent estimator of f_X. He proposed to modify the deconvolution kernel density estimator by replacing $\varphi_U^{-1}(t)$ at (1.3) by $\varphi_U^{-1}(t) \cdot I\{|\varphi_U(t)| \leq \rho\}$, for some small parameter $\rho > 0$. In other words, he proposed to remove, from the integration domain, the regions where $|\varphi_U|$ is too small. Instead of removing these regions from the integration domain, Meister (2007) proposed to replace $\hat{\varphi}_{X^*}(t)/\varphi_U(t)$ in those regions by using a polynomial approximation obtained from nearby points.

In the same context where φ_U has isolated zeros, Hall and Meister (2007) proposed to use a ridge procedure, where they replace φ_U by a ridge function when $|\varphi_U|$ gets too small. Using the fact that $\varphi_{X*}(t) = \varphi_X(t)\varphi_U(t)$ can be expressed as $\varphi_{X*}(t)\varphi_U(-t) = \varphi_X(t)|\varphi_U(t)|^2$, we could estimate $\varphi_X(t)$ by

$$\hat{\varphi}_X(t) = \frac{\hat{\varphi}_{X*}(t)\varphi_U(-t)}{\max\{|\varphi_U(t)|, \rho(t)\}^2}\,,$$

where $\rho(t) > 0$ is a ridge function that prevents the denominator from getting too close to zero. The advantage of multiplying the numerator and the denominator by $\varphi_U(-t)$ before ridging is that unlike $\varphi_U(t)$, $|\varphi_U(t)|$ is guaranteed to be nonnegative, so that $\rho(t)$ can be restricted to be positive.

More generally, they write $\varphi_{X*}(t)\varphi_U(-t)|\varphi_U(t)|^r = \varphi_X(t)|\varphi_U(t)|^{r+2}$ for some $r \geq 0$ and propose to estimate $\varphi_X(t)$ by

$$\hat{\varphi}_X(t) = \frac{\hat{\varphi}_{X*}(t)\varphi_U(-t)|\varphi_U(t)|^r}{\max\{|\varphi_U(t)|, \rho(t)\}^{r+2}}\,.$$

Then assuming that $|\varphi_U(t)|^{r+1}$ is integrable, they estimate $f_X(x)$ by

$$\hat{f}_X(x) = \frac{1}{2\pi}\int e^{-itx}\frac{\hat{\varphi}_{X*}(t)\varphi_U(-t)|\varphi_U(t)|^r}{\max\{|\varphi_U(t)|, \rho(t)\}^{r+2}}\,dt\,.$$

Since the integral exists for all x, this approach does not require a kernel function, although Hall and Meister (2007) showed that the estimator can actually be expressed in a kernel form. See Hall and Meister (2007) for how to choose the ridge function $\rho(t)$ in practice using cross-validation. They studied the convergence rates of this estimator under various settings and showed that in some cases, those rates are the same as when φ_U does not vanish anywhere (i.e., the same as those from Section 1.4). However, for errors as at (1.25), those rates are slower whereas the method proposed by Delaigle and Meister (2011) (see Section 1.7.1) has the same rates as when φ_U does not vanish anywhere.

Meister and Neumann (2010) considered the case where φ_U has some zeros and some replicated contaminated measurements are observed. Exploiting the replicates, they were able to propose yet another procedure.

1.7.3 Heteroscedatic errors

Delaigle and Meister (2008) considered the density estimation problem in cases where the errors are not identically distributed. There, the observations $X_1^*, \ldots, X_n^*$ are independent and for $i = 1, \ldots, n$, $X_i^* = X_i + U_i$ where X_i and U_i are independent, $X_i \sim f_X$ for all i, but $U_i \sim f_{U_i}$, where the f_{U_i}'s are not necessarily identical. In that case, for $j = 1, \ldots, n$ we have $\varphi_{X_j^*}(t) = \varphi_X(t)\varphi_{U_j}(t)$ so that $\sum_{j=1}^n \varphi_{X_j^*}(t) = \varphi_X(t)\sum_{j=1}^n \varphi_{U_j}(t)$. Using Fourier inversion, if $\sum_{j=1}^n \varphi_{U_j}$ never vanishes we have

$$f_X(x) = \frac{1}{2\pi}\int e^{-itx}\frac{\sum_{j=1}^n \varphi_{X_j^*}(t)}{\sum_{j=1}^n \varphi_{U_j}(t)}\,dt\,.$$

Mimicking the deconvolution kernel estimation procedure at (1.3), this suggests that we could estimate $f_X(x)$ by

$$\hat{f}_X(x) = \frac{1}{2\pi}\int e^{-itx}\frac{\hat{\varphi}_{X^*}(t)\varphi_K(ht)}{n^{-1}\sum_{j=1}^n \varphi_{U_j}(t)}\,dt\,,$$

where $\hat{\varphi}_{X^*}(t) = n^{-1}\sum_{j=1}^n e^{itX_j^*}$ and with K and h as in Section 1.3.

While this approach is simple, Delaigle and Meister (2008) showed that faster convergence rates can be obtained by exploiting the relationship $\varphi_{X_j^*}(t)\varphi_{U_j}(-t) = \varphi_X(t)|\varphi_{U_j}(t)|^2$. Then if $\sum_{j=1}^n |\varphi_{U_j}(t)|^2 \neq 0$ for all $t \in \mathbb{R}$,

$$f_X(x) = \frac{1}{2\pi}\int e^{-itx}\frac{\sum_{j=1}^n \varphi_{X_j^*}(t)\varphi_{U_j}(-t)}{\sum_{j=1}^n |\varphi_{U_j}(t)|^2}\,dt\,,$$

which suggests estimating $f_X(x)$ by

$$\hat{f}_X(x) = \frac{1}{2\pi}\int e^{-itx}\frac{\sum_{j=1}^n e^{itX_j^*}\varphi_{U_j}(-t)}{\sum_{j=1}^n |\varphi_{U_j}(t)|^2}\varphi_K(ht)\,dt\,.$$

They generalised their procedure to the case where the error densities are unknown but the noisy observations are replicated; see Section 12.2 in Delaigle and Van Keilegom (2021). McIntyre and Stefanski (2011) also considered this problem in the case where the U_{ij}'s are normally distributed. Hesse (1995, 1996) considered a related setting where only a fraction of the observations are contaminated by errors and the others are observed without noise.

1.7.4 Multivariate and dependent cases

In Masry (1991), the observations are a univariate sample $X_1^*, \ldots, X_n^*$ from the classical error model at (1.1), but the X_i's and the U_i's are allowed the have a dependence structure, whereas the X_i's remain independent of the U_i's. For $\mathbf{T} = \mathbf{X}$ and $T_j = X_j$, $\mathbf{T} = \mathbf{X}^*$ and $T_j = X_j^*$, and $\mathbf{T} = \mathbf{U}$ and $T_j = U_j$, with $j = 1, \ldots, p$, let $f_{\mathbf{T}}$ and $\varphi_{\mathbf{T}}$ denote, respectively, the density and the characteristic function of $\mathbf{T} = (T_1, \ldots, T_p)$, where $p < n$. For simplicity we omit the explicit dependence on p and all vectors in this section are assumed to be of size p. Assuming that the process $\{X_i\}_{i=-\infty}^{\infty}$ is stationary, the goal in Masry (1991) is to estimate $f_{\mathbf{X}}$ from the univariate data $X_1^*, \ldots, X_n^*$.

Masry (1991) generalised the deconvolution kernel estimator to this p-variate case, as follows. Since $\varphi_{\mathbf{X}^*}(\mathbf{t}) = \varphi_{\mathbf{X}}(\mathbf{t})\varphi_{\mathbf{U}}(\mathbf{t})$, if $\varphi_{\mathbf{U}}(\mathbf{t}) \neq 0$ for all $\mathbf{t}$ and all integrals below are well defined, by Fourier inversion we have

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^p} \int e^{-i\mathbf{t}\cdot\mathbf{x}} \varphi_{\mathbf{X}^*}(\mathbf{t})/\varphi_{\mathbf{U}}(\mathbf{t})\, d\mathbf{t}\,,$$

where $\mathbf{t} \cdot \mathbf{x} = \sum_{j=1}^p t_j x_j$. This suggests extending the deconvolution kernel density estimator to the p-variate case by taking

$$\hat{f}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^p} \int e^{-i\mathbf{t}\cdot\mathbf{x}} \varphi_{\mathbf{K}}(\mathbf{t}h) \frac{\hat{\varphi}_{\mathbf{X}^*}(\mathbf{t})}{\varphi_{\mathbf{U}}(\mathbf{t})}\, d\mathbf{t}\,,$$

where $\mathbf{K}$ is a p-variate kernel function, $h > 0$ is a bandwidth, and $\hat{\varphi}_{\mathbf{X}^*}(\mathbf{t}) = (n-p+1)^{-1} \sum_{j=1}^{n-p+1} e^{i\mathbf{t}\mathbf{X}_j^*}$, with $\mathbf{X}_j^* = (X_j^*, \ldots, X_{j+p-1}^*)$. Letting $\mathbf{K}_{\mathbf{U}}(\mathbf{x}) = (2\pi)^{-p} \int e^{-i\mathbf{t}\cdot\mathbf{x}} \varphi_{\mathbf{K}}(\mathbf{t})/\varphi_{\mathbf{U}}(\mathbf{t}/h)\, d\mathbf{t}$, we can express this estimator as

$$\hat{f}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(n-p+1)h^p} \sum_{j=1}^{n-p+1} \mathbf{K}_{\mathbf{U}}\Big(\frac{\mathbf{x} - \mathbf{X}_j^*}{h}\Big)\,.$$

This definition uses a single bandwidth h, but as in the error-free case, we could also define a more general version with a p-variate bandwidth matrix.

Masry (1991) studied L_2 properties of this estimator and derived a number of technical results that are useful for more general deconvolution problems. Asymptotic normality, strong consistency and further properties were derived in Masry (1993a,b, 2003) in the particular case where the U_k's are i.i.d. In that case, if φ_U denotes the common characteristic function of the U_k's, and we take $\varphi_{\mathbf{K}}(\mathbf{t}) = \prod_{k=1}^p \varphi_K(t_k)$ where K is a univariate kernel, we can write

$$\hat{f}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(n-p+1)h^p} \sum_{j=1}^{n-p+1} \prod_{k=0}^{p-1} K_U\Big(\frac{x_j - X_{j+k}^*}{h}\Big)\,,$$

with K_U at (1.5). In a related work where $p = 1$, Kulik (2008) studied properties of the deconvolution kernel density and distribution estimators in the case where the X_i's are dependent and the errors are ordinary smooth, and showed that the order of the bandwidth and central limit theorems can be influenced by the strength of the dependence.

Youndjé and Wells (2008) considered a related multivariate density estimation problem. There, the observations are i.i.d. p-vectors $\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*$, where for $i = 1, \ldots, n$, $\mathbf{X}_i^* = \mathbf{X}_i + \mathbf{U}_i$ and the p-variate $\mathbf{X}_i$'s are independent of the p-variate $\mathbf{U}_i$'s. This setting is different from the one above since here, for $\mathbf{T}_i = \mathbf{X}_i$ and $T_{ij} = X_{ij}$, $\mathbf{T}_i = \mathbf{X}_i^*$ and $T_{ij} = X_{ij}^*$, and $\mathbf{T}_i = \mathbf{U}_i$ and $T_{ij} = U_{ij}$, with $j = 1, \ldots, p$, the $\mathbf{T}_i$'s are defined by $\mathbf{T}_i = (T_{i1}, \ldots, T_{ip})$, so that the $\mathbf{T}_i$'s are independent; for $i = 1, \ldots, n$ and $\mathbf{T} = \mathbf{X}$, $\mathbf{X}^*$ and $\mathbf{U}$, we use $f_{\mathbf{T}}$ and $\varphi_{\mathbf{T}}$ to denote, respectively, the density and the characteristic function of $\mathbf{T}_i$. An estimator of $f_{\mathbf{X}}$ can essentially be constructed as above but with a different estimator of $\varphi_{\mathbf{X}^*}$. Specifically, we take $\hat{\varphi}_{\mathbf{X}^*}(\mathbf{t}) = n^{-1} \sum_{j=1}^n e^{i\mathbf{t}\mathbf{X}_j^*}$ and estimate $f_{\mathbf{X}}(\mathbf{x})$ by

$$\hat{f}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^p} \int e^{-i\mathbf{t}\cdot\mathbf{x}} \varphi_{\mathbf{K}}(\mathbf{t}h) \hat{\varphi}_{\mathbf{X}^*}(\mathbf{t}) / \varphi_{\mathbf{U}}(\mathbf{t})\, d\mathbf{t} = \frac{1}{nh^p} \sum_{j=1}^n \mathbf{K}_{\mathbf{U}}\Big(\frac{\mathbf{x} - \mathbf{X}_j^*}{h}\Big),$$

with $\mathbf{K}_{\mathbf{U}}$ as above. In the particular case where the U_{ij}'s are i.i.d. and we take $\varphi_{\mathbf{K}}(\mathbf{t}) = \prod_{j=1}^p \varphi_K(t_j)$, this estimator simplifies into

$$\hat{f}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{nh^p} \sum_{j=1}^n \prod_{k=1}^p K_U\Big(\frac{x_k - X_{jk}^*}{h}\Big),$$

with K_U at (1.5).

1.7.5 Further reading

Hall and Lahiri (2008) considered the estimation of a cumulative distribution function, its moments and its quantiles, when the contaminated data come from the classical error model. Minimax properties of the distribution estimation problem were studied by Dattner et al. (2011) in the ordinary smooth error case; the quantile estimation problem was also considered by Dattner et al. (2016).

Rachdi and Sabre (2000) considered the problem of mode estimation of a density in the nonparametric deconvolution problem. Zhang and Karunamuni (2000) and Zhang and Karunamuni (2009) considered deconvolution kernel density estimation in the case where f_X has a compact support, and Hall and Simar (2002), Goldenshluger and Tsybakov (2004), Delaigle and Gijbels (2006a,b), Meister (2006), Aarts et al. (2007) and Kneip et al. (2015) proposed ways to estimate the boundary of this support.

Holzmann et al. (2007) and Butucea et al. (2009) considered general tests of hypothesis on the density f_X; Meister (2009b) considered tests of local monotonicity and Carroll et al. (2011) considered estimating and testing shape-constrained nonparametric density and regression with measurement errors. Their tilting technique consists in assigning weights p_i to each X_i^*, such that

$p_i \geq 0$ and $\sum_{i=1}^n p_i = 1$ (whereas the standard deconvolution kernel estimator assigns the same weight $1/n$ to each observation), where the p_i's are chosen so that the density estimator satisfies the desired shape constraint. In a related work, Hazelton and Turlach (2009) proposed a weighted kernel density estimator $\hat{f}_X$ which assigns positive weights p_i to each observation, but uses a standard kernel K instead of the deconvolution kernel K_U. To choose the weights, they proposed to minimise the L_2 distance between the standard kernel density estimator of f_{X^*} computed from the X_i^*'s and $\hat{f}_{X^*} = \hat{f}_X * f_U$. They also proposed a multivariate version of their procedure.

Hazelton and Turlach (2010) considered a semiparametric kernel estimator that avoids the use of a deconvolution technique by incorporating some parametric information in the construction; they extended their technique to the multivariate case. Delaigle and Hall (2014) considered a parametrically assisted version of the deconvolution kernel density estimator, where some a priori parametric information is included in the estimation technique. Potgieter (2020) considered a related deconvolution problem in the case where we can reasonably assume that X has a generalised skew-symmetric distribution.

Some related work was also developed for other variants of the classical measurement error model. This includes Delaigle (2007), who considered kernel density estimation of a density contaminated by classical and Berkson errors, and Camirand Lemyre et al. (ress), who considered the kernel deconvolution problem with excess zeros. There, the interest is in a continuous random variable, typically representing the long term intake X of a nutrient, and the measured variables are either equal to a contaminated version X^* of X, if the measurement was taken on a consumption day, or to zero if the observed individual did not eat the nutrient on the day where the measurement was taken.

Acknowledgment

Delaigle acknowledges support from the Australia Research Council (DP170102434).

Bibliography

Aarts, L., P. Groeneboom, and G. Jongbloed (2007). Estimating the upper support point in deconvolution. *Scandinavian journal of statistics 34*(3), 552–568.

Achilleos, A. and A. Delaigle (2012). Local bandwidth selectors for deconvolution kernel density estimation. *Statistics and Computing 22*(2), 563–577.

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 14, pp. ? CRC.

Barry, J. and P. Diggle (1995). Choosing the smoothing parameter in a fourier approach to nonparametric deconvolution of a density estimate. *Journal of Nonparametric Statistics 4*(3), 223–232.

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, and Applications*. CRC press.

Butucea, C. (2004a). Asymptotic normality of the integrated square error of a density estimator in the convolution model. *SORT: Statistics and Operations Research Transactions 28*(1), 9–26.

Butucea, C. (2004b). Deconvolution of supersmooth densities with smooth noise. *Canadian Journal of Statistics 32*(2), 181–192.

Butucea, C., C. Matias, and C. Pouet (2009). Adaptive goodness-of-fit testing from indirect observations. In *Annales de l'IHP Probabilités et statistiques*, Volume 45, pp. 352–372.

Butucea, C. and A. B. Tsybakov (2008a). Sharp optimality in density deconvolution with dominating bias. i. *Theory of Probability & Its Applications 52*(1), 24–39.

Butucea, C. and A. B. Tsybakov (2008b). Sharp optimality in density deconvolution with dominating bias. ii. *Theory of Probability & Its Applications 52*(1), 237–249.

Camirand Lemyre, F., R. J. Carroll, and A. Delaigle (in press). Semiparametric estimation of the distribution of episodically consumed foods measured with error. *Journal of the American Statistical Association 0*(0), 1–13.

Carroll, R. J., A. Delaigle, and P. Hall (2011). Testing and estimating shape-constrained nonparametric density and regression in the presence of measurement error. *Journal of the American Statistical Association 106*(493), 191–202.

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.

Carroll, R. J., D. Ruppert, L. Stefanski, and C. Crainiceanu (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. CRC press.

Cook, J. R. and L. A. Stefanski (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical association 89*(428), 1314–1328.

Dattner, I., A. Goldenshluger, and A. Juditsky (2011). On deconvolution of distribution functions. *The Annals of Statistics*, 2477–2501.

Dattner, I., M. Reiß, and M. Trabs (2016). Adaptive quantile estimation in deconvolution with unknown error distribution. *Bernoulli 22*(1), 143–192.

Delaigle, A. (2007). Nonparametric density estimation from data with a mixture of berkson and classical errors. *Canadian Journal of Statistics 35*(1), 89–104.

Delaigle, A. (2008). An alternative view of the deconvolution problem. *Statistica Sinica 18*, 1025–1045.

Delaigle, A. (2014). Nonparametric kernel methods with errors-in-variables: constructing estimators, computing them, and avoiding common mistakes. *Australian & New Zealand Journal of Statistics 56*(2), 105–124.

Delaigle, A., J. Fan, and R. J. Carroll (2009). A design-adaptive local polynomial estimator for the errors-in-variables problem. *Journal of the American Statistical Association 104*(485), 348–359.

Delaigle, A. and I. Gijbels (2002). Estimation of integrated squared density derivatives from a contaminated sample. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(4), 869–886.

Delaigle, A. and I. Gijbels (2004a). Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *Annals of the Institute of Statistical Mathematics 56*(1), 19–47.

Delaigle, A. and I. Gijbels (2004b). Practical bandwidth selection in deconvolution kernel density estimation. *Computational statistics & data analysis 45*(2), 249–267.

Delaigle, A. and I. Gijbels (2006a). Data-driven boundary estimation in deconvolution problems. *Computational statistics & data analysis 50*(8), 1965–1994.

Delaigle, A. and I. Gijbels (2006b). Estimation of boundary and discontinuity points in deconvolution problems. *Statistica Sinica 16*, 773–788.

Delaigle, A. and I. Gijbels (2007). Frequent problems in calculating integrals and optimizing objective functions: a case study in density deconvolution. *Statistics and Computing 17*(4), 349–355.

Delaigle, A. and P. Hall (2006). On optimal kernel choice for deconvolution. *Statistics & Probability Letters 76*(15), 1594–1602.

Delaigle, A. and P. Hall (2008). Using simex for smoothing-parameter choice in errors-in-variables problems. *Journal of the American Statistical Association 103*(481), 280–287.

Delaigle, A. and P. Hall (2014). Parametrically assisted nonparametric estimation of a density in the deconvolution problem. *Journal of the American Statistical Association 109*(506), 717–729.

Delaigle, A. and P. Hall (2016). Methodology for non-parametric deconvolution when the error distribution is unknown. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*, 231–252.

Delaigle, A., P. Hall, and A. Meister (2008). On deconvolution with repeated measurements. *The Annals of Statistics 36*(2), 665–685.

Delaigle, A., T. Hyndman, and T. Wang (2020). Deconvolve: Deconvolution tools for measurement error problems. R package version 0.1.0.

Delaigle, A. and A. Meister (2007). Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *Journal of the American Statistical Association 102*, 1416–1426.

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

Delaigle, A. and A. Meister (2011). Nonparametric function estimation under fourier-oscillating noise. *Statistica Sinica*, 1065–1092.

Delaigle, A. and I. Van Keilegom (2021). Deconvolution with unknown error distribution. In G. Y. Yi, A. Delaigle, and P. Gustafson (Eds.), *Handbook of Measurement Error Models*, Chapter 12, pp. XXX–XXX. CRC.

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

Diggle, P. and P. Hall (1993). A fourier approach to nonparametric deconvolution of a density estimate. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 55*(2), 523–531.

Fan, J. (1991a). Asymptotic normality for deconvolution kernel density estimators. *Sankhyā: The Indian Journal of Statistics, Series A 53*, 97–110.

Fan, J. (1991b). Global behavior of deconvolution kernel estimates. *Statistica Sinica 1*, 541–551.

Fan, J. (1991c). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics 19*, 1257–1272.

Fan, J. (1992). Deconvolution with supersmooth distributions. *Canadian Journal of Statistics 20*(2), 155–169.

Fan, J. and Y. K. Truong (1993). Nonparametric regression with errors in variables. *The Annals of Statistics 21*, 1900–1925.

Goldenshluger, A. and A. Tsybakov (2004). Estimating the endpoint of a distribution in the presence of additive observation errors. *Statistics & Probability Letters 68*(1), 39–49.

Granovsky, B. L. and H.-G. Müller (1989). On the optimality of a class of polynomial kernel functions. *Statistics & Risk Modeling 7*(4), 301–312.

Granovsky, B. L. and H.-G. Müller (1991). Optimizing kernel methods: a unifying variational principle. *International Statistical Review/Revue Internationale de Statistique 59*, 373–388.

Granovsky, B. L., H.-G. Müller, and C. Pfeifer (1995). Some remarks on optimal kernel functions. *Statistics & Risk Modeling 13*(2), 101–116.

Groeneboom, P. and G. Jongbloed (2003). Density estimation in the uniform deconvolution model. *Statistica Neerlandica 57*(1), 136–157.

Hall, P. and S. N. Lahiri (2008). Estimation of distributions, moments and quantiles in deconvolution problems. *The Annals of Statistics 36*(5), 2110–2134.

Hall, P. and A. Meister (2007). A ridge-parameter approach to deconvolution. *The Annals of Statistics 35*(4), 1535–1558.

Hall, P. and R. D. Murison (1993). Correcting the negativity of high-order kernel density estimators. *Journal of Multivariate Analysis 47*(1), 103–122.

Hall, P. and L. Simar (2002). Estimating a changepoint, boundary, or frontier in the presence of observation error. *Journal of the American statistical Association 97*(458), 523–534.

Hazelton, M. L. and B. A. Turlach (2009). Nonparametric density deconvolution by weighted kernel estimators. *Statistics and Computing 19*(3), 217–228.

Hazelton, M. L. and B. A. Turlach (2010). Semiparametric density deconvolution. *Scandinavian Journal of Statistics 37*(1), 91–108.

Hesse, C. H. (1995). Deconvolving a density from partially contaminated observations. *Journal of multivariate analysis 55*(2), 246–260.

Hesse, C. H. (1996). How many "good"observations do you need for "fast" density deconvolution from supersmooth errors. *Sankhyā: The Indian Journal of Statistics, Series A 58*, 491–506.

Hesse, C. H. (1999). Data-driven deconvolution. *Journal of Nonparametric Statistics 10*(4), 343–373.

Holzmann, H., N. Bissantz, and A. Munk (2007). Density testing in a contaminated sample. *Journal of Multivariate Analysis 98*(1), 57–75.

Holzmann, H. and L. Boysen (2006). Integrated square error asymptotics for supersmooth deconvolution. *Scandinavian Journal of Statistics 33*(4), 849–860.

Kang, Y. and P. Qiu (2021). Nonparametric deconvolution by fourier transformation and other related approaches. In G. Y. Yi, A. Delaigle, and P. Gustafson (Eds.), *Handbook of Measurement Error Models*, Chapter 11, pp. ? CRC.

Kneip, A., L. Simar, and I. Van Keilegom (2015). Frontier estimation in the presence of measurement error with unknown variance. *Journal of Econometrics 184*(2), 379–393.

Kulik, R. (2008). Nonparametric deconvolution problem for dependent sequences. *Electronic Journal of Statistics 2*, 722–740.

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.

Liu, M. C. and R. L. Taylor (1990). Simulations and computations of nonparametric density estimates for the deconvolution problem. *Journal of Statistical Computation and Simulation 35*, 145–167.

Lütkenöner, B. (2015). A family of kernels and their associated deconvolving kernels for normally distributed measurement errors. *Journal of Statistical Computation and Simulation 85*(12), 2347–2363.

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

Masry, E. (1993a). Asymptotic normality for deconvolution estimators of multivariate densities of stationary processes. *Journal of Multivariate Analysis 44*(1), 47–68.

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

Masry, E. (2003). Deconvolving multivariate kernel density estimates from contaminated associated observations. *IEEE Transactions on Information Theory 49*(11), 2941–2952.

McIntyre, J. and L. Stefanski (2011). Density estimation with replicate heteroscedastic measurements. *Annals of the Institute of Statistical Mathematics 63*(1), 81–99.

Meister, A. (2006). Support estimation via moment estimation in presence of noise. *Statistics 40*(3), 259–275.

Meister, A. (2007). Deconvolution from fourier-oscillating error densities under decay and smoothness restrictions. *Inverse Problems 24*(1), 015003.

Meister, A. (2009a). *Deconvolution Problems in Nonparametric Statistics.* Springer.

Meister, A. (2009b). On testing for local monotonicity in deconvolution problems. *Statistics & Probability Letters 79*(3), 312–319.

Meister, A. and M. H. Neumann (2010). Deconvolution from non-standard error densities under replicated measurements. *Statistica Sinica 20*, 1609–1636.

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

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

Potgieter, C. J. (2020). Density deconvolution for generalized skew-symmetric distributions. *Journal of Statistical Distributions and Applications 7*(1), 1–20.

Rachdi, M. and R. Sabre (2000). Consistent estimates of the mode of the probability density function in nonparametric deconvolution problems. *Statistics & Probability Letters 47*(2), 105–114.

Song, W. (2010). Moderate deviations for deconvolution kernel density estimators with ordinary smooth measurement errors. *Statistics & Probability Letters 80*, 169–176.

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

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

Stefanski, L. and J. R. Cook (1995). Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association 90*(432), 1247–1256.

Van Es, A. and A. Kok (1998). Simple kernel estimators for certain nonparametric deconvolution problems. *Statistics & Probability Letters 39*(2), 151–160.

Van Es, A. and H.-W. Uh (2004). Asymptotic normality of nonparametric kernel type deconvolution density estimators: crossing the cauchy boundary. *Nonparametric Statistics 16*, 261–277.

Van Es, A. and H.-W. Uh (2005). Asymptotic normality of kernel-type deconvolution estimators. *Scandinavian Journal of Statistics 32*(3), 467–483.

Van Es, B. and S. Gugushvili (2010). Asymptotic normality of the deconvolution kernel density estimator under the vanishing error variance. *Journal of the Korean Statistical Society 39*, 103–115.

Youndjé, É. and M. T. Wells (2002). Least squares cross-validation for the kernel deconvolution density estimator. *Comptes Rendus Mathematique 334*(6), 509–513.

Youndjé, É. and M. T. Wells (2008). Optimal bandwidth selection for multivariate kernel deconvolution density estimation. *Test 17*(1), 138–162.

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

Zhang, S. and R. J. Karunamuni (2000). Boundary bias correction for nonparametric deconvolution. *Annals of the Institute of Statistical Mathematics 52*(4), 612–629.

Zhang, S. and R. J. Karunamuni (2009). Deconvolution boundary kernel method in nonparametric density estimation. *Journal of Statistical Planning and Inference 139*(7), 2269–2283.

Zu, Y. (2015). A note on the asymptotic normality of the kernel deconvolution density estimator with logarithmic chi-square noise. *Econometrics 3*(3), 561–576.