Electronic Journal of Statistics
Vol. 10 (2016) 3579–3620
ISSN: 1935-7524
DOI: 10.1214/16-EJS1210

Nonparametric modal regression in the presence of measurement error

Haiming Zhou

Division of Statistics, Northern Illinois University, DeKalb, IL 60115, USA
e-mail: zhouh@niu.edu

and

Xianzheng Huang

Department of Statistics, University of South Carolina, Columbia, SC 29208, USA
e-mail: huang@stat.sc.edu

Abstract: In the context of regressing a response Y on a predictor X, we consider estimating the local modes of the distribution of Y given $X = x$ when X is prone to measurement error. We propose two nonparametric estimation methods, with one based on estimating the joint density of (X, Y) in the presence of measurement error, and the other built upon estimating the conditional density of Y given $X = x$ using error-prone data. We study the asymptotic properties of each proposed mode estimator, and provide implementation details including the mean-shift algorithm for mode seeking and bandwidth selection. Numerical studies are presented to compare the proposed methods with an existing mode estimation method developed for error-free data naively applied to error-prone data.

MSC 2010 subject classifications: Primary 62G08; secondary 62G20.
Keywords and phrases: Bandwidth selection, deconvoluting kernel, Fourier transform, local mode, mean-shift algorithm.

Received May 2016.

1. Introduction

The majority of statistical literature on regression analysis focuses on inferring the mean function of the response Y given a predictor X. There are also a large body of work devoted to making inference for the quantiles of Y given X in the regression setting (e.g., Koenker, 2005). Recently, researchers started to investigate methods to infer local modes of Y given X (see Yao *et al.*, 2012; Yao and Li, 2014; Chen *et al.*, 2016, among others). These researchers pointed out valuable information about the association between the response and the predictor that conditional modes can provide yet the conditional mean/quantiles can miss. Advantages of modal regression compared to mean or quantile regression have been well appreciated in analyzing speed-flow data in traffic engineering (Einbeck and Tutz, 2006), studying temperature patterns (Hyndman *et al.*, 1996), investigating galaxy properties conditioning on a given environment (Bamford *et al.*, 2008), and in economics (Huang *et al.*, 2013) for instance. To address the practical issue of error-contaminated covariates, methods accounting for measurement error in mean regression (Carroll *et al.*, 2006) and methods for quantile regression in the presence of measurement error (He and Liang, 2000; Wei and Carroll, 2009; Ma and Yin, 2011; Wang *et al.*, 2012) have been developed. In contrast, there is no existing research on modal regression when X is prone to measurement error even though this issue often arises in the aforementioned applications, and ignoring measurement error in modal regression typically results in misleading inference. We tackle this important problem in our study.

We propose two nonparametric methods for estimating conditional modes of a response variable Y given an error-prone covariate X. The first method exploits a kernel density estimator of the joint density of (X, Y) that accounts for measurement error in X. The second method is based on a local linear estimator of the conditional density of Y given $X = x$. These methods are elaborated in Section 2. Asymptotic properties of the mode estimators resulting from these methods are presented in Section 3. We provide a data-driven method for bandwidth selection to facilitate the implementation of the proposed methods in Section 4. Numerical studies are presented in Section 5, which include simulation experiments and an application to dietary data. Section 6 provides a discussion on follow-up open research questions.

2. Nonparametric estimation of local modes

2.1. Data and measurement error model

Suppose one wishes to collect data for a response Y and a covariate X, $\{(X_j, Y_j), j = 1, \ldots, n\}$, consisting of n independent pairs from a bivariate distribution specified by the joint probability density function (pdf), $p(x, y)$. However, the reality is that $\mathbf{X} = \{X_j\}_{j=1}^n$ cannot be observed directly due to error contamination, and $\mathbf{W} = \{W_j\}_{j=1}^n$ are observed instead. More specifically, the observed covariate W relates to the underlying true covariate X via an additive measurement error model given by

$$W_j = X_j + U_j, \text{ for } j = 1, \ldots, n, \tag{2.1}$$

where $\mathbf{U} = \{U_j\}_{j=1}^n$ are nondifferential measurement errors (Carroll *et al.*, 2006, Section 2.5), meaning that $\mathbf{U} \perp \mathbf{X}$ and $\mathbf{U} \perp \mathbf{Y} = \{Y_j\}_{j=1}^n$. We assume a known distribution of U specified by its pdf $f_U(u)$ in this study. Estimating the distribution of U requires either replicate measures of each underlying X_j or external validation data. For instance, if replicate measures are available and one assumes $f_U(u)$ known up to some parameters, such as the variance parameter, then one can follow equation (4.3) in Carroll *et al.* (2006) to estimate the measurement error variance consistently. One may also estimate the characteristic function of U as proposed in Delaigle *et al.* (2008), which is all our proposed inference methods need in terms of information regarding the measurement error distribution.

The focal point of statistical inference presented in this study lies in local modes of the conditional pdf of Y given $X = x$, $p(y|x)$. Denote by $\mathscr{X}$ the support of X and by $\mathscr{Y}$ the support of Y. For a generic bivariate function $g(s,t)$, $g_s(s_1,t)$ refers to $(\partial/\partial s)g(s,t)$ evaluated at $s = s_1$, and notations for higher-order partial derivatives of $g(s,t)$ are similarly defined. Given $x \in \mathscr{X}$, a mode of $p(y|x)$, denoted by $y_{\scriptscriptstyle M}(x)$, is a value in $\mathscr{Y}$ such that $p_y\{y_{\scriptscriptstyle M}(x)|x\} = 0$ and $p_{yy}\{y_{\scriptscriptstyle M}(x)|x\} < 0$. Equivalently, $y_{\scriptscriptstyle M}(x)$ satisfies $p_y\{x, y_{\scriptscriptstyle M}(x)\} = 0$ and $p_{yy}\{x, y_{\scriptscriptstyle M}(x)\} < 0$. This latter viewpoint motivates our first proposed estimator of $y_{\scriptscriptstyle M}(x)$ described next. It is possible that $p(y|x)$ is multimodal at a given x, producing a mode set $M(x) = \{y \in \mathscr{Y} : p_y(x,y) = 0 \text{ and } p_{yy}(x,y) < 0\}$. Although multi-modality brings in certain challenges in the actual implementation of mode seeking, it adds little complication in asymptotics analyses of our proposed mode estimators. To avoid unnecessarily tedious notations, we assume a unimodal $p(y|x)$ for the methodology development and theoretical analysis in the main article. We repeat the key part of the preliminary theoretical development with the uni-modality assumption relaxed in Appendix I for illustration purposes. In addition, multimodal $p(y|x)$ is considered when illustrating the implementation of the proposed methods in Section 5.

Even though the conditional mode set $M(x)$ is formulated above in terms of the joint pdf $p(x,y)$, we shall point out that, as x moves in $\mathscr{X}$, the resultant conditional mode curve(s), $\{M(x),\, x \in \mathscr{X}\}$, characterize the mode structure of the conditional density of Y given X. Hence, they typically differ from density ridges (Genovese *et al.*, 2014) and principal curves (Hastie and Stuetzle, 1989; Ozertem and Erdogmus, 2011), which focus on certain structures of the joint pdf. Chen *et al.* (2016, Section 8) provided detailed explanations on the distinction between conditional mode curves and density ridges, which are also helpful for one to see how they differ from principal curves.

2.2. Local constant estimator

We first consider an estimator of $y_{\scriptscriptstyle M}(x)$, denoted by $\hat{y}_{\scriptscriptstyle M0}(x)$, as the solution to $\hat{p}_y(x,y) = 0$, where $\hat{p}_y(x,y)$ is a kernel-based estimator of $p_y(x,y)$. The construction of $\hat{p}_y(x,y)$ traces back to the kernel density estimator of $p(x,y)$ in the absence of measurement error considered in Wand and Jones (1993),

$$\tilde{p}(x,y) = \frac{1}{nh_1h_2}\sum_{j=1}^{n} K_1\left(\frac{X_j - x}{h_1}\right) K_2\left(\frac{Y_j - y}{h_2}\right), \tag{2.2}$$

where h_1 and h_2 are bandwidths, and $K_1(t)$ and $K_2(t)$ are kernels. In the presence of measurement error, with $\mathbf{W}$ observed in place of $\mathbf{X}$, we follow the idea of deconvoluting kernel (Carroll and Hall, 1988; Stefanski and Carroll, 1990) and propose an estimator of $p(x,y)$ that accounts for measurement error as follows,

$$\hat{p}(x,y) = \frac{1}{nh_1h_2}\sum_{j=1}^{n} K_{\scriptscriptstyle U,0}\left(\frac{W_j - x}{h_1}\right) K_2\left(\frac{Y_j - y}{h_2}\right), \tag{2.3}$$

where

$$K_{U,0}(t)=\frac{1}{2\pi}\int e^{-its}\frac{\phi_{K_1}(s)}{\phi_U(s/h_1)}ds \tag{2.4}$$

is the deconvoluting kernel, in which i is the imaginary unit, $\phi_{K_1}(\cdot)$ is the Fourier transform of $K_1(\cdot)$, and $\phi_U(\cdot)$ is the characteristic function of U, i.e., the Fourier transform of $f_U(\cdot)$. In this article, all integrations are over the entire real line unless otherwise specified. The estimator $\hat{p}(x,y)$ is motivated by the property of the deconvoluting kernel that $E[K_{U,0}\{(W-x)/h_1\}|X]=K_1\{(X-x)/h_1\}$. A direct implication of this property is that $E\{\hat{p}(x,y)\}=E\{\tilde{p}(x,y)\}$.

Differentiating (2.3) with respect to (w.r.t.) y yields an estimator of $p_y(x,y)$ based on $(\mathbf{W},\mathbf{Y})$,

$$\hat{p}_y(x,y)=\frac{-1}{nh_1h_2^2}\sum_{j=1}^{n}K_{U,0}\left(\frac{W_j-x}{h_1}\right)K_2'\left(\frac{Y_j-y}{h_2}\right), \tag{2.5}$$

where $K_2'(t)=(d/dt)K_2(t)$. To facilitate mode seeking, we choose $K_2(t)$ to be a radially symmetric kernel, that is, $K_2(t)=c_2k_2(t^2)$, in which $k_2(s)$ is a nonnegative-valued function, referred to as the profile of $K_2(t)$, and c_2 is a positive constant serving as a normalization constant so that $K_2(t)$ integrates to one. Furthermore, we choose $K_2(t)$ such that its profile $k_2(s)$ relates to the profile of another radially symmetric kernel $K_3(t)=c_3k_3(t^2)$ via $k_3(s)=-k_2'(s)=-(d/ds)k_2(s)$. The Epanechnikov kernel and the normal kernel are examples where a kernel possesses the above desirable features for $K_2(t)$. Using the so-chosen $K_2(t)$, (2.5) can be further elaborated as

$$\hat{p}_y(x,y)=\frac{2c_2}{nh_1h_2^2c_3}\sum_{j=1}^{n}K_{U,0}\left(\frac{W_j-x}{h_1}\right)K_3\left(\frac{Y_j-y}{h_2}\right)\left(\frac{Y_j-y}{h_2}\right).$$

For illustration purposes, throughout the article, we set $K_2(t)=\exp(-t^2/2)/\sqrt{2\pi}$. Then, with $c_2=1/(2\sqrt{2\pi})$ and $k_2(s)=2\exp(-s/2)$, one has $K_3(t)=K_2(t)$, with $c_3=1/\sqrt{2\pi}$ and $k_3(s)=\exp(-s/2)$, and the above estimator becomes

$$\hat{p}_y(x,y)=\frac{1}{nh_1h_2^2}\sum_{j=1}^{n}K_{U,0}\left(\frac{W_j-x}{h_1}\right)K_2\left(\frac{Y_j-y}{h_2}\right)\left(\frac{Y_j-y}{h_2}\right). \tag{2.6}$$

Based on $\hat{p}_y(x,y)$ in (2.6), an estimator of $y_M(x)$ is the solution to the equation

$$\sum_{j=1}^{n}K_{U,0}\left(\frac{W_j-x}{h_1}\right)K_2\left(\frac{Y_j-y}{h_2}\right)(Y_j-y)=0.$$

Rearranging terms in this equation yields a variant of the equation leading to the following updating formula that one evaluates iteratively until convergence in order to find a solution to it,

$$y^{(k+1)} = \frac{\sum_{j=1}^{n} K_{U,0}\left(\frac{W_j - x}{h_1}\right) K_2\left(\frac{Y_j - y^{(k)}}{h_2}\right) Y_j}{\sum_{j=1}^{n} K_{U,0}\left(\frac{W_j - x}{h_1}\right) K_2\left(\frac{Y_j - y^{(k)}}{h_2}\right)}, \tag{2.7}$$

where $y^{(k+1)}$ is the value resulting from the $(k+1)$th iteration as an update of the value from the previous iteration, $y^{(k)}$, in search for $\hat{y}_{M0}(x)$, for $k = 0, 1, \ldots$. One may view

$$\omega_j^{(k)} = \frac{K_{U,0}\{(W_j - x)/h_1\} K_2\{(Y_j - y^{(k)})/h_2\}}{\sum_{j'=1}^{n} K_{U,0}\{(W_{j'} - x)/h_1\} K_2\{(Y_{j'} - y^{(k)})/h_2\}}$$

as a weight associated with the jth data point (W_j, Y_j), for $j = 1, \ldots, n$, of which the magnitude depends on the proximity of this data point to $(x, y^{(k)})$. Following this viewpoint, one can see that the right-hand side of (2.7) is a weighted average of $\mathbf{Y}$, and one may interpret this updating formula as updating $y^{(k)}$ to $y^{(k+1)}$ using the weighted mean of the responses surrounding $(x, y^{(k)})$. In fact, this algorithm of finding an estimated mode is in line with the mean-shift algorithm used to search for modes of a distribution (Cheng, 1995; Comaniciu and Meer, 2002; Einbeck and Tutz, 2006; Chen *et al.*, 2016). Compared to the existing mean-shift algorithm and its application, the complication caused by measurement error is that the weight $\omega_j^{(k)}$ can be negative because the deconvoluting kernel $K_{U,0}(\cdot)$ is not guaranteed to be nonnegative (Stefanski and Carroll, 1990). However, with careful choices of the bandwidths and starting values for the algorithm, as to be elaborated in Sections 4 and 5, our mean-shift algorithm can converge and produce a mode estimate $\hat{y}_{M0}(x)$.

We call the so-obtained estimator $\hat{y}_{M0}(x)$ a local constant estimator of the mode because of the construction of $\hat{p}(x, y)$, which in nature is a local constant estimator of $p(x, y)$. This interpretation of $\hat{p}(x, y)$ is made clearer when compared to the way we estimate $p(y|x)$ in order to estimate the mode.

2.3. Local linear estimator

The second estimator of $y_M(x)$ we propose, denoted by $\hat{y}_{M1}(x)$, is a solution to $\hat{p}_y(y|x) = 0$, where $\hat{p}_y(y|x)$ is an estimator of $p_y(y|x)$ obtained as follows. We start from evoking the local linear estimator of $p(y|x)$ in the absence of measurement error proposed by Fan *et al.* (1996) given by

$$\tilde{p}(y|x) = \boldsymbol{e}_1^{\mathrm{T}} \mathbf{S}_n^{-1}(x) \mathbf{T}_n(x, y), \tag{2.8}$$

where $\boldsymbol{e}_1 = (1,\, 0)^{\mathrm{T}}$,

$$\mathbf{S}_n(x) = \begin{bmatrix} S_{n,0}(x) & S_{n,1}(x) \\ S_{n,1}(x) & S_{n,2}(x) \end{bmatrix}, \quad \mathbf{T}_n(x, y) = [T_{n,0}(x, y),\, T_{n,1}(x, y)]^{\mathrm{T}},$$

in which

$$S_{n,\ell}(x) = \frac{1}{nh_1}\sum_{j=1}^{n}\left(\frac{X_j - x}{h_1}\right)^{\ell} K_1\left(\frac{X_j - x}{h_1}\right), \text{ for } \ell = 0, 1, 2, \tag{2.9}$$

$$T_{n,\ell}(x,y) = \frac{1}{nh_1h_2}\sum_{j=1}^{n}\left(\frac{X_j - x}{h_1}\right)^{\ell} K_1\left(\frac{X_j - x}{h_1}\right) K_2\left(\frac{Y_j - y}{h_2}\right), \text{ for } \ell = 0, 1. \tag{2.10}$$

This estimator is motivated by the property that $E[K_2\{(Y-y)/h_2\}/h_2|X = x] \approx p(y|x)$ as $h_2 \to 0$, and hence $p(y|x)$ can be approximately viewed as the mean function when regressing $K_2\{(Y-y)/h_2\}/h_2$ on X. Adopting this standpoint, one can employ the general strategy of estimating a mean function via local polynomial estimators (Fan and Gijbels, 1996) to estimate $p(y|x)$, with $\{K_2\{(Y_j-y)/h_2\}/h_2\}_{j=1}^n$ being the response data and $\{X_j\}_{j=1}^n$ as the covariate data. In particular, the local linear estimator of $p(y|x)$ is as in (2.8).

In the presence of measurement error, we adjust $\tilde{p}(y|x)$ to account for measurement error in X and propose the following local linear estimator of $p(y|x)$,

$$\hat{p}(y|x) = \boldsymbol{e}_1^{\mathrm{T}}\hat{\mathbf{S}}_n^{-1}(x)\hat{\mathbf{T}}_n(x,y), \tag{2.11}$$

where

$$\hat{\mathbf{S}}_n(x) = \begin{bmatrix}\hat{S}_{n,0}(x) & \hat{S}_{n,1}(x)\\ \hat{S}_{n,1}(x) & \hat{S}_{n,2}(x)\end{bmatrix}, \quad \hat{\mathbf{T}}_n(x,y) = \left[\hat{T}_{n,0}(x,y),\, \hat{T}_{n,1}(x,y)\right]^{\mathrm{T}},$$

in which

$$\hat{S}_{n,\ell}(x) =\frac{1}{nh_1}\sum_{j=1}^{n} K_{U,\ell}\left(\frac{W_j - x}{h_1}\right), \text{ for } \ell = 0, 1, 2, \tag{2.12}$$

$$\hat{T}_{n,\ell}(x,y) =\frac{1}{nh_1h_2}\sum_{j=1}^{n} K_{U,\ell}\left(\frac{W_j - x}{h_1}\right) K_2\left(\frac{Y_j - y}{h_2}\right), \text{ for } \ell = 0, 1, \tag{2.13}$$

and, with $\phi_{K_1}^{(\ell)}(s)$ denoting the ℓ-th derivative of $\phi_{K_1}(s)$,

$$K_{U,\ell}(t) = i^{-\ell}\frac{1}{2\pi}\int e^{-its}\frac{\phi_{K_1}^{(\ell)}(s)}{\phi_U(s/h_1)}ds, \text{ for } \ell = 0, 1, 2. \tag{2.14}$$

The transform of $K_1(\cdot)$ in (2.14) is a generalization of the transform in (2.4) derived in Delaigle *et al.* (2009). This generalization leads to a generalized deconvoluting kernel $K_{U,\ell}(t)$ possessing the property that $E[K_{U,\ell}\{(W-x)/h_1\}|X] = \{(X-x)/h_1\}^{\ell}K_1\{(X-x)/h_1\}$, $\ell = 0, 1, 2, \ldots$. Thanks to this property, given $(\mathbf{X}, \mathbf{Y})$, (2.12) and (2.13) are unbiased estimators of their counterparts in the absence of measurement error in (2.9) and (2.10), respectively. Hence, $\hat{p}(y|x)$ is a sensible counterpart estimator of $\tilde{p}(y|x)$ in the presence of measurement error.

Using (2.11), an estimator of $p_y(y|x)$ follows immediately by differentiating $\hat{p}(y|x)$ w.r.t. y. This gives

$$\hat{p}_y(y|x) = \boldsymbol{e}_1^{\mathrm{T}} \hat{\mathbf{S}}_n^{-1}(x) \hat{\mathbf{T}}'_n(x, y), \tag{2.15}$$

where $\hat{\mathbf{T}}'_n(x, y) = [\hat{T}'_{n,0}(x, y), \hat{T}'_{n,1}(x, y)]^{\mathrm{T}}$, in which, for $\ell = 0, 1$,

$$\hat{T}'_{n,\ell}(x, y) = \frac{\partial}{\partial y} \hat{T}_{n,\ell}(x, y) = \frac{1}{n h_1 h_2^2} \sum_{j=1}^{n} K_{U,\ell}\left(\frac{W_j - x}{h_1}\right) K_2\left(\frac{Y_j - y}{h_2}\right)\left(\frac{Y_j - y}{h_2}\right).$$

Setting $\hat{p}_y(y|x) = 0$ gives an equation to which the solution is the mode estimator $\hat{y}_{M1}(x)$. Elaborating (2.15) reveals that $\hat{y}_{M1}(x)$ solves

$$\sum_{j=1}^{n} \left\{ K_{U,0}\left(\frac{W_j - x}{h_1}\right) \hat{S}_{n,2}(x) - K_{U,1}\left(\frac{W_j - x}{h_1}\right) \hat{S}_{n,1}(x) \right\} K_2\left(\frac{Y_j - y}{h_2}\right)(Y_j - y) = 0.$$

This suggests the following updating formula one may use iteratively until convergence to find the solution to the equation,

$$y^{(k+1)} = \frac{\sum_{j=1}^{n} \left\{ K_{U,0}\left(\frac{W_j - x}{h_1}\right) \hat{S}_{n,2}(x) - K_{U,1}\left(\frac{W_j - x}{h_1}\right) \hat{S}_{n,1}(x) \right\} K_2\left(\frac{Y_j - y^{(k)}}{h_2}\right) Y_j}{\sum_{j=1}^{n} \left\{ K_{U,0}\left(\frac{W_j - x}{h_1}\right) \hat{S}_{n,2}(x) - K_{U,1}\left(\frac{W_j - x}{h_1}\right) \hat{S}_{n,1}(x) \right\} K_2\left(\frac{Y_j - y^{(k)}}{h_2}\right)}.$$

Like the previous updating formula, the right-hand side of this updating formula can also be viewed as a weighted average of $\mathbf{Y}$, and thus this algorithm of searching for an estimated mode is also in the spirit of the mean-shift algorithm.

The common theme the above two proposed mode estimation methods follow is to solve an equation of the form $\hat{g}_y(x, y) = 0$, where $\hat{g}_y(x, y)$ an estimator of $g_y(x, y)$, and $g(x, y)$ is a function such that $g_y\{x, y_M(x)\} = 0$ and $g_{yy}\{x, y_M(x)\} < 0$. When $g(x, y) = p(x, y)$, the solution is $\hat{y}_{M0}(x)$; and, when $g(x, y) = p(y|x)$, solving the equation yields $\hat{y}_{M1}(x)$. Furthermore, this common theme closely relates to the idea of corrected score (Nakamura, 1990; Novick and Stefanski, 2002; Carroll *et al.*, 2006). More specifically, since $E\{\hat{p}(x, y)|(\mathbf{X}, \mathbf{Y})\} = \tilde{p}(x, y)$, one has $E\{\hat{p}_y(x, y)|(\mathbf{X}, \mathbf{Y})\} = \tilde{p}_y(x, y)$. Hence, the estimating equation one solves to obtain $\hat{y}_{M0}(x)$, i.e., $\hat{p}_y(x, y) = 0$, is the corrected score estimating equation corresponding to $\tilde{p}_y(x, y) = 0$, which is the equation one solves to estimate modes in the absence of measurement error. Although, given $(\mathbf{X}, \mathbf{Y})$, $\hat{p}_y(y|x)$ is not an unbiased estimator of $\tilde{p}_y(y|x)$, the building blocks in the former are unbiased scores of those in the latter, i.e., $E\{\hat{\mathbf{S}}_n(x)|\mathbf{X}\} = \mathbf{S}_n(x)$ and $E\{\hat{\mathbf{T}}_n(x, y)|(\mathbf{X}, \mathbf{Y})\} = \mathbf{T}_n(x, y)$. When the solution to an equation associated with a method is not unique, the method leads to an estimated mode set, denoted by $\hat{M}_0(x)$ and $\hat{M}_1(x)$ for the first and the second method, respectively. In what follows, we present asymptotic properties of these mode estimators. For notational simplicity, we assume $\hat{M}_0(x) = \{\hat{y}_{M0}(x)\}$ and $\hat{M}_1(x) = \{\hat{y}_{M1}(x)\}$ in the next section. Also, y_M is often used in place of $y_M(x)$ for brevity in the sequel, and $\hat{y}_M$ is used to refer to a mode estimator generically when we do not distinguish between $\hat{y}_{M0}(x)$ and $\hat{y}_{M1}(x)$.

3. Asymptotic properties

3.1. Preliminary

We focus on convergence rates of three forms of error associated with $\hat{y}_M$ in this section, first, the pointwise error defined by $\Delta_n(x) = |\hat{y}_M(x) - y_M(x)|$; second, the mean integrated squared error (MISE), $\text{MISE}(\hat{y}_M) = E\{\int_{\mathscr{X}} \Delta_n^2(x)dx\}$; and third, the uniform error, $\Delta_n = \sup_{x\in\mathscr{X}} \Delta_n(x)$. We show next that the convergence rate of $\Delta_n(x)$ hinges on the bias and variance of $\hat{g}_y(x, y_M)$. Given the pointwise error rate, the convergence rate of $\text{MISE}(\hat{y}_M)$ follows straightforwardly. Under regularity conditions pointed out along the way, the uniform error rate can be established using existing results regarding the uniform consistency of kernel-based estimators (Ginè and Guillou, 2002; Einmahl and Mason, 2005; Chen *et al.*, 2015).

Because $\hat{g}_y(x, \hat{y}_M) = 0$, by the mean-value theorem, one has $\hat{g}_y(x, y_M) = \hat{g}_y(x, y_M) - \hat{g}_y(x, \hat{y}_M) = (y_M - \hat{y}_M)\hat{g}_{yy}(x, y^*)$, where y^* lies between y_M and $\hat{y}_M$. Thus,

$$\hat{y}_M - y_M = -\frac{\hat{g}_y(x, y_M)}{\hat{g}_{yy}(x, y^*)}. \tag{3.1}$$

Provided that $\hat{g}_{yy}(x, y_M)$ and $g_{yy}(x, y_M)$ are bounded away from zero, one has

$$|\{\hat{g}_{yy}(x, y^*)\}^{-1} - \{g_{yy}(x, y_M)\}^{-1}| = O\left(\|\hat{g}_{yy} - g_{yy}\|_\infty\right),$$

where $\|\hat{g}_{yy} - g_{yy}\|_\infty = \sup_{(x,y)\in\mathscr{X}\times\mathscr{Y}} |\hat{g}_{yy}(x, y) - g_{yy}(x, y)|$. Then (3.1) implies that

$$\hat{y}_M - y_M = -\{g_{yy}(x, y_M)\}^{-1}\hat{g}_y(x, y_M) + O\left(\|\hat{g}_{yy} - g_{yy}\|_\infty\right)\hat{g}_y(x, y_M). \tag{3.2}$$

It follows that $\Delta_n(x) = |\{g_{yy}(x, y_M)\}^{-1}||\hat{g}_y(x, y_M)| + O(\|\hat{g}_{yy} - g_{yy}\|_\infty)|\hat{g}_y(x, y_M)|$, or, equivalently,

$$\frac{\Delta_n(x)}{|\{g_{yy}(x, y_M)\}^{-1}||\hat{g}_y(x, y_M)|} = 1 + O(\|\hat{g}_{yy} - g_{yy}\|_\infty)|g_{yy}(x, y_M)|. \tag{3.3}$$

Under conditions (CK2), (CK5) and (CK9) in Appendix A, by Lemma 10 in Chen *et al.* (2015), $\|\hat{g}_{yy} - g_{yy}\|_\infty$ converges to zero in probability. Therefore, under these conditions, (3.3) suggests that $\Delta_n(x)$ can be approximated by $|\{g_{yy}(x, y_M)\}^{-1}||\hat{g}_y(x, y_M)|$, and thus the convergence rate of $\Delta_n(x)$ can be revealed through studying the convergence rate of $|\hat{g}_y(x, y_M)|$.

Once we turn to studying the convergence rate of $|\hat{g}_y(x, y_M)|$, the bias and variance of $\hat{g}_y(x, y_M)$ become highly relevant. This connection can be explained by first noting that $g_y(x, y_M) = 0$, and thus one has

$$\begin{aligned} |\hat{g}_y(x, y_M)| &= |\hat{g}_y(x, y_M) - g_y(x, y_M)| \\ &\le |\hat{g}_y(x, y_M) - E\{\hat{g}_y(x, y_M)\}| + |E\{\hat{g}_y(x, y_M)\} - g_y(x, y_M)| \\ &= O_P\left[\sqrt{\text{Var}\{\hat{g}_y(x, y_M)\}}\right] + |\text{Bias}\{\hat{g}_y(x, y_M)\}|. \end{aligned} \tag{3.4}$$

The above preliminary asymptotic analysis leads to the road map we follow to study the error rates associated with $\hat{y}_M$, which is to, first, derive the asymptotic bias and variance of $\hat{g}_y(x, y_M)$, which leads to the pointwise error rate by (3.4); second, establish the convergence rate of MISE$(\hat{y}_M)$; third, provide the uniform error rate. This is also the order in which we present our theoretical findings for each of the proposed mode estimator in the next two subsections. Supporting materials of these findings are provided in the appendices. In particular, all coded conditions referenced henceforth are in Appendix A.

3.2. Convergence rates associated with $\hat{y}_{M0}(x)$

With $g(x, y) = p(x, y)$, we establish the following asymptotic bias result for $\hat{p}_y(x, y_M)$, with the proof given in Appendix C.

Lemma 3.1. *Under conditions (CP1) and (CK1)–(CK4),*

$$Bias\{\hat{p}_y(x, y_M)\} = 0.5\left\{p_{xxy}(x, y_M)\mu_2^{(1)}h_1^2 + p_{yyy}(x, y_M)h_2^2\right\} + o(h_1^2 + h_2^2), \tag{3.5}$$

where $p_{xxy}(x, y) = (\partial^3/\partial x^2 \partial y)p(x, y)$, $p_{yyy}(x, y) = (\partial^3/\partial y^3)p(x, y)$, *and* $\mu_2^{(1)} = \int t^2 K_1(t)dt$.

This bias result coincides with the result in Chacón *et al.* (2011), where kernel-based estimators of derivatives of a multivariate joint pdf are considered in the absence of measurement error. This is expected since $E\{\hat{p}_y(x, y)\} = E\{\tilde{p}_y(x, y)\}$, as pointed out in Section 2.2.

The variance of $\hat{p}_y(x, y_M)$ depends on the smoothness of the measurement error distribution. There are two levels of smoothness of $f_U(u)$ considered in measurement error literature (Fan, 1991a,b,c), ordinary smooth and super smooth. Their definitions are given below.

Definition 3.1. *The distribution of U is ordinary smooth of order b if*

$$\lim_{t\to+\infty} t^b \phi_U(t) = c \text{ and } \lim_{t\to+\infty} t^{b+1}\phi_U'(t) = -cb$$

for some positive constants b and c.

Definition 3.2. *The distribution of U is super smooth of order b if*

$$d_0|t|^{b_0}\exp(-|t|^b/d_2) \le |\phi_U(t)| \le d_1|t|^{b_1}\exp(-|t|^b/d_2) \text{ as } |t| \to \infty$$

for some positive constants d_0, d_1, d_2, b, b_0 and b_1.

In Appendix D, we derive the asymptotic variance of $\hat{p}_y(x, y_M)$ and the results are given in the next lemma.

Lemma 3.2. *Assume conditions required for Lemma 3.1 hold. When U is ordinary smooth of order b, under conditions (CU2) and (CK5)–(CK7), if*

$nh_1^{1+2b}h_2^3 \to \infty$, then

$$Var\{\hat{p}_y(x, y_M)\} = \frac{\eta_0 f_{W,Y}(x, y_M)}{4\sqrt{\pi}c^2 n h_1^{1+2b} h_2^3} + o\left(\frac{1}{nh_1^{1+2b}h_2^3}\right), \tag{3.6}$$

where $\eta_0 = \int |t|^{2b}|\phi_{K_1}(t)|^2 dt/(2\pi)$, and $f_{W,Y}(\cdot,\cdot)$ is the joint pdf of (W, Y).

When U is super smooth of order b, under conditions (CK5) and (CK8), if $nh_1^{1-2b_2}h_2^3 \exp(-2h_1^{-b}/d_2) \to \infty$, then

$$Var\{\hat{p}_y(x, y_M)\} \leq \frac{\exp\left(2h_1^{-b}/d_2\right)}{nh_1^{1-2b_2}h_2^3} C f_{W,Y}(x, y_M) + o\left\{\frac{\exp\left(2h_1^{-b}/d_2\right)}{nh_1^{1-2b_2}h_2^3}\right\}, \tag{3.7}$$

where C is some finite positive constant and $b_2 = b_0 I(b_0 < 0.5)$.

Putting results in Lemmas 3.1 and 3.2 together, one has the convergence rate of $|\hat{p}_y(x, y_M)|$, which leads to the pointwise error rates summarized in the following theorem.

Theorem 3.1. *Under the conditions in Lemma 3.1, Lemma 3.2, and (CK9), when U is ordinary smooth,*

$$\Delta_n(x) = \frac{|p_{xxy}(x, y_M)\mu_2^{(1)} h_1^2 + p_{yyy}(x, y_M)h_2^2|}{-2p_{yy}(x, y_M)} + O_P\left(\sqrt{\frac{1}{nh_1^{1+2b}h_2^3}}\right) + o(h_1^2 + h_2^2); \tag{3.8}$$

and, when U is super smooth,

$$\Delta_n(x) = \frac{|p_{xxy}(x, y_M)\mu_2^{(1)} h_1^2 + p_{yyy}(x, y_M)h_2^2|}{-2p_{yy}(x, y_M)} + O_P\left\{\frac{\exp(h_1^{-b}/d_2)}{\sqrt{nh_1^{1-2b_2}h_2^3}}\right\} + o(h_1^2 + h_2^2). \tag{3.9}$$

Chen *et al.* (2016) estimated local modes based on $\tilde{p}(x, y)$ in (2.2) with $h_1 = h_2 = h$, and they showed that the pointwise error rate of the resultant mode estimator is of order $O(h^2) + O_P\{\sqrt{1/(nh^4)}\}$. Comparing this with (3.8) and (3.9), one can see that the pointwise error tends to zero much slower with the added complication of measurement error, especially when it is super smooth. Hence, mode estimation in the presence of measurement error is substantially more challenging, as one would expect with noisier data.

The convergence rate of $\text{MISE}(\hat{y}_{M0})$ can also be deduced from Lemmas 3.1 and 3.2. To see this more clearly, first note that, assuming interchangeability of expectation and integration, one has $\text{MISE}(\hat{y}_M) = \int_{\mathscr{X}} \left\{\text{Bias}^2(\hat{y}_M) + \text{Var}(\hat{y}_M)\right\} dx$. As elaborated in Appendix E, the dominating terms in the integrated squared bias of $\hat{y}_{M0}$, $\int_{\mathscr{X}} \text{Bias}^2(\hat{y}_{M0})dx$, can be easily derived based on $\text{Bias}\{\hat{p}_y(x, y_M)\}$; and the dominating terms in the integrated variance of $\hat{y}_{M0}$, $\int_{\mathscr{X}} \text{Var}(\hat{y}_{M0})dx$, can be deduced from $\text{Var}\{\hat{p}_y(x, y_M)\}$. Combining these dominating terms, we reach the following conclusion regarding $\text{MISE}(\hat{y}_{M0})$.

Theorem 3.2. *Under conditions in Theorem 3.1, and assume that $p_{xxy}(x, y_M)$ and $p_{yyy}(x, y_M)$ are square integrable, then, when U is ordinary smooth,*

$$MISE(\hat{y}_{M0}) = O\{(h_1^2 + h_2^2)^2\} + O\left(\frac{1}{nh_1^{1+2b}h_2^3}\right);$$

when U is super smooth,

$$MISE(\hat{y}_{M0}) = O\{(h_1^2 + h_2^2)^2\} + O\left\{\frac{\exp\left(2h_1^{-b}/d_2\right)}{nh_1^{1-2b_2}h_2^3}\right\}.$$

Moving to the uniform error, Δ_n, it is helpful to note that, by (3.3), (3.4), and Lemma 3.1, one has

$$\Delta_n(x) = |\{p_{yy}(x, y_M)\}^{-1}| \left[|\hat{p}_y(x, y_M) - E\{\hat{p}_y(x, y_M)\}| + O(h_1^2 + h_2^2)\right] + o_P(1).$$

Thus

$$\Delta_n = \sup_{x \in \mathscr{X}} |\{p_{yy}(x, y_M)\}^{-1}||\hat{p}_y(x, y_M) - E\{\hat{p}_y(x, y_M)\}| + O(h_1^2 + h_2^2) + o_P(1).$$

Following similar methods in Ginè and Guillou (2002) and Einmahl and Mason (2005), one can establish the following result regarding Δ_n.

Theorem 3.3. *Under conditions in Theorem 3.1 and (CP2), for ordinary smooth U,*

$$\Delta_n = O(h_1^2 + h_2^2) + O_P\left(\sqrt{\frac{\log n}{nh_1^{1+2b}h_2^3}}\right);$$

and, for super smooth U,

$$\Delta_n = O(h_1^2 + h_2^2) + O_P\left(\sqrt{\frac{\exp(h_1^{-b}/d_2)\log n}{nh_1^{1-2b_2}h_2^3}}\right).$$

The uniform error rate associated with the mode estimator with $h_1 = h_2 = h$ considered in Chen *et al.* (2016) in the absence of measurement error is $O(h^2) + O_P\{\sqrt{\log n/(nh^4)}\}$. Compared with their result, Theorem 3.3 suggests the inevitable compromise in convergence rate due to measurement error.

3.3. Convergence rates associated with $\hat{y}_{M1}(x)$

With $g(x, y) = p(y|x)$, we present in Appendix F the bias analysis and in Appendix G the variance analysis of $\hat{p}_y(y_M|x)$. Results from these analyses are summarized in the following two lemmas.

Lemma 3.3. *When U is ordinary smooth, assume (CK4), (CK5), (CX1), and $(nh_1^{1+2b}h_2^3)^{-1/2} = O(h_1^4h_2^{-1} + h_2^3)$; when U is super smooth, assume (CK4),* *(CK6), (CX1), (CP1), and* $\exp(h_1^{-b}/d_2)(nh_1^{1-2b_2}h_2^3)^{-1/2} = O(h_1^4h_2^{-1}+h_2^3)$*, then one has*

$$Bias\{\hat{p}_y(y_M|x)\} = f_X^{-1}(x)\Big[\{0.5p_{xxy}(x,y_M) - f_X'(x)p_{xy}(y_M|x)\}\,\mu_2^{(1)}h_1^2 + 0.5p_{yyy}(x,y_M)h_2^2\Big] + O(h_1^4h_2^{-1}+h_2^3), \tag{3.10}$$

where $f_X(x)$ *is the pdf of* X *and* $f_X'(x) = (d/dx)f_X(x)$*.*

Lemma 3.4. *Under the conditions in Lemma 3.3, when* U *is ordinary smooth, if* $nh_1^{1+2b}h_2^3 \to \infty$*, then*

$$Var\{\hat{p}_y(y_M|x)\} = \frac{\eta_0 f_{W,Y}(x,y_M)}{4\sqrt{\pi}c^2nh_1^{1+2b}h_2^3f_X^2(x)} + o\left(\frac{1}{nh_1^{1+2b}h_2^3}\right); \tag{3.11}$$

and, when U *is super smooth, if* $nh_1^{1-2b_2}h_2\exp(-2h_1/d_2) \to \infty$*, then*

$$Var\{\hat{p}_y(y_M|x)\} \le \frac{C\exp\left(2h_1^{-b}/d_2\right)}{nh_1^{1-2b_2}h_2^3f_X^2(x)} + O\left\{\frac{\exp\left(2h_1^{-b}/d_2\right)}{nh_1^{1-2b_2}h_2}\right\}, \tag{3.12}$$

where C *is some finite positive constant.*

Although the order of Bias$\{\hat{p}_y(y_M|x)\}$ in (3.10) and that of Bias$\{\hat{p}_y(x,y_M)\}$ in (3.5) are both $O(h_1^2+h_2^2)$, the dependence of the dominating term on $f_X^{-1}(x)$ shown in (3.10) indicates that estimating $\hat{p}_y(y_M|x)$ at the (diminishing) tail of the X-distribution can be challenging. The residual term in (3.10) suggests that we need $h_1^4h_2^{-1} \to 0$ in order for $\hat{p}_y(y_M|x)$, and thus for $\hat{y}_{M1}$, to be consistent. A sufficient condition for $h_1^4h_2^{-1} \to 0$ is to have $h_1 \to 0$ faster than $h_2 \to 0$ as $n \to \infty$. This may indicate that a sensible bandwidth selection procedure tends to choose $h_1 < h_2$, which is indeed observed in our simulation study when we apply the data-driven bandwidth selection method described in Section 4.

Comparing (3.11) with (3.6), one can see that the dominating variance of $\hat{p}_y(y_M|x)$ can be higher than that of $\hat{p}_y(x,y_M)$, and the former can be large at the (diminishing) tail of $f_X(x)$. This suggests that estimating $p_y(y_M|x)$ via $\hat{p}_y(y_M|x)$, and thus estimating $y_M(x)$ via $\hat{y}_{M1}(x)$, will be subject to high uncertainty if data surrounding x are scarce.

Based on these bias and variance results, we establish the following theorem regarding the pointwise error rate of $\hat{y}_{M1}(x)$.

Theorem 3.4. *Under conditions in Lemma 3.3, Lemma 3.4, (CK2) and (CK9), when* U *is ordinary smooth,*

$$\Delta_n(x) = \frac{|\{p_{xxy}(x,y_M) - 2f_X'(x)p_{xy}(y_M|x)\}\,\mu_2^{(1)}h_1^2 + p_{yyy}(x,y_M)h_2^2|}{-2f_X(x)p_{yy}(y_M|x)} + O_P\left(\sqrt{\frac{1}{nh_1^{1+2b}h_2^3}}\right) + O(h_1^4h_2^{-1}+h_2^3); \tag{3.13}$$

and, when U is super smooth,

$$\Delta_n(x) = \frac{|\{p_{xxy}(x, y_{\scriptscriptstyle M}) - 2f'_X(x)p_{xy}(y_{\scriptscriptstyle M}|x)\}\mu_2^{(1)}h_1^2 + p_{yyy}(x, y_{\scriptscriptstyle M})h_2^2|}{-2f_X(x)p_{yy}(y_{\scriptscriptstyle M}|x)} + O_P\left\{\frac{\exp(h_1^{-b}/d_2)}{\sqrt{nh_1^{1-2b_2}h_2^3}}\right\} + O(h_1^4h_2^{-1} + h_2^3). \tag{3.14}$$

Following the same line of arguments as those leading to MISE($\hat{y}_{\scriptscriptstyle M0}$) in Section 3.2, we show the same convergence rate for MISE($\hat{y}_{\scriptscriptstyle M1}$) as those stated in Theorem 3.2 under slightly different conditions. Now we need to assume all conditions stated in Lemmas 3.3 and 3.4, (CP2), and that $p_{xxy}(x, y_{\scriptscriptstyle M})$, $p_{yyy}(x, y_{\scriptscriptstyle M})$, and $p_{xy}(y_{\scriptscriptstyle M}|x)$ are square integrable.

Finally, for the uniform error Δ_n associated with $\hat{y}_{\scriptscriptstyle M1}$, using the elaboration of $\hat{p}_y(y_{\scriptscriptstyle M}|x)$ in Appendix F (Section F.1 to be specific), one can see that the dominating term of $\hat{p}_y(y_{\scriptscriptstyle M}|x)$ is simply $\hat{p}_y(x, y_{\scriptscriptstyle M})$ divided by $f_X(x)$. Hence, the convergence rate of Δ_n associated with $\hat{y}_{\scriptscriptstyle M1}$ is the same as that of Δ_n associated with $\hat{y}_{\scriptscriptstyle M0}$ stated in Theorem 3.3 under the same set of conditions, in addition to the assumption that $f_X(x)$ is bounded away from zero over the range of x of interest, $[x_{\scriptscriptstyle L}, x_{\scriptscriptstyle U}]$.

3.4. Asymptotic optimal bandwidths

With the asymptotic error rates of the proposed mode estimators provided in Sections 3.2 and 3.3, the asymptotic optimal (in some sense) bandwidths are readily available. In particular, taking MISE as the metric to optimize w.r.t. $\boldsymbol{h} = (h_1, h_2)^{\mathrm{T}}$, we show that, for both proposed mode estimators, the optimal rate of MISE($\hat{y}_{\scriptscriptstyle M}$) is of order $O(h_1^4)$ for both ordinary smooth U and super smooth U. The orders of the asymptotic optimal h_1 and h_2 (as $n \to \infty$) are given in a corollary next, where "$\asymp$" refers to "tending to zero or infinity at the same rate." Note that explicit expressions of the asymptotic optimal $\boldsymbol{h}$ are not available except for $\hat{y}_{\scriptscriptstyle M0}$ when U is ordinary smooth.

Corollary 3.1. *Under conditions in Theorem 3.2, when U is ordinary smooth of order b, the asymptotic optimal bandwidths for $\hat{y}_{\scriptscriptstyle M0}$ satisfy $h_1 = r_1h_2$ and $h_2 = r_2n^{-1/(2b+8)}$, where*

$$r_1 = \left\{\frac{(b-1)I_1 + \sqrt{(b-1)^2I_1^2 + 3(2b+1)I_2I_3}}{3\mu_2^{(1)}I_2}\right\}^{1/2},$$

$$r_2 = \left\{\frac{3\eta_0I_4}{4\sqrt{\pi}c^2r_1^{2b+1}(r_1^2\mu_2^{(1)}I_1 + I_3)}\right\}^{1/(2b+8)},$$

in which

$$
\begin{aligned}
I_1 &= \int_{\mathscr{X}} p_{yy}^{-2}(x, y_M) p_{xxy}(x, y_M) p_{yyy}(x, y_M) dx, \\
I_2 &= \int_{\mathscr{X}} p_{yy}^{-2}(x, y_M) p_{xxy}^2(x, y_M) dx, \\
I_3 &= \int_{\mathscr{X}} p_{yy}^{-2}(x, y_M) p_{yyy}^2(x, y_M) dx, \\
I_4 &= \int_{\mathscr{X}} p_{yy}^{-2}(x, y_M) f_{W,\,Y}(x, y_M) dx.
\end{aligned}
$$

When U is super smooth of order b, the asymptotic optimal bandwidths for $\hat{y}_{M0}$ satisfy $h_1 \asymp h_2^{2/(b+2)}$ and $\exp(2h_1^{-b}/d_2)/h_1^{3b/2-2b_2+6} \asymp n$. The rates of the asymptotic optimal h_1 and h_2 for $\hat{y}_{M1}$ are the same as those for $\hat{y}_{M0}$ under each type of U. The corresponding optimal rate of MISE($\hat{y}_M$) is of order $O(h_1^4)$ under each type of U for both $\hat{y}_{M0}$ and $\hat{y}_{M1}$.

These rates of the asymptotic optimal bandwidths for mode estimators are also the rates of the asymptotic optimal bandwidths for the corresponding density derivative estimators. This is not surprising considering the connection between the proposed mode estimators and density derivative estimators pointed out in Section 3.1. For instance, when U follows a Laplace distribution, which is ordinary smooth of order $b = 2$, the asymptotic optimal h_1 and h_2 are of order $O(n^{-1/12})$ for $\hat{p}_y(x, y)$, and the corresponding optimal MISE of $\hat{p}_y(x, y)$ is of order $O(h_1^4) = O(n^{-1/3})$. In the absence of measurement error, Chacón *et al.* (2011, Theorem 3) showed that the asymptotic optimal bandwidths for the kernel-based estimator of $p_y(x, y)$ is of order $O(n^{-1/4})$, and the corresponding optimal MISE of the density derivative estimator is of order $O(h_1^2) = O(n^{-1/2})$. This comparison highlights that measurement error inflate the optimal MISE rate of density derivative estimators, and also lead to much larger optimal bandwidths.

4. Bandwidth selection

The choice of bandwidths can noticeably affect finite sample performance of almost all kernel-based estimators. The explicit expression of the asymptotic optimal $\boldsymbol{h}$ for $\hat{y}_{M0}$ in Corollary 3.1 is not ready to use for choosing bandwidths given a finite sample until reliable estimators of unknown quantities, such as I_1, I_2, and I_3, are available. In this section we present a strategy of choosing $\boldsymbol{h}$ that mostly follows the idea of incorporating cross validation (CV) and simulation extrapolation (SIMEX, Cook and Stefanski, 1994) proposed by Delaigle and Hall (2008). This strategy is based on the CV method of choosing bandwidth for estimating conditional density in the absence of measurement error developed by Fan and Yim (2004) and Hall *et al.* (2004). These authors constructed a CV criterion based on the weighted integrated squared error (ISE) associated with a kernel-based estimator of $p(y|x)$, $\tilde{p}(y|x)$, defined by

$$
\begin{aligned}
\text{ISE} &= \int_{\mathscr{Y}}\int_{\mathscr{X}} \{\tilde{p}(y|x) - p(y|x)\}^2 f_X(x)\omega(x)dxdy \\
&= \int_{\mathscr{Y}}\int_{\mathscr{X}} \tilde{p}(y|x)^2 f_X(x)\omega(x)dxdy - 2\int_{\mathscr{Y}}\int_{\mathscr{X}} \tilde{p}(y|x)p(y|x)f_X(x)\omega(x)dxdy \\
&\quad + \int_{\mathscr{Y}}\int_{\mathscr{X}} p(y|x)^2 f_X(x)\omega(x)dxdy,
\end{aligned}
$$

where $\omega(\cdot)$ is a nonnegative weight function used to avoid estimating $p(y|x)$ at an x around which data are scarce. Given the observed data $\mathbf{X}$, a reasonable choice of $\omega(\cdot)$ is simply $\omega(x) = I(x \in [x_L, x_U])$, where x_L and x_U are the 2.5th and 97.5th percentile of $\mathbf{X}$, respectively, and $I(\cdot)$ is the indicator function. Noting that the third term in the above elaboration of ISE does not depend on $\boldsymbol{h}$ and thus can be ignored when minimizing ISE with respect to $\boldsymbol{h}$, they proposed the following estimator of the first two terms as a CV criterion,

$$
\text{CV}(\tilde{p}, \boldsymbol{h}, \mathbf{X}, \mathbf{Y}, \omega) = \frac{1}{n}\sum_{j=1}^{n}\omega(X_j)\int_{\mathscr{Y}} \tilde{p}_{-j}(y|X_j)^2 dy - \frac{2}{n}\sum_{j=1}^{n}\omega(X_j)\tilde{p}_{-j}(Y_j|X_j), \tag{4.1}
$$

where $\tilde{p}_{-j}(y|X_j)$ is the estimate of $p(y|X_j)$ based on data $(\mathbf{X}, \mathbf{Y})$ excluding (X_j, Y_j), for $j = 1, \ldots, n$. It is worth pointing out that, if one uses the kernel density estimator of $p(y|x)$ with the kernel associated with y, i.e., $K_2(\cdot)$, being the standard normal pdf, then the integral in (4.1) can be derived explicitly.

Clearly, the components in (4.1) that involve X_j cannot be evaluated in the presence of measurement error. To account for measurement error in $\mathbf{X}$ (but not in $\mathbf{Y}$), we first set h_2 at $\hat{h}_2 = 1.06 s_Y n^{-1/5}$ according to the normal reference rule (Silverman, 1986), where s_Y is the sample standard deviation of $\mathbf{Y}$; then we adopt the CV-SIMEX method to find an approximation of

$$
h_1 = \text{argmin}_{h_1>0}\text{CV}(\hat{p}, \boldsymbol{h}, \mathbf{X}, \mathbf{Y}, \omega), \tag{4.2}
$$

where $\hat{p}$ denotes the estimate of $p(y|x)$ based on $(\mathbf{W}, \mathbf{Y})$. Implementation of the CV-SIMEX method involves the following steps.

Step 1: Generate B sets of further contaminated data according to $\mathbf{W}_b^* = \mathbf{W} + \mathbf{U}_b^*$, where $\mathbf{U}_b^* = \{U_{b,j}^*\}_{j=1}^n$ are i.i.d. from $f_U(u)$, for $b = 1, \ldots, B$.

Step 2: Viewing $\mathbf{W}$ as the unobserved true covariate values, and $\mathbf{W}_b^*$ as the error-contaminated surrogate of $\mathbf{W}$, find

$$
h_1^* = \text{argmin}_{h_1>0}\frac{1}{B}\sum_{b=1}^{B}\text{CV}(\hat{p}_b^*, \boldsymbol{h}, \mathbf{W}, \mathbf{Y}, \tilde{\omega}), \tag{4.3}
$$

where $\hat{p}_b^*$ is the estimate of $p(y|x)$ based on $(\mathbf{W}_b^*, \mathbf{Y})$, and $\tilde{\omega}(w) = I(w \in [w_L, w_U])$, with w_L and w_U being the 2.5th and 97.5th percentile of $\mathbf{W}$, respectively.

Step 3: Generate another B sets of further contaminated data, $\mathbf{W}_b^{**} = \mathbf{W}_b^* + \mathbf{U}_b^{**}$, where $\mathbf{U}_b^{**} = \{U_{b,j}^{**}\}_{j=1}^n$ are i.i.d. from $f_U(u)$, for $b = 1, \ldots, B$.

Step 4: Viewing $\mathbf{W}_b^*$ as the unobserved true covariate values, and $\mathbf{W}_b^{**}$ as the the error-contaminated surrogate of $\mathbf{W}_b^*$, find

$$h_1^{**} = \text{argmin}_{h_1>0} \frac{1}{B} \sum_{b=1}^{B} \text{CV}(\hat{p}_b^{**}, \boldsymbol{h}, \mathbf{W}_b^*, \mathbf{Y}, \tilde{\omega}_b), \tag{4.4}$$

where $\hat{p}_b^{**}$ is the estimate of $p(y|x)$ based on $(\mathbf{W}_b^{**}, \mathbf{Y})$, and $\tilde{\omega}_b(w) = I(w \in [w_{Lb}^*, w_{Ub}^*])$, with w_{Lb}^* and w_{Ub}^* being the 2.5th and 97.5th percentile of $\mathbf{W}_b^*$, respectively.

Step 5: Set $\hat{h}_1 = h_1^{*2}/h_1^{**}$ as the final choice of h_1 for estimating $p(y|x)$ based on $(\mathbf{W}, \mathbf{Y})$.

The rationale behind the CV-SIMEX method is that the way h_1^* in (4.3) relates to h_1 in (4.2) is similar to the way h_1^{**} (4.4) relates to h_1^*. In particular, Delaigle and Hall (2008) showed that, as $\text{Var}(U) \to 0$, $\log(h_1) - \log(h_1^*) \approx \log(h_1^*) - \log(h_1^{**})$, and thus $h_1 \approx h_1^{*2}/h_1^{**}$, which suggests Step 5. We then set h_1 at $\hat{h}_1$ when estimating local modes using $(\mathbf{W}, \mathbf{Y})$. For the local constant mode estimator $\hat{y}_{M0}(x)$, $\hat{p}(y|x)$ in (4.2) is $\hat{p}(x, y)/\hat{f}_X(x)$, where $\hat{p}(x, y)$ is given in (2.3) and $\hat{f}_X(x)$ is the deconvoluting density estimator of $f_X(x)$ as in Stefanski and Carroll (1990). When considering the local linear mode estimator $\hat{y}_{M1}(x)$, $\hat{p}(y|x)$ in (4.2) is given by (2.11).

5. Empirical evidence

In this section we first present simulation studies to demonstrate the performance of the two proposed mode estimators, and compare them with the mode estimator resulting from naively applying the method in Chen *et al.* (2016) to error-contaminated data. Then we apply these methods to a real data example. In Chen *et al.* (2016), it is assumed that $h_1 = h_2$. We do not impose this constraint when implementing their method for a fair comparison with our methods.

5.1. *Simulation design*

In the simulation experiment, we consider the following two true model configurations:

(C1) $[Y|X = x] \sim 0.5N\left(m(x) - 2\sigma(x),\, 2.5^2\sigma^2(x)\right) + 0.5N\left(m(x),\, 0.5^2\sigma^2(x)\right)$, where $m(x) = x + x^2$, $\sigma(x) = 0.5 + e^{-x^2}$, $X \sim \text{Uniform}(-2, 2)$, $U \sim \text{Laplace}(0, \sigma_u/\sqrt{2})$. In this case, $p(y|x)$ is unimodal with $M(x) \approx \{m(x)\}$, $\forall x \in [-2,\, 2]$.

(C2) $[Y|X = x] \sim 0.5N\left(m_1(x),\, 0.5^2\right) + 0.5N\left(m_2(x),\, 0.5^2\right)$, where $m_1(x) = x + x^2$, $m_2(x) = m_1(x) - 6$, $X \sim \text{Uniform}(-2, 2)$, and $U \sim \text{Laplace}(0, \sigma_u/\sqrt{2})$. In this case, $p(y|x)$ is bimodal with $M(x) \approx \{m_1(x),\, m_2(x)\}$, $\forall x \in [-2,\, 2]$.

Under each true model configuration, we vary $\text{Var}(U) = \sigma_u^2$ to achieve the reliability ratio $\lambda = \text{Var}(X)/\{\text{Var}(X) + \sigma_u^2\}$ equal to 0.75, 0.85, and 0.95. Given each of the six simulation settings, we generate 500 Monte Carlo (MC) replicates, each of sample size $n = 500$, from the true model of (W, Y). When implementing our proposed methods, we choose $K_1(\cdot)$ of which the Fourier transform is $\phi_{K_1}(t) = (1-t^2)^3 I(t \in [-1, 1])$, and choose $K_2(\cdot)$ to be the standard normal pdf. For the method in Chen *et al.* (2016), both $K_1(\cdot)$ and $K_2(\cdot)$ are the standard normal pdfs.

To focus on comparing different mode estimators without being distracted by data-driven bandwidth selection, we first use approximated theoretical optimal bandwidths for each method to mitigate the confounding effect of data-driven bandwidth selection on the estimation quality. Denote by $\hat{M}(x)$ a mode set estimator generically. Given a candidate $\boldsymbol{h}$, we obtain $\hat{M}(x)$ for a sequence of grid points in $[x_L, x_U]$, $\{x_k = x_L + k\Delta\}_{k=1}^{\mathcal{M}}$, where Δ is the partition resolution, and $\mathcal{M}$ is the largest integer no larger than $(x_U - x_L)/\Delta$. Then the approximated theoretical optimal $\boldsymbol{h}$ associated with $\hat{M}(x)$ is obtained by minimizing with respect to $\boldsymbol{h}$ the empirical ISE, $\text{ISE} = \sum_{k=0}^{\mathcal{M}} \{\text{Haus}(\hat{M}(x_k), M(x_k))\}^2 \Delta$, where $\text{Haus}(S_1, S_2)$ denotes the Hausdorff distance between sets S_1 and S_2, which is defined by $\text{Haus}(S_1, S_2) = \inf\{r : S_1 \subset S_2 \oplus r, S_2 \subset S_1 \oplus r\}$, in which $S_\ell \oplus r = \{x : (\inf_{y \in S_\ell} |x - y|) \le r\}$, for $\ell = 1, 2$. Simply put, $\hat{M}(x)$ and $M(x)$ are close according to the Hausdorff distance if and only if every point in either set is close to some point in the other set, where the closeness of two points is assessed by the Euclidean distance.

For any given finite sample, besides the choice of $\boldsymbol{h}$, the starting values one uses in the mean-shift algorithm also influence $\text{Haus}(\hat{M}(x), M(x))$. A starting mode too far from the majority of the data cloud around x can cause numerical trouble in this iterative algorithm, and thus we suggest exercising great care in choosing starting values. One way that works well in our simulation study to set starting values is as follows. Given x at which $M(x)$ is of interest, define an index set $I_x = \{j : |W_j - x| < e\}$, where e is a positive small value chosen so that the number of elements in I_x is relatively large, say, 30. Then the starting values for estimating $M(x)$ via the mean-shift algorithm are set to be the percentiles of $\{Y_j : j \in I_x\}$ equally spaced between the 10th and 90th percentiles. For example, if one chooses to start with three initial values for an x, then one may set the starting values to be the 10th, 50th, and 90th percentiles of $\{Y_j : j \in I_x\}$. To avoid missing a mode, the number of starting values, denoted by N, can be slightly bigger than one's visual impression of the number of clusters of the observed data cloud.

5.2. *Simulation results*

Table 1 shows the MC average of ISE of each mode (set) estimate across 500 MC replicates and the associated standard error under each simulation setting, with $[x_L, x_U] = [-2, 2]$ and $N = 4$. In terms of both ISE and variability, our mode estimates, $\hat{M}_0(x)$ and $\hat{M}_1(x)$, outperform the naive mode estimate, de-

Table 1

Averages of ISE across 500 MC replicates using approximated theoretical optimal bandwidths. Numbers in parentheses are (10× standard errors) associated with the averages

	(C1)			(C2)		
λ	0.75	0.85	0.95	0.75	0.85	0.95
$\hat{M}_N(x)$	1.50 (0.21)	1.05 (0.15)	0.64 (0.10)	1.52 (0.29)	0.81 (0.12)	0.34 (0.05)
$\hat{M}_0(x)$	1.15 (0.17)	0.88 (0.13)	0.62 (0.10)	0.91 (0.16)	0.57 (0.09)	0.30 (0.04)
$\hat{M}_1(x)$	0.43 (0.10)	0.32 (0.07)	0.22 (0.05)	0.93 (0.31)	0.51 (0.13)	0.21 (0.04)

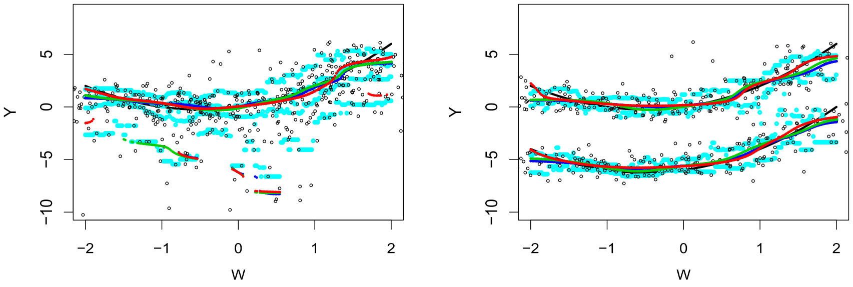
Table 2

Averages of ISE across 500 MC replicates using $\boldsymbol{h}$ chosen by the CV-SIMEX method. Numbers in parentheses are (10× standard errors) associated with the averages

	(C1)			(C2)		
λ	0.75	0.85	0.95	0.75	0.85	0.95
$\hat{M}_N(x)$	0.83 (0.19)	0.51 (0.12)	0.25 (0.06)	1.20 (0.51)	0.51 (0.11)	0.21 (0.03)
$\hat{M}_0(x)$	0.61 (0.15)	0.42 (0.11)	0.24 (0.05)	0.68 (0.37)	0.44 (0.39)	0.33 (0.40)
$\hat{M}_1(x)$	0.44 (0.12)	0.29 (0.08)	0.18 (0.07)	1.07 (0.54)	0.53 (0.29)	0.35 (0.60)

noted by $\hat{M}_N(x)$, in all six settings. And the improvement of our estimates over the naive estimate is more substantial when the error contamination is more severe (i.e., for smaller λ). In addition, the local linear estimate $\hat{M}_1(x)$ performs much better than the local constant estimate $\hat{M}_0(x)$ under (C1), while the advantage of the former is less obvious under (C2). In fact, $\hat{M}_1(x)$ deteriorates, although still outperforms $\hat{M}_N(x)$, faster than $\hat{M}_0(x)$ does as λ decreases. The boxplots of ISEs and several estimated mode curves from each method are given in Appendix H, which clearly show that there are more outliers for $\hat{M}_1(x)$ compared to $\hat{M}_0(x)$. This comparison between the two proposed mode estimators indicates that the local linear estimator may be more suitable when $p(y|x)$ is unimodal, and can be subject to more numerical instability when applying to data from a multimodal distribution. This is reminiscent of a remark in Einbeck and Tutz (2006), who recommended use the local linear mode estimator (in the absence of measurement error in their study) “only for the case of functional dependence, i.e. where the mode is unique.”

Acknowledging the fact that the approximated theoretical optimal $\boldsymbol{h}$ is not available in practice, we carry out a second round of simulation under the same six settings, with h_2 fixed at $\hat{h}_2$ for all three methods, h_1 used in our estimators chosen via the CV-SIMEX method with $B = 15$, and h_1 for the naive method chosen via naive CV as if $\mathbf{W}$ were $\mathbf{X}$ in (4.2). Table 2 shows the MC averages of ISE of the three considered estimates across 500 MC replicates, along with the corresponding standard errors, with $[x_L, x_U] = [-1.8, 1.8]$. One can see that, under (C1), the proposed estimates still outperform the naive estimate. However, this is not as clear-cut under (C2). It appears that the number of modes in an estimated mode set (for any of these methods) can be sensitive to h_2, and a smaller h_2 tends to give a bigger estimated mode set, creating more estimated local modes that can be far away from the true modes. As pictorial demonstration of these results, boxplots of these ISEs and several estimated mode curves from each method are provided in Appendix H.



(a)

(b)

Fig 1. *Estimated mode curves using one data set generated from each of (C1) (panel (a)) and (C2) (panel (b)), incorporating the data-driven bandwidth selection and data-dependent starting values (in turquoise). Each panel contains the truth $M(x)$ (black lines), $\hat{M}_N(x)$ (blue lines), $\hat{M}_0(x)$ (green lines), and $\hat{M}_1(x)$ (red lines).*

Besides the choice of h_2, as pointed out in Section 5.1, the starting values for the mean-shift algorithm also affect finite sample performance of these estimators. We use the data-dependent starting values described in Section 5.1 in the first round of simulations where the approximated theoretical optimal $\boldsymbol{h}$ is used. With the data-driven $\boldsymbol{h}$ used in the estimates as in the second round of simulations, the quality of an estimate is even more sensitive to the choice of starting values. Figure 1 depicts the estimated mode curves from the three methods based on one simulated data set under each of (C1) and (C2), with the data-dependent starting values highlighted in turquoise. From there one can see that, if one starts at a starting mode value far away from the truth, the mean-shift algorithm can fail to converge or/and result in an inferior mode estimate. To avoid the interaction effects between the data-driven bandwidth selection and the data-dependent starting values, allowing us to focus on assessing the performance of the data-driven bandwidth selection, we set the starting values to be $\{m(x) \pm 0.5\}$ under (C1), and $\{m_1(x) \pm 0.5, m_2(x) \pm 0.5\}$ under (C2) when estimating $M(x)$ in the second round of simulations that produce Table 2.

5.3. Dietary data

For illustration purposes, we consider estimating local modes of the food frequency questionnaire (FFQ) intake given one's long-term usual intake using dietary data. The data set to be analyzed contains the FFQ intake, measured as percent calories from fat (Y), and six 24-hour food recalls from 271 subjects in the Women's Interview Survey of Health. The covariate of interest, the long-term usual intake (X), cannot be observed directly. A common practice in epidemiology studies is to use data from 24-hour food recalls to construct a surrogate (W) of the true covariate. For instance, Liang and Wang (2005) used the average of two 24-hour food recalls from a subject as W and studied the mean FFQ intake conditioning on X and other error-free covariates; Wang *et al.*

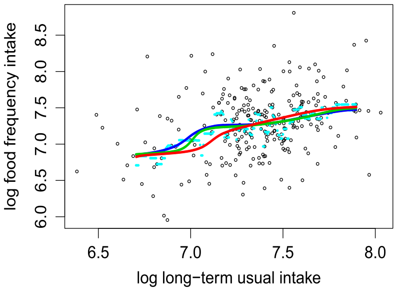


Fig 2. *The dietary data (black circles) and three estimated mode curves: the naive estimate $\hat{M}_N(x)$ (blue line), and our proposed estimates accounting for measurement error, $\hat{M}_0(x)$ (green line) and $\hat{M}_1(x)$ (red line). Turquoise points are the data-dependent starting values for the mean-shift algorithm.*

(2012) used the average of six 24-hour food recalls as W and estimated conditional quantiles of the FFQ intake. All intake values are on the log scale in these studies. We followed the construction of W in Wang *et al.* (2012), associated with which the estimated reliability ratio is 0.737.

Figure 2 presents the naive estimated mode curve and the two non-naive estimated mode curves from the two proposed methods. Both visual inspection of the scatter plot and the mean-shift algorithm seem to suggest a unimodal $p(y|x)$. Empirical evidence from simulation experiments suggest that this is the scenario where the local linear mode estimator $\hat{M}_1(x)$ can substantially improve over the naive estimator $\hat{M}_N(x)$, and the local constant mode estimator $\hat{M}_0(x)$ can also correct $\hat{M}_N(x)$ to some extent. The discrepancy between the three estimated mode curves at the lower segments in Figure 2 can be due to bias correction from the two non-naive estimates, with the correction from $\hat{M}_1(x)$ more noticeable than that from $\hat{M}_0(x)$.

6. Discussion

The study presented in this article fills in an important gap in the measurement error literature by providing mode estimation in the presence of measurement error. We rigorously study the asymptotic properties of the proposed mode estimators and develop a data-driven bandwidth selection method. This line of research leads us to more interesting open questions that we have started to investigate upon the completion of this project.

An immediate extension of this research is to allow the response prone to measurement error, with or without measurement error in covariates. An easy revision of the current estimator of the conditional (or joint) density is to replace $K_2(\cdot)$ with the corresponding deconvoluting kernel. The complication then, at least from the implementation standpoint, is that one will lose the mean-shift algorithm updating formula because the deconvoluting kernel for Y is no longer radially symmetric, the key feature of $K_2(\cdot)$ that leads to the updating formula in all existing mean-shift algorithm applications. A different mode seeking algorithm is needed in this case.

An even more involved problem arises from our development of data-driven bandwidth selection methods. In this study, we employed the CV-SIMEX method to select the bandwidth in the x-direction, h_1, with the bandwidth in the y-direction, h_2, fixed at the normal reference that only depends on the response data. We conjecture that a more sensible way to choose these two bandwidths is to choose them jointly according to some CV criterion tailored for mode estimation, as opposed to choosing them in two separate steps using two different CV criteria that are designed for density estimation. It is unclear at this point how to implement such joint selection while bearing in mind that h_1 and h_2 play very different roles, with one relating to an error-prone predictor and the other corresponding to an error-free response. Chen *et al.* (2016) assumed $h_1 = h_2 = h$ and then chose h to minimize the volume of the estimated prediction set, a statistic constructed to strive for a balance between the number of estimated local modes and the distance between the estimated mode and $\mathbf{Y}$. Following their idea, one may incorporate in the CV-SIMEX method the following CV criterion defined by (in the absence of measurement error)

$$\mathrm{CV}(\tilde{M}, \boldsymbol{h}, \mathbf{X}, \mathbf{Y}, \omega) = \frac{1}{n}\sum_{j=1}^{n} d\left(Y_j,\, \tilde{M}_{-j}(X_j)\right) \tilde{N}_{-j}(X_j)\omega(X_j),$$

where $d(x, S) = \inf_{y\in S} |x - y|$ for a set S, $\tilde{M}(x)$ represents the estimated mode set at x based on $(\mathbf{X}, \mathbf{Y})$, $\tilde{N}(x)$ is the number of distinct elements in $\tilde{M}(x)$, and $\omega(\cdot)$ is some weight function. However, theoretical justification of this CV criterion has not been established, and we did not see improvement from this bandwidth selection method over our current version of CV-SIMEX method in the simulation study (not shown here).

Besides the need for a new CV criterion for the purpose of mode estimation, we also believe that mode estimation, with or without measurement error, can benefit from using bandwidths that depend on x. We gain this intuition from simulation study with multimodal $p(y|x)$, where we encountered more difficulty in estimating modes when multiple true mode curves are steeper and close to each other. This difficulty is expected and can be illustrated by Figure 3, where two pairs of curves are shown, with the left pair flat and the right pair steep (as functions of x). Fixing at an x, the separation (in y-direction) between two curves within each pair is the same in this figure, and the variability of Y around each mode curve is also the same. But, within a given window (of a fixed width) in the x-direction, the data points (in red in Figure 3) surrounding the two steep curves are much harder to be separated into two clusters compared to the (red) data points around the two flat mode curves. And indeed in our simulation experiments (not shown here), the $\boldsymbol{h}$ that works well for identifying the flat pair of mode curves does poorly in revealing the steep pair of the mode curves, and vice versa. Hence, if the multiple mode curves associated with $p(y|x)$ show different steepness and different amount of separation along the y-direction over

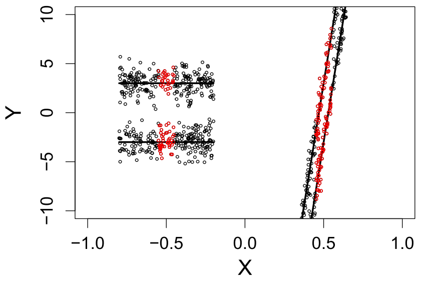


Fig 3. *Illustration of two pairs of mode curves, one pair being steeper than the other pair as functions of x. The red data points fall within a window of width 0.1 in the x-direction centering at $x = -0.5$ for the left pair of curves and $x = 0.5$ for the right pair of curves.*

different regions along the x-direction, it seems more sensible to apply different bandwidths along $\mathscr{X}$.

In light of the need for variable bandwidths, the asymptotic optimal global bandwidths provided in Corollary 3.1 may seem irrelevant for the purpose of bandwidth selection, although they are theoretically valuable because they lead to optimal rates of MISE of the proposed mode estimators. Whether or not these optimal rates reach the minimax rate (given a well-defined class of mode estimators and a class of distributions or mode functions) remains an open problem. We conjecture that the key to solving this problem lies in the minimax convergence rate of MISE associated with nonparametric density derivative estimators in the presence of measurement error, which itself is an open problem that we will tackle next.

Appendix A: Technical conditions

Here we provide technical conditions that are needed at different parts of the theoretical development for different estimators considered in the main article.

(I) Conditions on the joint probability density $p(x, y)$

(CP1) The joint density $p(x, y)$ is four times continuously differentiable with all partial derivatives bounded in absolute value by a finite positive constant C_p.

(CP2) For $(x, y) \in \mathscr{X} \times \mathscr{Y}$ where $p_y(x, y) = 0$, there exists a finite positive constant λ_2 such that $|p_{yy}(x, y)| > \lambda_2$.

Condition (CP1) is an ordinary smoothness condition. Condition (CP2) is a sharpness condition imposed on all critical points and implies no saddle points for $p(x, y)$.

(II) Conditions on kernels $K_1(t)$ and $K_2(t)$

(CK1) $K_\ell(-t)$ is an even function and $\int K_\ell(t)dt = 1$, for $\ell = 1, 2$.

(CK2) $K_2(t)$ is four times continuous differentiable with all derivatives bounded in absolute value, and $\int \{K_2^{(k)}(t)\}^2 dt < \infty$, $\int t^2 K_2^{(k)}(t)dt < \infty$, for $k = 0, 1, 2$.

(CK3) $\sup_t |\phi_{K_1(t)}/\phi_U(t/h_1)| < \infty$ and $\int |\phi^{(k)}_{K_1(t)}/\phi_U(t/h_1)|dt < \infty$, for $k = 0, 1, 2$.

(CK4) $|\phi_{K_1}(t)|_\infty < \infty$, $|\phi'_{K_1}(t)|_\infty < \infty$.

(CK5) $\int (|t|^b + |t|^{b-1})\{|\phi_{K_1}(t)| + |\phi'_{K_1}(t)|\}dt < \infty$, and $\int |t|^4 |\phi_{K_1}(t)|dt < \infty$.

(CK6) $|\phi_{K_1}(t)|$ is supported on $[-1, 1]$.

(CK7) $\phi_{K_1}(t)$ is even and real.

(CK8) $\phi^{(k)}_{K_1}(t)$ is not identically 0, for $k = 0, 1, 2$.

(CK9) The class of functions defined by

$$\mathscr{K}_4 = \cup_{k=0}^{4} \{\boldsymbol{v} \mapsto K_{U,0}\{(v_1 - x)/h_1\} K_2^{(k)}\{(v_2 - y)/h_2\} : \boldsymbol{v} = (v_1, v_2)^{\mathrm{T}} \in \mathbb{R}^2\}$$

is a VC-type class (van der Varrt and Wellner, 1996), that is, there exist A, $v > 0$, and a constant envelop τ such that $\sup_Q N(\mathscr{K}_4, \mathscr{L}^2(Q), \tau\epsilon) \leq (A/\epsilon)^v$, where $N(T, d_T, \epsilon)$ is the ϵ-covering number for a semi-metric space (T, d) and Q is any probability measure.

Define $\mu_k^{(\ell)} = \int t^k K_\ell(t)dt$ for $k = 0, 1, \ldots$. Then (CK1) implies that $\mu_k^{(\ell)} = 0$ for all odd k and $\mu_0^{(\ell)} = 1$, for $\ell = 1, 2$. Condition (CK2) and (CK5) are needed to derive the uniform error rates of the proposed mode estimator. With $k = 0$, (CK3) gives equation (1.2) in Stefanski and Carroll (1990), the assumption necessary for a well-behaved deconvoluting kernel $K_{U,0}(t)$. Condition (CK3) with $k = 1, 2$ and (CK8) are needed to derive the mean and variance of $\hat{p}_y(y|x)$. Conditions (CK4) and (the first half of) (CK5) are needed in Lemma B.4 in Delaigle *et al.* (2009), which we use to derive $\text{Var}\{\hat{g}_y(x, y)\}$ when U is ordinary smooth. Conditions (CK4) and (CK6) are needed in Lemma B.9 in Delaigle *et al.* (2009), which we evoke to derive $\text{Var}\{\hat{g}_y(x, y)\}$ when U is super smooth. Condition (CK7), along with (CU1) given later, are imposed so that $K_{U,0}(t)$ is real. Finally, (CK9) is needed to obtain the uniform consistency of the kernel-based estimators involved in the study.

(III) Conditions on measurement error U

(CU1) $\phi_U(t) \neq 0$, $\forall t$, and it is an even function.

(CU2) $|\phi'_U(t)|_\infty < \infty$.

Condition (CU1) is needed for a well-defined real-valued $K_{U,\ell}(t)$, for $\ell = 0, 1, \ldots$, and (CU2) is imposed in Lemma B.4 in Delaigle *et al.* (2009).

(IV) Conditions on the density of X, $f_X(x)$:

(CX1) $f_X(x) > 0, \forall x \in \mathscr{X}$; $f_X(x)$ is twice differentiable and $f_X^{(k)}(x)$ is bounded in absolute value by a finite positive constant C_f, for $k = 0, 1, 2$.

This condition is needed for deriving the mean and variance of $\hat{p}_y(y|x)$.

Appendix B: Relevant existing lemmas

In deriving $\mathrm{Var}\{\hat{g}_y(x, y)\}$ we evoke Lemma B.4, Lemma B.6 (for ordinary smooth U) and Lemma B.9 (for super smooth U) in Delaigle *et al.* (2009). For completeness, these lemmas are restated next.

Lemma B.4: Assume that, for $\ell = \ell_1, \ell_2$, $\|\phi_K^{(\ell)}\|_\infty < \infty$, $\|\phi_K^{(\ell+1)}\|_\infty < \infty$, $\|\phi_U'\|_\infty < \infty$, $\int(|t|^b + |t|^{b-1})\{|\phi_K^{(\ell)}| + |\phi_K^{(\ell+1)}|\}dt < \infty$, and $\int |t|^b|\phi_K^{(\ell)}|dt < \infty$, then, for a bounded function g,

$$\lim_{n\to\infty} h^{2b} \int K_{U,\ell_1}(v)K_{U,\ell_2}(v)g(x - hv)dv$$
$$= i^{-\ell_1-\ell_2}(-1)^{-\ell_2} \frac{g(x)}{c^2} \frac{1}{2\pi} \int |t|^{2b}\phi_K^{(\ell_1)}(t)\phi_K^{(\ell_2)}(t)dt.$$

Lemma B.9: Suppose that $\phi_K(t)$ is supported on $[-1, 1]$, and, for $\ell = \ell_1$ and ℓ_2, $\|\phi_K^{(\ell)}(t)\|_\infty < \infty$. Then $|\int_{-\infty}^{\infty} K_{U,\ell_1}(v)K_{U,\ell_2}(v)dv| \le Ch^{2b_2} \exp(2h^{-b}/d_2)$, where $b_2 = b_0 I(b_0 < 1/2)$.

Appendix C: Asymptotic bias of $\hat{p}_y(x, y)$

Assuming (CK3), it is shown that $E[K_{U,0}\{(W - x)/h_1\}|X] = K_1\{(X - x)/h_1\}$ (Carroll and Hall, 1988; Stefanski and Carroll, 1990). Hence, $E\{\hat{p}_y(x, y)\} = E\{\tilde{p}_y(x, y)\}$, thus $\mathrm{Bias}\{\hat{p}_y(x, y)\} = \mathrm{Bias}\{\tilde{p}_y(x, y)\}$. We next focus on deriving $E\{\tilde{p}_y(x, y)\}$.

Recall that

$$\tilde{p}_y(x, y) = \frac{1}{nh_1h_2^3} \sum_{j=1}^{n} K_1\left(\frac{X_j - x}{h_1}\right) K_2\left(\frac{Y_j - y}{h_2}\right)(Y_j - y).$$

It follows that

$$E\{\tilde{p}_y(x, y)\} = \int\int \frac{1}{h_1h_2^2} K_1\left(\frac{u - x}{h_1}\right) K_2\left(\frac{v - y}{h_2}\right)\left(\frac{v - y}{h_2}\right) p(u, v)dudv$$
$$= h_2^{-1} \int\int tK_1(s)K_2(t)p(x + h_1s, y + h_2t)dsdt. \tag{C.1}$$

Under (CP1), $p(x + h_1s, y + h_2t)$ has the third-order Taylor expansion around (x, y) as follows,

$$
\begin{aligned}
&p(x,y)+p_x(x,y)h_1s+p_y(x,y)h_2t\\
&+\frac{1}{2}\{p_{xx}(x,y)h_1^2s^2+2p_{xy}(x,y)h_1h_2st+p_{yy}(x,y)h_2^2t^2\}\\
&+\frac{1}{3!}\left\{p_{xxx}(x,y)h_1^3s^3+\frac{3!}{2}p_{xxy}(x,y)h_1^2h_2s^2t+\frac{3!}{2}p_{xyy}(x,y)h_1h_2^2st^2+p_{yyy}(x,y)h_2^3t^3\right\}\\
&+r_1h_1^3s^3+r_2h_1^2h_2s^2t+r_3h_1h_2^2st^2+r_4h_2^3t^3, \qquad \text{(C.2)}
\end{aligned}
$$

where r_1, r_2, r_3, and r_4 approach zero as $h_1, h_2 \to 0$, $p_{xx}(x,y) = (\partial^2/\partial x^2)p(x,y)$, $p_{xxy}(x,y) = (\partial^3/\partial x^2\partial y)p(x,y)$, and other partial derivatives are similarly denoted. Given (CK1), using this third-order Taylor expansion in (C.1) leads to

$$
\begin{aligned}
E\{\tilde{p}_y(x,y)\} = {} & \mu_2^{(2)}p_y(x,y) + \frac{1}{2}p_{xxy}\mu_2^{(1)}\mu_2^{(2)}h_1^2 + \frac{1}{3!}p_{yyy}(x,y)\mu_4^{(2)}h_2^2\\
& + r_2\mu_2^{(1)}\mu_2^{(2)}h_1^2 + r_4\mu_4^{(2)}h_2^2,
\end{aligned}
$$

which is equal to

$$
p_y(x,y) + 0.5\{p_{xxy}(x,y)\mu_2^{(1)}h_1^2 + p_{yyy}(x,y)h_2^2\} + r_2\mu_2^{(1)}h_1^2 + 3r_4h_2^2
$$

when $K_2(t)$ is the standard normal density, the choice we make in the main article. Hence,

$$
\text{Bias}\{\hat{p}_y(x,y)\} = 0.5\{p_{xxy}(x,y)\mu_2^{(1)}h_1^2 + p_{yyy}(x,y)h_2^2\} + o(h_1^2+h_2^2). \qquad \text{(C.3)}
$$

Setting $y = y_{\scriptscriptstyle M}$ gives Lemma 3.1 in the main article.

Appendix D: Asymptotic variance of $\hat{p}_y(x,y)$

Recall that

$$
\hat{p}_y(x,y) = \frac{1}{nh_1h_2^3}\sum_{j=1}^{n} K_{\scriptscriptstyle U,0}\left(\frac{W_j-x}{h_1}\right)K_2\left(\frac{Y_j-y}{h_2}\right)(Y_j-y).
$$

It follows that

$$
\begin{aligned}
\text{Var}\{\hat{p}_y(x,y)\} &= \frac{1}{nh_1^2h_2^6}\text{Var}\left\{K_{\scriptscriptstyle U,0}\left(\frac{W_j-x}{h_1}\right)K_2\left(\frac{Y_j-y}{h_2}\right)(Y_j-y)\right\}\\
&= \frac{1}{nh_1^2h_2^6}E\left\{K^2_{\scriptscriptstyle U,0}\left(\frac{W_j-x}{h_1}\right)K_2^2\left(\frac{Y_j-y}{h_2}\right)(Y_j-y)^2\right\}\\
&\quad - \frac{1}{nh_1^2h_2^6}\left[E\left\{K_{\scriptscriptstyle U,0}\left(\frac{W_j-x}{h_1}\right)K_2\left(\frac{Y_j-y}{h_2}\right)(Y_j-y)\right\}\right]^2. \qquad \text{(D.1)}
\end{aligned}
$$

From the bias analysis in Appendix C, we have

$$E\left\{K_{U,0}\left(\frac{W_j-x}{h_1}\right)K_2\left(\frac{Y_j-y}{h_2}\right)(Y_j-y)\right\}$$
$$= h_1h_2^3[p_y(x,y)+0.5\{p_{xxy}(x,y)\mu_2^{(1)}h_1^2+p_{yyy}(x,y)h_2^2\}+o(h_1^2+h_2^2)],$$

thus

$$\left[E\left\{K_{U,0}\left(\frac{W_j-x}{h_1}\right)K_2\left(\frac{Y_j-y}{h_2}\right)(Y_j-y)\right\}\right]^2$$
$$= h_1^2h_2^6\left[p_y^2(x,y)+p_y(x,y)\left\{p_{xxy}(x,y)\mu_2^{(1)}h_1^2+p_{yyy}(x,y)h_2^2\right\}+o(h_1^2+h_2^2)\right]. \tag{D.2}$$

This suggests that the second term in (D.1) is of order $O(n^{-1})$ if $p_y(x,y)\neq 0$, and it is of order $o\{n^{-1}(h_1^2+h_2^2)\}$ if $p_y(x,y)=0$. We next look into the first term in (D.1).

With nondifferential measurement error, the joint pdf of (W,Y) is

$$f_{W,Y}(w,y)=\int p(x,y)f_U(w-x)dx. \tag{D.3}$$

It follows that

$$E\left\{K_{U,0}^2\left(\frac{W_j-x}{h_1}\right)K_2^2\left(\frac{Y_j-y}{h_2}\right)(Y_j-y)^2\right\}$$
$$= h_2^2\int\int K_{U,0}^2\left(\frac{u-x}{h_1}\right)K_2^2\left(\frac{v-y}{h_2}\right)\left(\frac{v-y}{h_2}\right)^2 f_{W,Y}(u,v)dudv$$
$$= h_1h_2^3\int\int K_{U,0}^2(s)K_2^2(t)t^2 f_{W,Y}(x+h_1s,y+h_2t)dsdt$$
$$= h_1h_2^3\int\int K_{U,0}^2(s)K_2^2(t)t^2\int p(r,y+h_2t)f_U(x+h_1s-r)drdsdt$$
$$= h_1h_2^3\int K_{U,0}^2(s)\int f_U(x+h_1s-r)\int K_2^2(t)t^2p(r,y+h_2t)dtdrds. \tag{D.4}$$

Using the first-order Taylor expansion of $p(r,y+h_2t)$ around (r,y) in the innermost integral in (D.4) gives

$$\int K_2^2(t)t^2p(r,y+h_2t)dt=\int K_2^2(t)t^2\{p(r,y)+p_y(r,y)h_2t+O(h_2^2)\}dt$$
$$= p(r,y)\nu_2^{(2)}+p_y(r,y)\nu_3^{(2)}h_2+O(h_2^2),$$

where $\nu_k^{(2)}=\int t^kK_2^2(t)dt$, for $k=2,3$. Putting this elaboration of the innermost integral in (D.4) and using the first-order Taylor expansion of $f_U(x+h_1s-r)$ around $x-r$ in (D.4) gives

$$E\left\{K_{U,0}^2\left(\frac{W_j-x}{h_1}\right)K_2^2\left(\frac{Y_j-y}{h_2}\right)(Y_j-y)^2\right\}$$

$$= h_1 h_2^3 \int K_{U,0}^2(s) \int \{f_U(x-r) + f_U'(x-r)h_1 s + O(h_1^2)\}$$
$$\times \{p(r,y)\nu_2^{(2)} + p_y(r,y)\nu_3^{(2)} h_2 + O(h_2^2)\} dr ds.$$

This reveals the dominating term of the expectation above as

$$h_1 h_2^3 \nu_2^{(2)} \int K_{U,0}^2(s) \int f_U(x-r)p(r,y) dr ds = h_1 h_2^3 \nu_2^{(2)} f_{W,Y}(x,y) \int K_{U,0}^2(s) ds. \tag{D.5}$$

When U is ordinary smooth of order b, by Lemma B.4 in Delaigle *et al.* (2009), which is repeated under Appendix B above, (D.5) indicates that

$$E\left\{K_{U,0}^2\left(\frac{W_j - x}{h_1}\right) K_2^2\left(\frac{Y_j - y}{h_2}\right)(Y_j - y)^2\right\}$$
$$= h_1^{1-2b} h_2^3 c^{-2} \eta_0 \nu_2^{(2)} f_{W,Y}(x,y) + o(h_1^{1-2b} h_2^3), \tag{D.6}$$

where $\eta_0 = (2\pi)^{-1} \int |t|^{2b} |\phi_{K_1}(t)|^2 dt$. Because (D.6) tends to zero slower than $O(h_1^2 h_2^6)$, (D.1) is dominated by the first term there. Hence, assuming $nh_1^{1+2b} h_2^3 \to \infty$, we have

$$\text{Var}\{\hat{p}_y(x,y)\} = \frac{\eta_0 \nu_2^{(2)} f_{W,Y}(x,y)}{nh_1^{1+2b} h_2^3 c^2} + o\left(\frac{1}{nh_1^{1+2b} h_2^3}\right).$$

Setting $y = y_M$ gives the first half of Lemma 3.2 in the main article, where $\nu_2^{(2)}$ is replaced by $1/(4\sqrt{\pi})$ because $K_2(\cdot)$ is the standard normal pdf in the main article.

When U is super smooth of order b, by Lemma B.9 in Delaigle *et al.* (2009), repeated in Appendix B above, (D.5) suggests that

$$E\left\{K_{U,0}^2\left(\frac{W_j - x}{h_1}\right) K_2^2\left(\frac{Y_j - y}{h_2}\right)(Y_j - y)^2\right\}$$
$$\leq h_1^{1+2b_2} \exp(2h_1^{-b}/d_2) h_2^3 C \nu_2^{(2)} f_{W,Y}(x,y), \tag{D.7}$$

where C is a finite positive constant. Hence, if $nh_1^{1-2b_2} h_2^3 \exp(-2h_1^{-b}/d_2) \to \infty$, (D.1) suggests

$$\text{Var}\{\hat{p}_y(x,y)\} \leq \frac{1}{nh_1^2 h_2^6} E\left\{K_{U,0}^2\left(\frac{W_j - x}{h_1}\right) K_2^2\left(\frac{Y_j - y}{h_2}\right)(Y_j - y)^2\right\}$$
$$\leq \frac{\exp\left(2h_1^{-b}/d_2\right)}{nh_1^{1-2b_2} h_2^3} C \nu_2^{(2)} f_{W,Y}(x,y) + o\left\{\frac{\exp\left(2h_1^{-b}/d_2\right)}{nh_1^{1-2b_2} h_2^3}\right\}.$$

Setting $y = y_M$ and absorbing $\nu_2^{(2)} = 1/(4\sqrt{\pi})$ in C gives the second half of Lemma 3.2 in the main article.

Appendix E: Convergence rate of MISE($\hat{\boldsymbol{y}}_{M0}$)

Assuming interchangeability of integrations, the mean integrated squared error (MISE) of $\hat{y}_{M0}(x)$ can be decomposed into the sum of two parts,

$$\text{MISE}(\hat{y}_M) = \int_{\mathscr{X}} \text{Bias}^2(\hat{y}_{M0})dx + \int_{\mathscr{X}} \text{Var}(\hat{y}_M)dx.$$

By equation (17) in the main article, one has

$$\text{Bias}(\hat{y}_{M0}) = -\{p_{yy}(x, y_M)\}^{-1} E\left\{\hat{p}_y(x, y_M)\right\} + E\left\{O\left(\|\hat{p}_{yy} - p_{yy}\|_\infty\right)\hat{p}_y(x, y_M)\right\}, \tag{E.1}$$

$$\text{Var}(\hat{y}_{M0}) = \{p_{yy}(x, y_M)\}^2 \text{Var}\left\{\hat{p}_y(x, y_M)\right\} + o(1). \tag{E.2}$$

To derive the integrated squared bias, using (C.3) in (E.1), one has that, provided that $\|\hat{p}_{yy} - p_{yy}\|_\infty$ tends to zero,

$$\begin{aligned}\text{Bias}^2(\hat{y}_M) = 0.25p_{yy}^{-2}(x, y_M)\Big\{ &\left(\mu_2^{(1)}\right)^2 h_1^4 p_{xxy}^2(x, y_M) + h_2^4 p_{yyy}^2(x, y_M) \\ &+ 2\mu_2^{(1)} h_1^2 h_2^2 p_{xxy}(x, y_M) p_{yyy}(x, y_M)\Big\} + o\left\{\left(h_1^2 + h_2^2\right)^2\right\}.\end{aligned}$$

Under (CP2), and assuming $p_{xxy}(x, y_M)$ and $p_{yyy}(x, y_M)$ square integrable, one has

$$\int_{\mathscr{X}} \text{Bias}^2(\hat{y}_{M0})dx = O\left\{\left(h_1^2 + h_2^2\right)^2\right\}.$$

To derive the integrated variance, using the variance analysis Appendix D in (E.2), one has that, when U is ordinary smooth,

$$\int_{x\in\mathscr{X}} \text{Var}(\hat{y}_{M0})dx = O\left(\frac{1}{nh_1^{1+2b}h_2^3}\right);$$

and, when U is super smooth,

$$\int_{x\in\mathscr{X}} \text{Var}(\hat{y}_M)dx = O\left\{\frac{\exp\left(2h_1^{-b}/d_2\right)}{nh_1^{1-2b_2}h_2^3}\right\}.$$

Putting the above integrated squared bias of $\hat{y}_{M0}(x)$ and integrated variance of $\hat{y}_{M0}(x)$ together gives Theorem 3.2 in the main article.

Appendix F: Asymptotic bias of $\hat{\boldsymbol{p}}_{\boldsymbol{y}}(\boldsymbol{y}_M|\boldsymbol{x})$

F.1. Outline of deriving $E\{\hat{\boldsymbol{p}}_{\boldsymbol{y}}(\boldsymbol{y}_M|\boldsymbol{x})\}$

Recall that $\hat{p}_y(y|x) = \boldsymbol{e}_1^{\mathrm{T}}\hat{\mathbf{S}}_n^{-1}(x)\hat{\mathbf{T}}'_n(x, y)$. Denote by $[\hat{S}_n^{0,0}(x),\ \hat{S}_n^{0,1}(x)]$ the first row of $\hat{\mathbf{S}}_n^{-1}(x)$, then one has

$$\hat{p}_y(y|x) = \sum_{\ell=0}^{1} \hat{S}_n^{0,\ell}(x)\hat{T}'_{n,\ell}(x,y). \tag{F.1}$$

The derivations of the dominating bias of $\hat{p}_y(y|x)$ involve two tasks. First, revealing the dominating terms in $\hat{S}_n^{0,\ell}(x)$. Under (CX1), (CU1), and (CK1), (CK3)–(CK5), (CK8), Delaigle *et al.* (2009) showed that

$$\hat{\mathbf{S}}^{-1} = \mathbf{S}^{-1}f_X^{-1}(x) - h_1\mathbf{S}^{-1}\widetilde{\mathbf{S}}\mathbf{S}^{-1}f'_X(x)f_X^{-2}(x) + O_P(h_1^2),$$

where

$$\mathbf{S} = \begin{bmatrix} \mu_0^{(1)} & \mu_1^{(1)} \\ \mu_1^{(1)} & \mu_2^{(1)} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \mu_2^{(1)} \end{bmatrix}, \qquad \widetilde{\mathbf{S}} = \begin{bmatrix} \mu_1^{(1)} & \mu_2^{(1)} \\ \mu_2^{(1)} & \mu_3^{(1)} \end{bmatrix} = \begin{bmatrix} 0 & \mu_2^{(1)} \\ \mu_2^{(1)} & 0 \end{bmatrix}.$$

Elaborating the above result yields

$$\hat{S}_n^{0,0}(x) = f_X^{-1}(x) + O_P(h_1^2),\ \hat{S}_n^{0,1}(x) = -h_1 f'_X(x)f_X^{-2}(x) + O_P(h_1^2). \tag{F.2}$$

This completes the first task.

The second task is to reveal the dominating terms in $\hat{T}'_{n,\ell}(x,y)$ according to the following decomposition,

$$\hat{T}'_{n,\ell}(x,y) = E\left\{\hat{T}'_{n,\ell}(x,y)\right\} + O_P\left[\sqrt{\mathrm{Var}\left\{\hat{T}'_{n,\ell}(x,y)\right\}}\right]. \tag{F.3}$$

Next, we first look into $E\{\hat{T}'_{n,\ell}(x,y)\}$, then we study $\mathrm{Var}\{\hat{T}'_{n,\ell}(x,y)\}$, and show that the former dominates that latter under certain conditions.

F.2. Deriving $E\{\hat{T}'_{n,\ell}(x,y)\}$

Because

$$\hat{T}'_{n,\ell}(x,y) = \frac{1}{nh_1h_2^2}\sum_{j=1}^{n} K_{U,\ell}\left(\frac{W_j - x}{h_1}\right)K_2\left(\frac{Y_j - y}{h_2}\right)\left(\frac{Y_j - y}{h_2}\right), \tag{F.4}$$

and

$$E\left\{K_{U,\ell}\left(\frac{W_j - x}{h_1}\right)\middle| X_j\right\} = \left(\frac{X_j - x}{h_1}\right)^{\ell} K_1\left(\frac{X_j - x}{h_1}\right),$$

we have

$$\begin{aligned} &E\{\hat{T}'_{n,\ell}(x,y)\} \\ &\quad = \frac{1}{h_1h_2^2}E\left\{K_{U,\,\ell}\left(\frac{W_j - x}{h_1}\right)K_2\left(\frac{Y_j - y}{h_2}\right)\left(\frac{Y_j - y}{h_2}\right)\right\} \\ &\quad = \frac{1}{h_1h_2^2}E\left\{\left(\frac{X_j - x}{h_1}\right)^{\ell}K_1\left(\frac{X_j - x}{h_1}\right)K_2\left(\frac{Y_j - y}{h_2}\right)\left(\frac{Y_j - y}{h_2}\right)\right\} \end{aligned} \tag{F.5}$$

$$
\begin{aligned}
&= \frac{1}{h_1 h_2^2} \int\int \left(\frac{u-x}{h_1}\right)^{\ell} K_1\left(\frac{u-x}{h_1}\right) K_2\left(\frac{v-y}{h_2}\right)\left(\frac{v-y}{h_2}\right) p(u,v)dudv \\
&= h_2^{-1} \int\int s^{\ell} t K_1(s) K_2(t) p(x+h_1 s,\, y+h_2 t) dsdt.
\end{aligned}
$$

Inserting (C.2) in the above integrand leads to

$$
\begin{aligned}
&E\{\hat{T}'_{n,\ell}(x,y)\} \\
&\quad = h_2^{-1}\int\int s^{\ell} t K_1(s)K_2(t)\Big[p(x,y)+p_x(x,y)h_1 s+p_y(x,y)h_2 t \\
&\qquad + \frac{1}{2}\left\{p_{xx}(x,y)h_1^2 s^2+2p_{xy}(x,y)h_1h_2 st+p_{yy}(x,y)h_2^2t^2\right\} \\
&\qquad + \frac{1}{3!}\Big\{p_{xxx}(x,y)h_1^3s^3+\frac{3!}{2}p_{xxy}(x,y)h_1^2h_2s^2t \\
&\qquad\quad + \frac{3!}{2}p_{xyy}(x,y)h_1h_2^2st^2+p_{yyy}(x,y)h_2^3t^3\Big\} \\
&\qquad + O\left\{r_4(\boldsymbol{h})\right\}\Big]dsdt,
\end{aligned}
\tag{F.6}
$$

where $r_4(\boldsymbol{h}) = \sum_{r_1,r_2=0,\ldots,4}^{r_1+r_2=4} h_1^{r_1} h_2^{r_2}$, in which $\boldsymbol{h} = (h_1,\, h_2)^{\mathrm{T}}$.

Focusing on the case with $y = y_M$, noting that $p_y(x, y_M) = 0$ and $\mu_k^{(\ell)} = 0$ when k is odd, for $\ell = 1, 2$, (F.6) reduces to

$$
\begin{cases}
E\{\hat{T}'_{n,0}(x,y_M)\} = 0.5\left\{p_{xxy}(x,y_M)\mu_2^{(1)}h_1^2+p_{yyy}(x,y_M)h_2^2\right\}+O\left\{h_2^{-1}r_4(\boldsymbol{h})\right\}, \\
E\{\hat{T}'_{n,1}(x,y_M)\} = p_{xy}(x,y_M)\mu_2^{(1)}h_1+O\left\{h_2^{-1}r_4(\boldsymbol{h})\right\}.
\end{cases}
\tag{F.7}
$$

F.3. Deriving Var$\{\hat{T}'_{n,\ell}(x,y)\}$

By (F.4), we have

$$
\begin{aligned}
&\mathrm{Var}\{\hat{T}'_{n,\ell}(x,y)\} \\
&= \frac{1}{nh_1^2h_2^4}\mathrm{Var}\left\{K_{U,\ell}\left(\frac{W_j-x}{h_1}\right)K_2\left(\frac{Y_j-y}{h_2}\right)\left(\frac{Y_j-y}{h_2}\right)\right\} \\
&\le \frac{1}{nh_1^2h_2^4}E\left\{K_{U,\ell}^2\left(\frac{W_j-x}{h_1}\right)K_2^2\left(\frac{Y_j-y}{h_2}\right)\left(\frac{Y_j-y}{h_2}\right)^2\right\} \qquad \text{(F.8)} \\
&= \frac{1}{nh_1^2h_2^4}\iint K_{U,\ell}^2\left(\frac{w-x}{h_1}\right)K_2^2\left(\frac{v-y}{h_2}\right)\left(\frac{v-y}{h_2}\right)^2 f_{W,\,Y}(w,v)dwdv \\
&= \frac{1}{nh_1^2h_2^4}\iiint K_{U,\ell}^2\left(\frac{w-x}{h_1}\right)K_2^2\left(\frac{v-y}{h_2}\right)\left(\frac{v-y}{h_2}\right)^2 f_U(w-u)p(u,v)dudwdv \\
&= \frac{1}{nh_1h_2^3}\iiint K_{U,\ell}^2(s)K_2^2(t)t^2 f_U(x+h_1s-u)p(u,y+h_2t)dudsdt
\end{aligned}
$$

$$= \frac{1}{nh_1h_2^3}\int\left\{\int K_{U,\ell}^2(s)f_U(x+h_1s-u)ds\right\}\left\{\int t^2K_2^2(t)p(u,y+h_2t)dt\right\}du. \tag{F.9}$$

Next, we elaborate the two inner integrals, one w.r.t. s and the other w.r.t. t, in (F.9).

For the inner integral w.r.t. t, using the second-order Taylor expansion of $p(u,y+h_2t)$ around (u,y), one has

$$\begin{aligned}&\int K_2^2(t)t^2p(u,y+h_2t)dt\\&\quad=\int K_2^2(t)t^2\left\{p(u,y)+p_y(u,y)h_2t+0.5p_{yy}(u,y)h_2^2t^2+O(h_2^3)\right\}dt.\end{aligned}$$

Setting $y=y_M$, the above gives

$$\begin{aligned}&\int K_2^2(t)t^2p(u,y_M+h_2t)dt &\text{(F.10)}\\&\quad=\int K_2^2(t)t^2\left\{p(u,y_M)+0.5p_{yy}(u,y_M)h_2^2t^2+O(h_2^3)\right\}dt\\&\quad=p(u,y_M)\nu_2^{(2)}+0.5p_{yy}(u,y_M)\nu_4^{(2)}h_2^2+O(h_2^3). &\text{(F.11)}\end{aligned}$$

When it comes to the inner integral w.r.t s in (F.9), one shall distinguish between ordinary smooth U and super smooth U. If U is ordinary smooth of order b, under conditions (CK4) and (CK5), Lemma B.4 in Delaigle *et al.* (2009) implies that,

$$\int K_{U,\ell}^2(s)f_U(x+h_1s-u)ds=h_1^{-2b}c^{-2}\eta_\ell f_U(x-u)+o(h_1^{-2b}), \tag{F.12}$$

where $\eta_\ell=(2\pi)^{-1}\int|t|^{2b}\{\phi_{K_1}^{(\ell)}(t)\}^2dt$, for $\ell=0,1$. By (F.11) and (F.12), (F.9) is equal to

$$\begin{aligned}&\frac{\eta_\ell}{nh_1h_2^3c^2}\int\left\{h_1^{-2b}f_U(x-u)+o(h_1^{-2b})\right\}\\&\times\{p(u,y_M)\nu_2^{(2)}+0.5p_{yy}(u,y_M)\nu_4^{(2)}h_2^2+O(h^3)\}du\\&\quad=\frac{\eta_\ell}{nh_1h_2^3c^2}\Big[\{p(\cdot,y_M)*f_U\}(x)\nu_2^{(2)}h_1^{-2b}+0.5\{p_{yy}(\cdot,y_M)*f_U\}(x)\nu_4^{(2)}h_1^{-2b}h_2^2\\&\qquad+o(h_1^{-2b})\Big], &\text{(F.13)}\end{aligned}$$

where "$*$" is the convolution operator, that is, $\{g(\cdot,y)*f_U\}(x)=\int f_U(x-u)g(u,y)du$. Because the dominating term within the square brackets in (F.13) is

$$\{p(\cdot,y_M)*f_U\}(x)\nu_2^{(2)}h_1^{-2b},$$

which is equal to $f_{W,Y}(x,y_M)\nu_2^{(2)}h_1^{-2b}$ by (D.3), (F.9) indicates that, for $\ell=0,1$,

$$\text{Var}\{\hat{T}'_{n,\ell}(x,y_M)\}\le\frac{f_{W,Y}(x,y_M)\nu_2^{(2)}\eta_\ell}{nh_1^{1+2b}h_2^3c^2}+o\left(\frac{1}{nh_1^{1+2b}h_2^3}\right). \tag{F.14}$$

It follows that, when U is ordinary smooth, the first term in (F.3) dominates the second term there if $(nh_1^{1+2b}h_2^3)^{-1/2} = O\{h_2^{-1}r_4(\boldsymbol{h})\}$.

If U is super smooth of order b, by Lemma B.9 in Delaigle *et al.* (2009), under conditions (CK4) and (CK6), one has

$$\int K_{U,\ell}^2(t)dt \le Ch_1^{2b_2}\exp(2h_1^{-b}/d_2),$$

where $b_2 = b_0 I(b_0 < 0.5)$, and C is some positive finite constant. Hence,

$$\begin{aligned}
&\operatorname{Var}\{\hat{T}'_{n,\ell}(x, y_M)\}\\
&\quad\le \frac{1}{nh_1h_2^3}\int K_{U,\ell}^2(s)\int f_U(x+sh_1-u)\{p(u,y_M)\nu_2^{(2)} + 0.5p_{yy}(u,y_M)\nu_4^{(2)}h_2^2\\
&\qquad + O(h_2^3)\}duds\\
&\quad\le \frac{1}{nh_1h_2^3}\int K_{U,\ell}^2(s)\int f_U(x+sh_1-u)\{C_p\nu_2^{(2)} + O(h_2^2)\}duds, \text{ by (CP1)},\\
&\quad\le \frac{\exp(2h_1^{-b}/d_2)}{nh_1^{1-2b_2}h_2^3}\{CC_p\nu_2^{(2)} + O(h_2^2)\}\\
&\quad= \frac{\exp(2h_1^{-b}/d_2)CC_p\nu_2^{(2)}}{nh_1^{1-2b_2}h_2^3} + O\left\{\frac{\exp(2h_1^{-b}/d_2)}{nh_1^{1-2b_2}h_2}\right\}.
\end{aligned} \tag{F.15}$$

Hence, when U is super smooth, the first term in (F.3) dominates the second term there if $\{\exp(2h_1^{-b}/d_2)/(nh_1^{1-2b_2}h_2^3)\}^{1/2} = O\{h_2^{-1}r_4(\boldsymbol{h})\}$.

F.4. Concluding Lemma 3.3

Based on the mean and variance analysis of $\hat{T}'_{n,\ell}(x, y_M)$ in Sections F.2 and F.3, we now reach the conclusion that, if $(nh_1^{1+2b}h_2^3)^{-1/2} = O\{h_2^{-1}r_4(\boldsymbol{h})\}$ when U ordinary smooth, or if $\exp(h_1^{-b}/d_2)/\sqrt{nh_1^{1-2b_2}h_2^3} = O\{h_2^{-1}r_4(\boldsymbol{h})\}$ when U is super smooth, then

$$\begin{cases}
\hat{T}'_{n,0}(x,y_M) = 0.5\left\{p_{xxy}(x,y_M)\mu_2^{(1)}h_1^2 + p_{yyy}(x,y_M)h_2^2\right\} + O_P\left\{h_2^{-1}r_4(\boldsymbol{h})\right\},\\
\hat{T}'_{n,1}(x,y_M) = p_{xy}(x,y_M)\mu_2^{(1)}h_1 + O_P\left\{h_2^{-1}r_4(\boldsymbol{h})\right\}.
\end{cases} \tag{F.16}$$

This completes the second task stated in Section F.1 in order to derive $E\{\hat{p}_y(y_M|x)\}$.

Using (F.2) and (F.16) in (F.1), we have

$$\begin{aligned}
\hat{p}_y(y_M|x) &= \hat{S}_n^{0,0}(x)\hat{T}'_{n,0}(x,y_M) + \hat{S}_n^{0,1}(x)\hat{T}'_{n,1}(x,y_M)\\
&= f_X^{-1}(x)\left[0.5\left\{p_{xxy}(x,y_M)\mu_2^{(1)}h_1^2 + p_{yyy}(x,y_M)h_2^2\right\}\right.\\
&\quad \left. - f_X^{-1}(x)f'_X(x)p_{xy}(x,y_M)\mu_2^{(1)}h_1^2\right] + O_P\left\{h_2^{-1}r_4(\boldsymbol{h})\right\}.
\end{aligned} \tag{F.17}$$

Because the dominating term in (F.17) is a non-random quantity, this dominating term is also the dominating bias of $\hat{p}_y(y_{\scriptscriptstyle M}|x)$. This proves Lemma 3.3 in the main article.

Appendix G: Asymptotic variance of $\hat{p}_y(y_{\scriptscriptstyle M}|x)$

By (F.2),

$$
\begin{aligned}
&\hat{p}_y(y_{\scriptscriptstyle M}|x)\\
&\quad=\sum_{\ell=0}^{1}\hat{S}_n^{0,\ell}(x)\hat{T}'_{n,\ell}(x,y_{\scriptscriptstyle M})\\
&\quad=\left\{f_{\scriptscriptstyle X}^{-1}(x)+O_{\scriptscriptstyle P}(h_1^2)\right\}\hat{T}'_{n,0}(x,y_{\scriptscriptstyle M})+\left\{-h_1f'_{\scriptscriptstyle X}(x)f_{\scriptscriptstyle X}^{-2}(x)+O_{\scriptscriptstyle P}(h_1^2)\right\}\hat{T}'_{n,1}(x,y_{\scriptscriptstyle M})\\
&\quad=\frac{1}{nh_1h_2^2}\sum_{j=1}^{n}\left[\left\{f_{\scriptscriptstyle X}^{-1}(x)+O_{\scriptscriptstyle P}(h_1^2)\right\}K_{\scriptscriptstyle U,0}\left(\frac{W_j-x}{h_1}\right)K_2\left(\frac{Y_j-y_{\scriptscriptstyle M}}{h_2}\right)\left(\frac{Y_j-y_{\scriptscriptstyle M}}{h_2}\right)\right.\\
&\qquad\left.+\left\{-h_1f'_{\scriptscriptstyle X}(x)f_{\scriptscriptstyle X}^{-2}(x)+O_{\scriptscriptstyle P}(h_1^2)\right\}K_{\scriptscriptstyle U,1}\left(\frac{W_j-x}{h_1}\right)K_2\left(\frac{Y_j-y_{\scriptscriptstyle M}}{h_2}\right)\left(\frac{Y_j-y_{\scriptscriptstyle M}}{h_2}\right)\right].
\end{aligned}
$$

Extracting the dominating terms in the above expression reveals that, to find the dominating variance of $\hat{p}_y(y_{\scriptscriptstyle M}|x)$, it suffices to look into the variance of

$$
\begin{aligned}
&\frac{1}{nh_1h_2^2}\sum_{j=1}^{n}\left\{f_{\scriptscriptstyle X}^{-1}(x)K_{\scriptscriptstyle U,0}\left(\frac{W_j-x}{h_1}\right)K_2\left(\frac{Y_j-y_{\scriptscriptstyle M}}{h_2}\right)\left(\frac{Y_j-y_{\scriptscriptstyle M}}{h_2}\right)\right.\\
&\qquad\left.-h_1f'_{\scriptscriptstyle X}(x)f_{\scriptscriptstyle X}^{-2}(x)K_{\scriptscriptstyle U,1}\left(\frac{W_j-x}{h_1}\right)K_2\left(\frac{Y_j-y_{\scriptscriptstyle M}}{h_2}\right)\left(\frac{Y_j-y_{\scriptscriptstyle M}}{h_2}\right)\right\}. \qquad (G.1)
\end{aligned}
$$

This leads us to study the following variance and covariance,

$$
\begin{aligned}
&\mathrm{Var}\left\{K_{\scriptscriptstyle U,\ell}\left(\frac{W_j-x}{h_1}\right)K_2\left(\frac{Y_j-y_{\scriptscriptstyle M}}{h_2}\right)\left(\frac{Y_j-y_{\scriptscriptstyle M}}{h_2}\right)\right\},\ \text{for } \ell=0,1, \qquad (G.2)\\
&\mathrm{Cov}\left\{K_{\scriptscriptstyle U,0}\left(\frac{W_j-x}{h_1}\right)K_2\left(\frac{Y_j-y_{\scriptscriptstyle M}}{h_2}\right)\left(\frac{Y_j-y_{\scriptscriptstyle M}}{h_2}\right),\right.\\
&\qquad\left.K_{\scriptscriptstyle U,1}\left(\frac{W_j-x}{h_1}\right)K_2\left(\frac{Y_j-y_{\scriptscriptstyle M}}{h_2}\right)\left(\frac{Y_j-y_{\scriptscriptstyle M}}{h_2}\right)\right\}. \qquad (G.3)
\end{aligned}
$$

G.1. Deriving the variance in (G.2)

For $\ell=0,1$,

$$
\begin{aligned}
&\mathrm{Var}\left\{K_{\scriptscriptstyle U,\ell}\left(\frac{W_j-x}{h_1}\right)K_2\left(\frac{Y_j-y_{\scriptscriptstyle M}}{h_2}\right)\left(\frac{Y_j-y_{\scriptscriptstyle M}}{h_2}\right)\right\}\\
&=E\left\{K_{\scriptscriptstyle U,\ell}^2\left(\frac{W_j-x}{h_1}\right)K_2^2\left(\frac{Y_j-y_{\scriptscriptstyle M}}{h_2}\right)\left(\frac{Y_j-y_{\scriptscriptstyle M}}{h_2}\right)^2\right\} \qquad (G.4)
\end{aligned}
$$

$$
- \left[E \left\{ K_{U,\ell} \left(\frac{W_j - x}{h_1} \right) K_2 \left(\frac{Y_j - y_M}{h_2} \right) \left(\frac{Y_j - y_M}{h_2} \right) \right\} \right]^2 . \tag{G.5}
$$

The expectation in (G.4) is considered in Section F.3 (see (F.8)). From there, we have shown that, if U is ordinary smooth of order b, then (relating to (F.14))

$$
\begin{aligned}
& E \left\{ K^2_{U,\ell} \left(\frac{W_j - x}{h_1} \right) K^2_2 \left(\frac{Y_j - y_M}{h_2} \right) \left(\frac{Y_j - y_M}{h_2} \right)^2 \right\} \\
& \quad = h_1^{1-2b} h_2 c^{-2} \nu_2^{(2)} \eta_\ell f_{W,\,Y}(x, y_M) + o\left(h_1^{1-2b} h_2 \right) ;
\end{aligned} \tag{G.6}
$$

and, if U is super smooth of order b, then (relating to (F.15))

$$
\begin{aligned}
& E \left\{ K^2_{U,\ell} \left(\frac{W_j - x}{h_1} \right) K^2_2 \left(\frac{Y_j - y_M}{h_2} \right) \left(\frac{Y_j - y_M}{h_2} \right)^2 \right\} \\
& \quad \le h_1^{1+2b_2} h_2 \exp\left(2h_1^{-b}/d_2 \right) C C_p \nu_2^{(2)} + O\left\{ h_1^{1-2b_2} h_2^3 \exp\left(2h_1^{-b}/d_2 \right) \right\} .
\end{aligned} \tag{G.7}
$$

The expectation in (G.5) is considered in Section F.2 (see (F.5)). By (F.7), this expectation is of order $O\{h_1 h_2^2 (h_1^2 + h_2^2)\}$ when $\ell = 0$, and it is of order $O(h_1^2 h_2^2)$ when $\ell = 1$. Hence, for both $\ell = 0, 1$, (G.5) tends to zero faster than (G.6) and (G.7).

It follows that that, for ordinary smooth U,

$$
\begin{aligned}
& \operatorname{Var} \left\{ K_{U,\ell} \left(\frac{W_j - x}{h_1} \right) K_2 \left(\frac{Y_j - y_M}{h_2} \right) \left(\frac{Y_j - y_M}{h_2} \right) \right\} \\
& \quad = h_1^{1-2b} h_2 \nu_2^{(2)} c^{-2} \eta_\ell f_{W,\,Y}(x, y_M) + o\left(h_1^{1-2b} h_2 \right) ;
\end{aligned} \tag{G.8}
$$

and, for super smooth U,

$$
\begin{aligned}
& \operatorname{Var} \left\{ K_{U,\ell} \left(\frac{W_j - x}{h_1} \right) K_2 \left(\frac{Y_j - y_M}{h_2} \right) \left(\frac{Y_j - y_M}{h_2} \right) \right\} \\
& \quad \le h_1^{1+2b_2} h_2 \exp\left(2h_1^{-b}/d_2 \right) C C_p \nu_2^{(2)} + O\left\{ h_1^{1-2b_2} h_2^3 \exp\left(2h_1^{-b}/d_2 \right) \right\} .
\end{aligned} \tag{G.9}
$$

G.2. Deriving the covariance in (G.3)

The covariance in (G.3) is equal to

$$
E \left\{ K_{U,0} \left(\frac{W_j - x}{h_1} \right) K_{U,1} \left(\frac{W_j - x}{h_1} \right) K^2_2 \left(\frac{Y_j - y_M}{h_2} \right) \left(\frac{Y_j - y_M}{h_2} \right)^2 \right\} \tag{G.10}
$$

$$
\begin{aligned}
& - E \left\{ K_{U,0} \left(\frac{W_j - x}{h_1} \right) K_2 \left(\frac{Y_j - y_M}{h_2} \right) \left(\frac{Y_j - y_M}{h_2} \right) \right\} \\
& \quad \times E \left\{ K_{U,1} \left(\frac{W_j - x}{h_1} \right) K_2 \left(\frac{Y_j - y_M}{h_2} \right) \left(\frac{Y_j - y_M}{h_2} \right) \right\} .
\end{aligned} \tag{G.11}
$$

In Section F.2, we have shown that the product in (G.11) is of order $O(h_1^3 h_2^4)$, which tends to zero faster than (G.10), as to be revealed next.

When U is ordinary smooth, by Lemma B.4 in Delaigle *et al.* (2009),

$$E\left\{K_{U,0}\left(\frac{W_j-x}{h_1}\right)K_{U,1}\left(\frac{W_j-x}{h_1}\right)K_2^2\left(\frac{Y_j-y_M}{h_2}\right)\left(\frac{Y_j-y_M}{h_2}\right)^2\right\}$$
$$=h_1^{1-2b}h_2\nu_2^{(2)}c^{-2}\eta_{01}f_{W,Y}(x,y_M)+o\left(h_1^{1-2b}h_2\right),$$

where $\eta_{01}=\int|t|^{2b}\phi_{K_1}(t)\phi'_{K_1}(t)dt$. When U is super smooth, by Lemma B.9 in Delaigle *et al.* (2009),

$$E\left\{K_{U,0}\left(\frac{W_j-x}{h_1}\right)K_{U,1}\left(\frac{W_j-x}{h_1}\right)K_2^2\left(\frac{Y_j-y_M}{h_2}\right)\left(\frac{Y_j-y_M}{h_2}\right)^2\right\}$$
$$\leq h_1^{1+2b_2}h_2\exp\left(2h_1^{-b}/d_2\right)CC_p\nu_2^{(2)}+O\left\{h_1^{1+2b_2}h_2^3\exp\left(2h_1^{-b}/d_2\right)\right\}.$$

Hence, the covariance in (G.3) and variances in (G.2) are of the same order.

G.3. Concluding Lemma 3.4

Since (G.2) and (G.3) are of the same order, the variance of (G.1) with $y=y_M$ is dominated by

$$\frac{1}{nh_1^2h_2^4}\text{Var}\left\{K_{U,0}\left(\frac{W_j-x}{h_1}\right)K_2\left(\frac{Y_j-y_M}{h_2}\right)\left(\frac{Y_j-y_M}{h_2}\right)\right\}.$$

Hence, for ordinary smooth U, by (G.8) with $\ell=0$,

$$\text{Var}\left\{\hat{p}_y(y_M|x)\right\}=\frac{\eta_0\nu_2^{(2)}f_{W,Y}(x,y_M)}{nh_1^{1+2b}h_2^3c^2f_X^2(x)}+o\left(\frac{1}{nh_1^{1+2b}h_2^3}\right);$$

and, for super smooth U, by (G.9),

$$\text{Var}\left\{\hat{p}_y(y_M|x)\right\}\leq\frac{\exp\left(2h_1^{-b}/d_2\right)CC_p\nu_2^{(2)}}{nh_1^{1-2b_2}h_2^3f_X^2(x)}+O\left\{\frac{\exp\left(2h_1^{-b}/d_2\right)}{nh_1^{1-2b_2}h_2}\right\}.$$

This proves Lemma 3.4 in the main article.

Appendix H: Pictorial demonstrations of simulation results

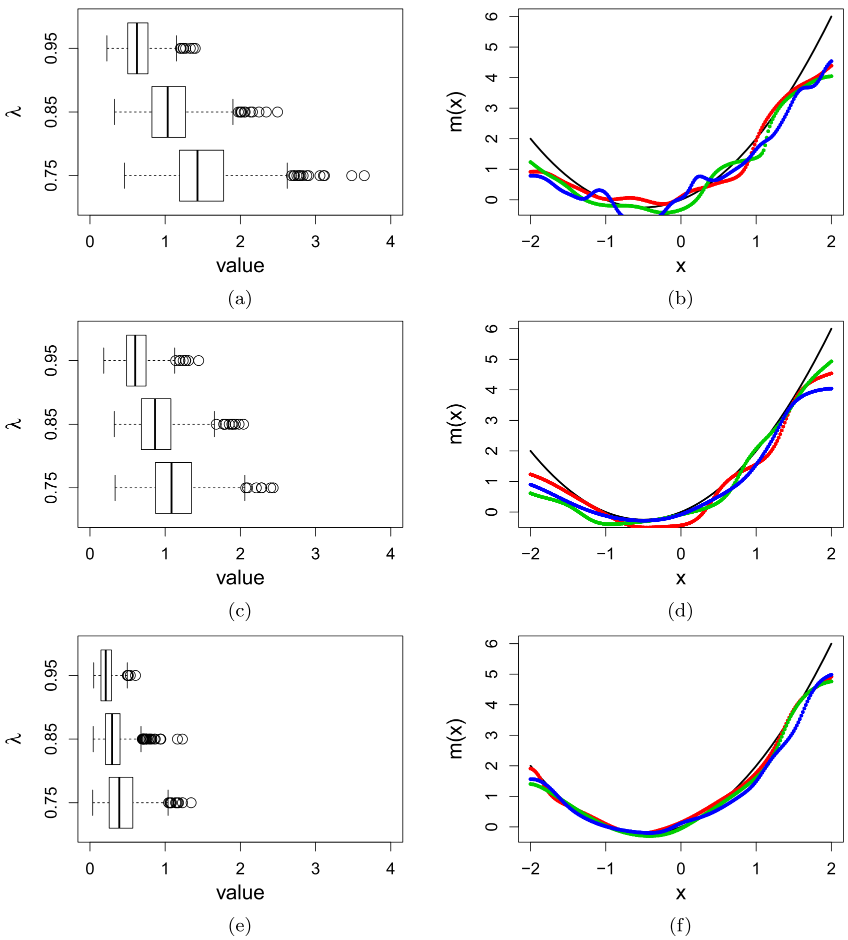
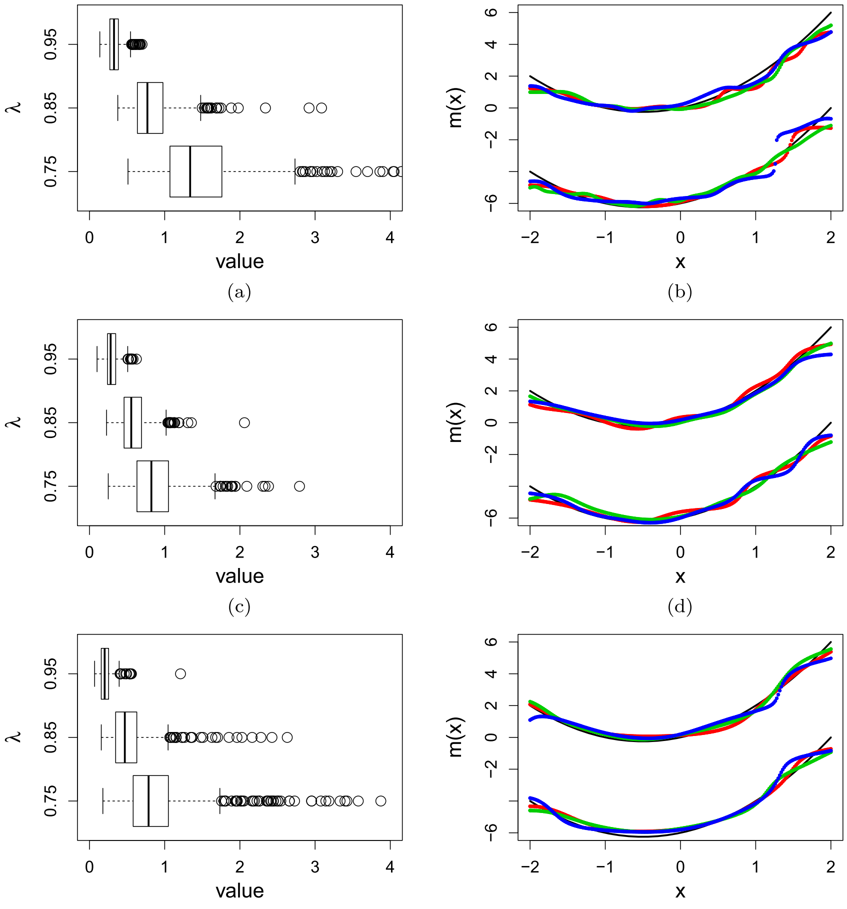


Fig H.1. *Results under (C1) using approximated theoretical optimal bandwidths. Panels (a), (c), and (e): boxplots of ISEs versus* λ *for* $\hat{M}_N(x)$, $\hat{M}_0(x)$, *and* $\hat{M}_1(x)$, *respectively. Panels (b), (d), and (f): estimated mode curves,* $\hat{M}_N(x)$, $\hat{M}_0(x)$, *and* $\hat{M}_1(x)$, *respectively, when* $\lambda = 0.85$. *In each panel with estimated mode curves associated with an estimator, the black line depicts the true mode curve, the red, green, and blue lines are three estimated mode curves from the same method that yield ISE being the first, second, and third quantiles among the 500 ISEs for that method from the simulation, respectively.*

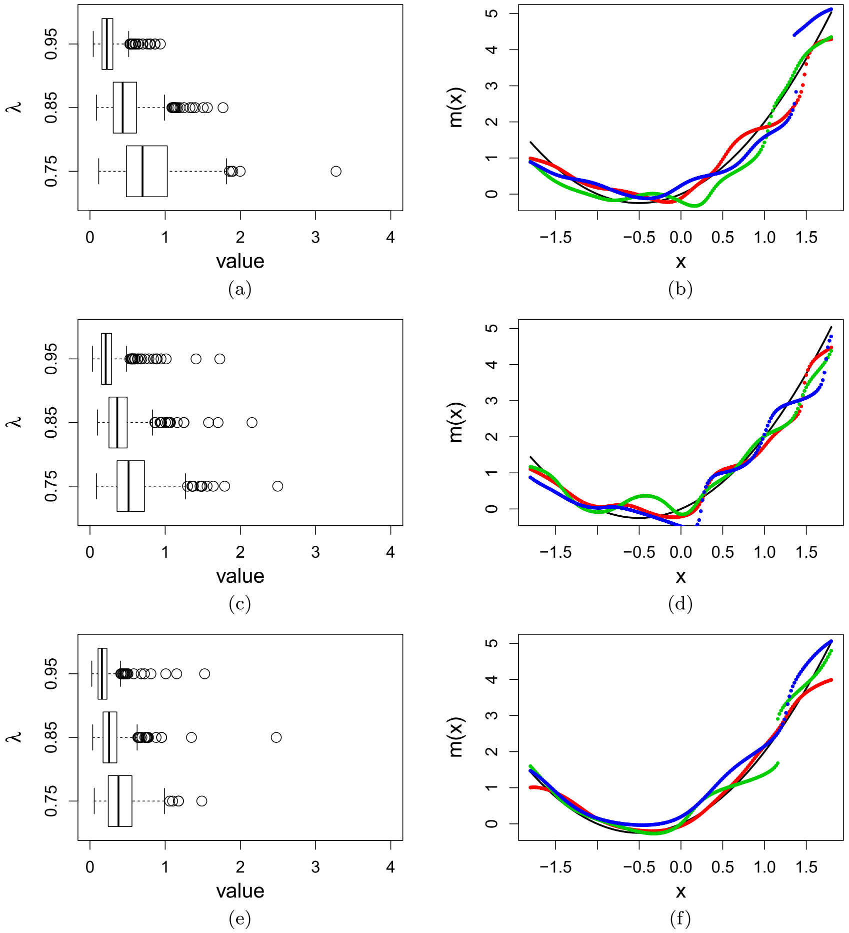


(a) (b)

(c) (d)

(e) (f)

Fig H.2. *Results under (C2) using approximated theoretical optimal bandwidths. Panels (a), (c), and (e): boxplots of ISEs versus* λ *for* $\hat{M}_N(x)$, $\hat{M}_0(x)$, *and* $\hat{M}_1(x)$, *respectively. Panels (b), (d), and (f): estimated mode curves,* $\hat{M}_N(x)$, $\hat{M}_0(x)$, *and* $\hat{M}_1(x)$, *respectively, when* $\lambda = 0.85$. *In each panel with estimated mode curves associated with an estimator, the black lines depict the true mode curves, the red, green, and blue lines are three estimated mode curves from the same method that yield ISE being the first, second, and third quantiles among the 500 ISEs for that method from the simulation, respectively.*

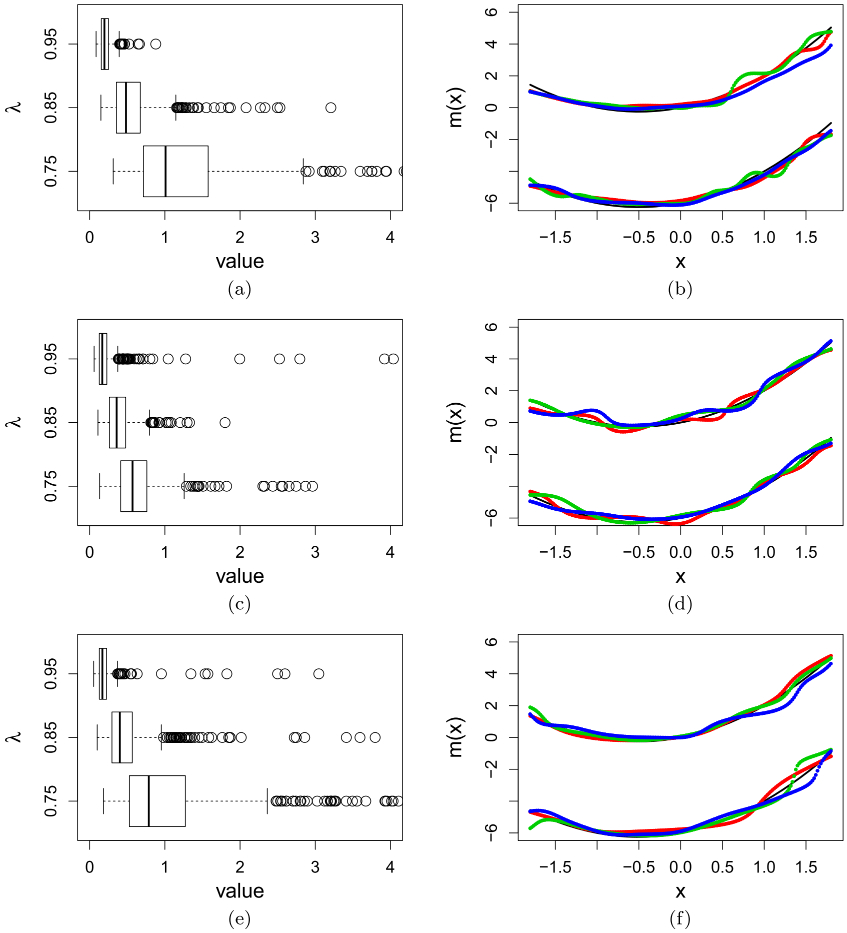


(a) (b)

(c) (d)

(e) (f)

Fig H.3. *Results under (C1) using the CV-SIMEX bandwidth selection. Panels (a), (c), and (e): boxplots of ISEs versus* λ *for* $\hat{M}_N(x)$, $\hat{M}_0(x)$, *and* $\hat{M}_1(x)$, *respectively. Panels (b), (d), and (f): estimated mode curves,* $\hat{M}_N(x)$, $\hat{M}_0(x)$, *and* $\hat{M}_1(x)$, *respectively, when* $\lambda = 0.85$. *In each panel with estimated mode curves associated with an estimator, the black line depicts the true mode curve, the red, green, and blue lines are three estimated mode curves from the same method that yield ISE being the first, second, and third quantiles among the 500 ISEs for that method from the simulation, respectively.*



(a) (b)

(c) (d)

(e) (f)

FIG H.4. *Results under (C2) using the CV-SIMEX bandwidth selection. Panels (a), (c), and (e): boxplots of ISEs versus* λ *for* $\hat{M}_N(x)$, $\hat{M}_0(x)$, *and* $\hat{M}_1(x)$, *respectively. Panels (b), (d), and (f): estimated mode curves,* $\hat{M}_N(x)$, $\hat{M}_0(x)$, *and* $\hat{M}_1(x)$, *respectively, when* $\lambda = 0.85$. *In each panel with estimated mode curves associated with an estimator, the black lines depict the true mode curves, the red, green, and blue lines are three estimated mode curves from the same method that yield ISE being the first, second, and third quantiles among the 500 ISEs for that method from the simulation, respectively.*

Appendix I: A sketch of the arguments in Section 3.1 with uni-modality assumption relaxed

Suppose $p(y|x)$ has K modes, with the mode set $M(x) = \{y_{M,1}(x), \ldots, y_{M,K}(x)\}$, and the estimated mode set $\hat{M}(x) = \{\hat{y}_{M,1}(x), \ldots, \hat{y}_{M,K}(x)\}$. Then the pointwise error is $\Delta_n(x) = \max_{1\le k\le K} |\hat{y}_{M,k}(x) - y_{M,k}(x)|$. By the mean-value theorem, for each $k = 1, \ldots, K$, one has

$$\hat{y}_{M,k}(x) - y_{M,k}(x) = -\{g_{yy}(x, y_{M,k}\}^{-1}\hat{g}_y(x, y_{M,k}) + O(\|\hat{g}_{yy} - g_{yy}\|_\infty)\hat{g}_y(x, y_{M,k}),$$

as in (3.2). It follows that

$$\begin{aligned}\Delta_n(x) = &\max_{1\le k\le K} |\{g_{yy}(x, y_{M,k})\}^{-1}\hat{g}_y(x, y_{M,k})| \\ &+ O(\|\hat{g}_{yy} - g_{yy}\|_\infty) \max_{1\le k\le K} |\hat{g}_y(x, y_{M,k})|,\end{aligned}$$

and thus

$$\begin{aligned}&\frac{\Delta_n(x)}{\max_{1\le k\le K} |\{g_{yy}(x, y_{M,k})\}^{-1}\hat{g}_y(x, y_{M,k})|} \\ &\quad = 1 + O(\|\hat{g}_{yy} - g_{yy}\|_\infty)\frac{\max_{1\le k\le K} |\hat{g}_y(x, y_{M,k})|}{\max_{1\le k\le K} |\{g_{yy}(x, y_{M,k})\}^{-1}\hat{g}_y(x, y_{M,k})|} \\ &\quad = 1 + O(\|\hat{g}_{yy} - g_{yy}\|_\infty),\end{aligned}$$

where the last equality results from assumption (CP2).

Hence, under the same conditions that supporting (3.3), $\Delta_n(x)$ can be approximated by $\max_{1\le k\le K} |\{g_{yy}(x, y_{M,k})\}^{-1}\hat{g}_y(x, y_{M,k})|$, and thus the convergence rate of $\Delta_n(x)$ is the same as that of $\max_{1\le k\le K} |\hat{g}_y(x, y_{M,k})|$. From this point on, all arguments in Section 3 regarding $\hat{g}_y(x, y_M)$, which is $\hat{p}_y(x, y_M)$ in Section 3.2 and is $\hat{p}_y(y_M|x)$ in Section 3.3, carry over to $\hat{g}_y(x, y_{M,k})$ for each $k = 1, \ldots, K$. And as long as K is finite, the convergence rate of $\max_{1\le k\le K} |\hat{g}_y(x, y_{M,k})|$ is the same as that of $|\hat{g}_y(x, y_{M,k})|$ for a $k \in \{1, \ldots, K\}$.

References

Bamford, S.P., Rojas, A.L., Genovese, C.R., Miller, C.E., Nichol R., and Wasserman, L. (2008) Revealing components of the galaxy population through nonparametric techniques. *Mon. Not. R. Astron. Soc.*, **391**, 607–616.

Carroll, R. and Hall, P. (1988) Optimal rates of convergence for deconvoluting a density. *J. Am. Statist. Ass.*, **83**, 1184–1186. MR0997599

Carroll, R., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. (2006) *Measurement error in nonlinear models: A model perspective*. Second edition. Chapman & Hall/CRC. Boca Raton, FL. MR2243417

Chacón, J.E., Duong, T. and Wand, M.P. (2011) Asymptotics for general multivariate kernel density derivative estimators. *Stat. Sinica*, **21**, 807–840. MR2829857

Chen, Y., Genovese, C.R., Tibshirani, R.J., and Wasserman, L. (2015) Asymptotic theory for density ridges. *Ann. Statist.*, **43**, 1896–1928. MR3375871

Chen, Y., Genovese, C.R., Tibshirani, R.J., and Wasserman, L. (2016) Nonparametric modal regression. *Ann. Statist.*, **44**, 489–514. MR3476607

Cheng, Y. (1995) Mean shift, mode seeking, and clustering. *IEEE. T. Pattern. Anal.*, **17**, 790–799.

Comaniciu, D. and Meer, P. (2002) Mean shift: a robust approach toward feature space analysis. *IEEE. T. Pattern. Anal.*, **24**, 603–619.

Cook, J.R. and Stefanski, L.A. (1994) Simulation-extrapolation estimation in parametic measurement error models. *J. Am. Statist. Ass.*, **89**, 1314–1328.

Delaigle, A., Fan, J., and Carroll, R. (2009) A design-adaptive local polynomial estimator for the error-in-variables problem. *J. Am. Statist. Ass.*, **104**, 348–359. MR2504382

Delaigle, A. and Hall, P. (2008) Using SIMEX for smoothing-parameter choice in errors-in-variables problems. *J. Am. Statist. Ass.*, **103**, 280-287. MR2394636

Delaigle, A., Hall, P., and Meister, A. (2008) On deconvolution with repeated measurements. *Ann. Statist.*, **36**, 665–687. MR2396811

Einmahl, U. and Mason, D.M. (2005) Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.*, **33**, 1380–1403. MR2195639

Einbeck, J. and Tutz, G. (2006) Modelling beyond regression functions: an application of multimodal regression to space-flow data. *J. R. Statist. Soc.*B, **55**, 461–475. MR2242274

Fan, J. (1991) Asymptotic normality for deconvolution kernel density estimators. *Sankhya* A, **53**, 97–110. MR1177770

Fan, J. (1991) Global behavior of deconvolution kernel estimates. *Stat. Sinica*, **1**, 541–551. MR1130132

Fan, J. (1991) On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, **19**, 1257–1272. MR1126324

Fan, J. and Gijbels, I. (1996) *Local polynomial modelling and its applications*, Chapman and Hall/CRC, Boca Raton. MR1383587

Fan, J. and Truong, Y.K. (1993) Nonparametric regression with errors in variables. *Ann. Statist.*, **21**, 1900–1925. MR1245773

Fan, J., Yao, Q., and Tong, H. (1996) Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, **83**, 189–206. MR1399164

Fan, J. and Yim, T.H. (2004) A cross validation method for estimating conditional densities. *Biometrika*, **91**, 819–834. MR2126035

Genovese, C.R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2014) Nonparametric ridge estimation. *Ann. Statist.*, **42** 1511–1545. MR3262459

Ginè, E. and Guillou, A. (2002) Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. H. Poincarè Probab. Statist*, **38** 907–921. MR1955344

Hall, P., Racine, J., and Li, Q. (2004) Cross-validation and the estimation of conditional probability densities. *J. Am. Statist. Ass.*, **99**, 1015–1026. MR2109491

Hastie, T. and Stuezle, S. (1989) Principal curves. *J. Am. Statist. Ass.*, **44**, 489–514.

He, X., and Liang, H. (2000) Quantile regression estimates for a class of linear and partially linear errors-in-variables models. *Statist. Sin.*, **10**, 129–140. MR1742104

Huang, M., Li, R., and Wang, S. (2013) Nonparametric mixture of regression models. *J. Am. Statist. Ass.*, **108**, 929–941. MR3174674

Hyndman, R.J., Bashtannyk, D.M. and Grunwald, G.K. (1996) Estimating and visualizing conditional densities. *J. Comput. Graph. Stat.*, **5**, 315–336. MR1422114

Koenker, R. (2005) *Quantile regression.* Cambridge University Press. MR2268657

Liang, H. and Wang, N. (2005) Partially linear single-index measurement error models. *Stat. Sinica*, **15**, 99–116. MR2125722

Ma, Y. and Yin, G. (2011) Censored quantile regression with covariate measurement errors. *Stat. Sinica*, **21**, 949–971. MR2829862

Nakamura, T. (1990) Corrected score functions for errors-in-variables models: methodology and applications to generalized linear models. *Biometrika*, **77**, 127–137. MR1049414

Novick, S.J. and Stefanski, L.A. (2002) Corrected score estimation via complex variable simulation extrapolation. *J. Am. Statist. Ass.*, **97**, 472–481. MR1941464

Ozertem, U. and Erdogmus, D. (2011) Locally defined principal curves and surfaces. *J. Mach. Learn. Res.*, **12**, 1249–1286. MR2804600

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall. MR0848134

Stefanski, L.A. and Carroll, R.J. (1990) Deconvoluting kernel density estimators. *Statistics*, **21**, 169–184. MR1054861

van der Vaart, A.W. and Wellner, J.A. (1996) *Weak convergence and empirical process: with applications to statistics.* Springer, New York. MR1385671

Wand, M.P. and Jones, M.C. (1993) Comparison of smoothing parameterizations in bivariate kernel density estimation. *J. Am. Statist. Ass.*, **88**, 520–528. MR1224377

Wang, H.J., Stefanski, L.A. and Zhu, Z. (2012) Corrected-loss estimation for quantile regression with covariate measurement errors. *Biometrika*, **99**, 405–421. MR2931262

Wei, Y. and Raymond, C.J. (2009) Quantile regression with measurement error. *J. Am. Statist. Ass.*, **104**, 1129–1143. MR2562008

Yao, W. and Li L. (2014) A new regression model: modal linear regression. *Scand. J. Stat.*, **41**, 656–671. MR3249422

Yao, W., Lindsay, B., and Li, R. (2012) Local modal regression. *J. Nonparametr. Stat.*, **24**, 647–663. MR2968894