BANDWIDTH SELECTION FOR NONPARAMETRIC REGRESSION WITH ERRORS-IN-VARIABLES

HAO DONG, TAISUKE OTSU, AND LUKE TAYLOR

Abstract. We propose two novel bandwidth selection procedures for the nonparametric regression model with classical measurement error in the regressors. Each method is based on evaluating the prediction errors of the regression using a second (density) deconvolution. The first approach uses a typical leave-one-out cross validation criterion, while the second applies a bootstrap approach and the concept of out-of-bag prediction. We show the asymptotic validity of both procedures and compare them to the SIMEX method of Delaigle and Hall (2008) in a Monte Carlo study. As well as enjoying advantages in terms of computational cost, the methods proposed in this paper lead to lower mean integrated squared error compared to the current state-of-the-art.

1. Introduction

Measurement error is rife in the social sciences where survey data are common and imprecise measurement instruments are used (see, for example, Blattman *et al.*, 2016). As well as being ubiquitous, if measurement error is not accounted for, estimation bias can be introduced, masking the true relationship between variables and rendering testing procedures invalid. Moreover, measurement error can be particularly troublesome when using nonparametric methods, which are now commonplace in applied work due to increases in computing power and data availability. A vital concern when using any nonparametric technique is the choice of bandwidth, leading to a great demand for robust, data-driven methods to select this parameter.

To this end, we consider bandwidth selection in the nonparametric estimation of a regression model with errors-in-variables:

$$Y = m(X) + U, \qquad E[U|X] = 0, \tag{1}$$

where $Y \in \mathbb{R}$ is a response variable, $X \in \mathbb{R}$ is an error-free but unobservable covariate, $U \in \mathbb{R}$ is an error term, and $m(\cdot) = E[Y|X = \cdot]$ is the conditional mean function of Y given X. We wish to estimate m using an independent and identically distributed (i.i.d.) sample $\{Y_j, W_j\}_{j=1}^n$ of (Y, W), where W is a noisy measurement of X generated by

$$W = X + \epsilon, \tag{2}$$

with $\epsilon \in \mathbb{R}$ a classical measurement error independent of (Y, X) with known density f_ϵ.

Let $g^{\mathrm{ft}}(t) = \int e^{\mathrm{i}tx} g(x)dx$ denote the Fourier transform of a function g with $\mathrm{i} = \sqrt{-1}$. One of the most popular estimators of the regression function m is the deconvolution kernel estimator

The authors acknowledge financial supports from the SMU Dedman College Research Fund (12-412268) (Dong) and the Aarhus University Research Fund (AUFF-26852) (Taylor).

(Fan and Truong, 1993)

$$\hat{m}(x;h) = \frac{\sum_{j=1}^{n} \mathbb{K}_h\left(\frac{x-W_j}{h}\right) Y_j}{\sum_{j=1}^{n} \mathbb{K}_h\left(\frac{x-W_j}{h}\right)},$$

where $\mathbb{K}_h$ is the deconvolution kernel defined as

$$\mathbb{K}_h(u) = \frac{1}{2\pi} \int e^{-\mathrm{i}tu} \frac{K^{\mathrm{ft}}(t)}{f_\epsilon^{\mathrm{ft}}(t/h)} dt,$$

with an (ordinary) kernel function $K : \mathbb{R} \to \mathbb{R}$ and bandwidth parameter h. Even for this simple estimator in this standard model, only one method currently exists to select the bandwidth parameter: a SIMEX approach due to Delaigle and Hall (2008).[1]

In this paper, we provide two alternative bandwidth selection procedures for the deconvolution regression estimator. Each method is based on evaluating the regression prediction errors using a second (density) deconvolution. The first procedure uses a typical leave-one-out cross validation criterion, while the second applies a bootstrap approach and the concept of out-of-bag prediction (Breiman, 2001). Relative to the SIMEX approach of Delaigle and Hall (2008), both methods reduce computational cost by an order of magnitude. This is particularly pertinent for nonparametric deconvolution estimators which are computationally expensive. Moreover, our simulation results show that while neither the out-of-bag nor leave-one-out method dominates the other, they both lead to reduced mean integrated squared error (MISE) in comparison to the SIMEX approach.

Deconvolution methods within statistics have predominantly focused on density and regression estimation in the presence of measurement errors. Carroll and Hall (1988) and Stefanski and Carroll (1990) introduced the deconvolution kernel density estimator which has been extended in many directions in the last three decades. Their approach was adapted to the problem of regression estimation with mismeasured regressors by Fan and Truong (1993), which has subsequently also been extended in several directions, most notably to the heteroskedastic error case (for example, Delaigle and Meister, 2007) and to settings where the distribution of the measurement error is unknown (for example, Delaigle, Hall and Meister, 2008). For a detailed review on this ever-growing subject see, for example, Schennach (2016).

Throughout this literature, it has been widely acknowledged that the performance of kernel deconvolution estimators depends sensitively on the choice of bandwidth. In response to this, several papers have studied procedures to choose this tuning parameter. For density deconvolution, Fan (1991) suggested a simple rule-of-thumb method, Stefanski and Carroll (1990) proposed a plug-in approach based on minimising the asymptotic MISE, and Delaigle and Gijbels (2004a) developed a bootstrap method. However, there has been far less work on the notoriously difficult problem of bandwidth selection in nonparametric regression in the presence of measurement error. Delaigle and Hall (2008) is one of the few exceptions to this, proposing a SIMEX-type approach to this issue; we compare their method to those developed in this paper in Section 3. Finally, Chichignoud *et al.* (2017) developed an adaptive data-driven selector for the wavelet

[1]Delaigle, Hall and Jamshidi (2015) and Kato and Sasaki (2019) also provide methods (closely related to Delaigle and Hall, 2008) for bandwidth selection in this model; however, these methods are designed to ensure the validity of confidence bands rather than optimising the estimation of m.

resolution in the wavelet deconvolution regression. However, their method assumes that the distribution of the regression error U is normal with known variance, and that the support of the error-free covariate X is known.

This paper proceeds as follows. In Section 2, we give details of the bandwidth selection mechanisms proposed and outline their theoretical properties. Section 3 provides results for the small sample properties of our procedures and compares them to the SIMEX approach of Delaigle and Hall (2008). Finally, in Section 4, we apply our method to real data to describe the relationship between blood pressure and cognitive ability. All mathematical proofs and auxiliary lemmas are relegated to Appendix.

2. Methodology

In this section, we present our bandwidth selection procedures. As a population criterion for determining the optimal bandwidth, we consider the mean squared prediction error for the $(n+1)^{\text{th}}$ observation

$$R(h) = E[\{Y_{n+1} - \hat{m}(X_{n+1}; h)\}^2]. \tag{3}$$

Our aim is to estimate this function and select the bandwidth h which minimises this.[2] In the absence of measurement error (i.e., X is observable), $R(h)$ could be estimated by the leave-one-out cross validation estimator

$$\hat{R}_{\text{infeasible}}(h) = \frac{1}{n}\sum_{j=1}^{n}\{Y_j - \hat{m}_j(X_j; h)\}^2,$$

where $\hat{m}_j(\cdot; h)$ is the leave-j-out counterpart of $\hat{m}(\cdot; h)$. However, when X is mismeasured, this approach is infeasible, and an alternative strategy must be found which allows for the estimation of $R(h)$ based only on the observables (Y, W) generated by (1) and (2).

Below, we present two approaches to estimate the mean squared prediction error $R(h)$: the leave-one-out approach (Section 2.1) and the out-of-bag approach (Section 2.2).

2.1. Leave-One-Out Approach.

Note that the mean squared prediction error can be expressed as

$$R(h) = E\left[\iint \{y - \hat{m}(x; h)\}^2 f_{YX}(y, x)dydx\right], \tag{4}$$

where f_{YX} is the joint density function of (Y, X), and the expectation is taken with respect to the observables used to compute $\hat{m}(\cdot; h)$. The joint density f_{YX} can be estimated by the deconvolution kernel density estimator

$$\hat{f}_{YX}(y, x) = \frac{1}{nh_y h_{1,x}}\sum_{j=1}^{n} K_y\left(\frac{y - Y_j}{h_y}\right)\mathbb{K}_{h_{1,x}}\left(\frac{x - W_j}{h_{1,x}}\right),$$

[2]Note that we only consider global bandwidth selection; the choice of a local bandwidth which changes over the range of the regressor is beyond the scope of this paper. For the local bandwidth choice, one may extend the conventional approach to minimise an estimate of the (approximate) MSE $E[\{\hat{m}(x; h) - m(x)\}^2]$ for given x. Also it is interesting to see whether we can extend recent developments for coverage error optimal bandwidths (Calonico, Cattaneo and Farrell, 2018, 2020) to the nonparametric deconvolution context.

where K_y is an ordinary kernel function and $(h_y, h_{1,x})$ are bandwidth parameters for this density estimation. Since Y is error-free, we apply the deconvolution kernel $\mathbb{K}_{h_{1,x}}$ only for W. If K_y is a higher-order kernel satisfying $\int K_y(a) = 1$ and $\int a^l K_y(a) = 0$ for $l = 1, 2$, then the integral in (4) can be estimated by[3]

$$\iint \{y - \hat{m}(x;h)\}^2 \hat{f}_{YX}(y,x)dydx = \frac{1}{nh_{1,x}} \sum_{j=1}^{n} \int \{Y_j - \hat{m}(x;h)\}^2 \mathbb{K}_{h_{1,x}} \left(\frac{x - W_j}{h_{1,x}} \right) dx.$$

Motivated by this expression, we can estimate $R(h)$ using the leave-one-out approach as

$$\hat{R}_{LOO}(h) = \frac{1}{nh_{1,x}} \sum_{j=1}^{n} \int \{Y_j - \hat{m}_j(x;h)\}^2 \mathbb{K}_{h_{1,x}} \left(\frac{x - W_j}{h_{1,x}} \right) dx. \tag{5}$$

In practice, the integration with respect to x in (5) is commonly conducted over a compact set $\mathcal{X}$ instead of $\mathbb{R}$. To this end, instead of $\hat{R}_{LOO}(h)$, we focus on the following truncated estimator

$$\tilde{R}_{LOO}(h) = \frac{1}{nh_{1,x}} \sum_{j=1}^{n} \int_{\mathcal{X}} \{Y_j - \hat{m}_j(x;h)\}^2 \mathbb{K}_{h_{1,x}} \left(\frac{x - W_j}{h_{1,x}} \right) dx, \tag{6}$$

and the optimal bandwidth, denoted as h^*_{LOO}, is chosen as the minimiser of $\tilde{R}_{LOO}(h)$, i.e.,

$$h^*_{LOO} = \underset{h \in [L_{1,n}, H_{1,n}]}{\operatorname{argmin}} \tilde{R}_{LOO}(h),$$

where $L_{1,n}$ and $H_{1,n}$ are deterministic sequences characterising the upper and lower bounds of the region to search for h^*_{LOO}, respectively.

In order to implement $\tilde{R}_{LOO}(h)$, an auxiliary bandwidth $h_{1,x}$ for estimation of f_{YX} must be chosen. This is typical in bandwidth selection procedures in the presence of measurement error. For example, Delaigle and Gijbels (2004a) require an initial bandwidth to estimate a criterion function for a density estimator bandwidth choice procedure, as do Delaigle and Hall (2008) for their SIMEX approach. Both Delaigle and Gijbels (2004a) and Delaigle and Hall (2008) suggest using the normal reference bandwidth of Stefanski and Carroll (1990). In Sections 3 and 4, we use the bandwidth selection procedure of Delaigle and Gijbels (2004a), which itself uses a normal reference pilot bandwidth. In Section 3, we also provide results on the sensitivity of our procedure to this initial bandwidth choice.

To establish the asymptotic validity of $\tilde{R}_{LOO}(h)$, based on Wong (1983), we focus on the following integrated squared error loss[4]

$$R_n(h) = \int_{\mathcal{X}} \{\hat{m}(x;h) - m(x)\}^2 f(x)dx, \tag{7}$$

[3]It is interesting to note that our deconvolution approach to estimate $R(h)$ in (4) may be applied to other estimation methods to construct $\hat{m}(x;h)$, where the meaning of the tuning parameter h changes. For example, Davezies and Barbanchon (2017) and Bartalotti, Brummet and Dieterle (2020) proposed nonparametric regression estimators in the context of regression discontinuity designs, where auxiliary data are available. Although they allow non-classical measurement errors, it is interesting to see whether our approach can be adapted to suggest bandwidth selectors for their estimators under the classical measurement error case.

[4]For the error-free case, Wong (1983) considered the average squared error loss $n^{-1} \sum_{j=1}^{n} \{\hat{m}(X_j;h) - m(X_j)\}^2$ as the criterion to select the bandwidth. Since X is unobservable in our contaminated case, it is natural to consider the integrated squared error loss $R_n(h)$.

where f is the marginal density of X. In particular, we shall prove consistency of the form $R_n(h^*_{LOO}) \xrightarrow{p} 0$ as $n \to \infty$, i.e., the integrated squared error loss converges to zero with the optimal bandwidth. To this end, we impose the following assumptions. Let $r_\epsilon(a) = \left\{\inf_{|t|\leq a^{-1}} |f^{\text{ft}}_\epsilon(t)|\right\}^{-1}$.

Assumption.

(1): *$\{Y_j, W_j\}_{j=1}^n$ is an i.i.d. sample of (Y, W) generated by (1) and (2), $E[Y^8] < \infty$, ϵ is independent of (Y, X) with zero mean, f_ϵ is known, and $f^{\text{ft}}_\epsilon(t) \neq 0$ for all $t \in \mathbb{R}$.*

(2): *$E[Y^2|X = \cdot]$, the regression function m, and the density f of X are p-times continuously differentiable with bounded and integrable derivatives, f is bounded away from zero over $\mathcal{X}$, and $E[Y^4|X = \cdot]$ is bounded.*

(3): *K is symmetric around zero and satisfies $\int K(u)du = 1$, $\int K(u)u^p du \neq 0$, and $\int K(u)u^q du = 0$ for all positive integers $q < p$. Also, $K^{\text{ft}}(t)$ is supported on $[-1, 1]$ and bounded.*

(4): *$(n^{1/2}h_{1,x})^{-1} \log(1/\sqrt{h_{1,x}}) \max\{n^{-1/2}r^2_\epsilon(h_{1,x}), 1\} \to 0$ as $n \to \infty$.*

(5): *$(n^{1/2}L_{1,n})^{-1} \log(1/\sqrt{L_{1,n}}) \max\{n^{-1/2}r^2_\epsilon(L_{1,n}), 1\} \to 0$, $H_{1,n} \to 0$, $L_{1,n} \leq h_{1,x} \leq H_{1,n}$, and $(n^{3/4}h_{1,x}L_{1,n})^{-1} r_\epsilon(h_{1,x}) r_\epsilon(L_{1,n}) \to 0$ as $n \to \infty$.*

Assumption (1) requires random sampling, some regularity conditions, and a classical measurement error with known distribution. In particular, $E[Y^8] < \infty$ is used in Lemma 3 in the Appendix to control the order of $\max_{1\leq j\leq n} |Y_j|^4$. The non-vanishing condition for f^{ft}_ϵ is commonly employed in kernel-based deconvolution methods and is satisfied for many distributions. Our method may be extended to the case where f^{ft}_ϵ is allowed to take zeros by introducing an additional ridge parameter (see, for example, Hall and Meister, 2007, and Meister, 2009). Assumption (2) imposes smoothness restrictions on the first and second conditional moments of Y and the density of X, bounded fourth conditional moments of Y, and that the density of X is non-vanishing over $\mathcal{X}$. Assumption (3) is a higher-order kernel assumption, which, together with the smoothness restrictions imposed in Assumption (2), are used to reduce the estimation bias. Due to the regularisation used in deconvolution problems, the Fourier transform of K is further required to be compactly supported. Assumption (4) imposes restrictions on the auxiliary bandwidth $h_{1,x}$.[5] In particular, to establish the uniform rate of convergence for the estimands based on $h_{1,x}$, we need Lemma 6 in the Appendix, which requires $(n^{1/2}h_{1,x})^{-1} \log(1/\sqrt{h_{1,x}}) \to 0$. The other condition is used to ensure the derived rate converges to zero as $n \to \infty$. Assumption (5) considers the upper and lower bounds of the region to search for h^*_{LOO}. For the upper bound $H_{1,n}$, it is required to be no smaller than the auxiliary bandwidth $h_{1,x}$ and go to zero as $n \to \infty$. However, the conditions on the lower bound $L_{1,n}$ are more complicated; it depends on both the

[5] In the ordinary smooth case when f_ϵ is of order α and K is of order β, the MSE optimal bandwidth is given by $h^*_{1,x} \sim n^{-1/(1+2\alpha+2\beta)}$ and $r_\epsilon(h^*_{1,x}) \sim h^{*-\alpha}_{1,x} = n^{\alpha/(1+2\alpha+2\beta)}$. Then Assumption (4) can be satisfied by $h^*_{1,x}$ if $\alpha > 0$, $\beta > 0$, and

$$n^{(\alpha+\beta-1/2)/(1+2\alpha+2\beta)} \max\{n^{(-1/2+\alpha-\beta)/(1+2\alpha+2\beta)}, 1\} \log n \to 0.$$

If $\alpha < \beta + 1/2$, this holds true if $\alpha + \beta > 1/2$. Thus, for the MSE optimal bandwidth $h^*_{1,x}$ to satisfy our Assumption (4), we only need a mild smoothness condition on f_ϵ, such as $\alpha > 1/2$ for smoothness of f_ϵ and $\beta > \alpha - 1/2$ for the order of K. Similar results can be obtained for the supersmooth case.

choice of the auxiliary bandwidth $h_{1,x}$ (no greater than the auxiliary bandwidth $h_{1,x}$ in particular) and the smoothness of the error distribution, reflected by $r_\epsilon(\cdot)$. Also note that Assumption (4) is implied by Assumption (5) if $r_\epsilon(a)$ is decreasing in a for small a, which is satisfied by both the Laplace and normal distribution among others.

It is worth noting at this stage that we do not split our discussion based on the decay rate of the tail of the error characteristic function f_ϵ^{ft}, as is typical in the nonparametric measurement error literature. By maintaining generality, our results can be applied to both ordinary smooth and supersmooth error distributions. These assumptions lead to the following consistency result.

Theorem 1. *Under Assumptions (1)-(5),*

$$R_n(h^*_{LOO}) \xrightarrow{p} 0.$$

Theorem 1 establishes the consistency of h^*_{LOO} with respect to the integrated squared error loss R_n (in an analogous sense to Wong, 1983). Since R_n is defined by integrating x over $\mathcal{X}$ rather than $\mathbb{R}$, h^*_{LOO} could be inconsistent; thus, integrating x over a truncated region does carry a cost. However, this cost will be small when working with a large enough $\mathcal{X}$ (so that $\mathcal{X}$ is close to the support of X).

It is also worth noting that the consistency result presented here is derived from $R_n(h^*_{LOO}) \leq R_n(h_{1,x}) + O\left(\sup_{h\in[L_{1,n},H_{1,n}]} |\Delta_n(h)|\right)$ (see eq. (2) in Appendix A.1), where $\Delta_n(\cdot)$ depends on the auxiliary bandwidth, $h_{1,x}$, the smoothness of the conditional moments and densities reflected by p, and the smoothness of the measurement error distribution reflected by $r_\epsilon(\cdot)$. From Lemma 6 in the Appendix, $R_n(h_{1,x}) = O_p(r_n^2(h_{1,x}))$, where $r_n(h_{1,x}) = (nh_{1,x})^{-1/2} r_\epsilon(h_{1,x})\sqrt{\log(1/\sqrt{h_{1,x}})}$. Thus, if we further impose $\sup_{h\in[L_{1,n},H_{1,n}]} |\Delta_n(h)| r_n^{-2}(h_{1,x}) \to 0$, then $R_n(h^*_{LOO}) \leq R_n(h_{1,x})(1 + o(1))$, which shows that h^*_{LOO} asymptotically leads to a value for R_n no greater than that achieved by the auxiliary bandwidth $h_{1,x}$, so h^*_{LOO} is at least as good as the auxiliary bandwidth.

2.2. **Out-Of-Bag Approach.** In this section, we present an alternative bootstrap-based procedure. Take a bootstrap sample of size n (with replacement) from the original data and estimate m using this bootstrap sample (denoted by $\hat{m}_b(\cdot; h)$ for $b = 1, \ldots, B$). Let $\mathcal{I}_b$ be the set of observations in the bootstrap sample b, $\mathcal{I}_b^c$ be the complement of this set, i.e. the out-of-bag observations, and n_b^c be the cardinality of the set $\mathcal{I}_b^c$. On average, these out-of-bag observations include 36.8% of the total observations, irrespective of sample size (Breiman, 2001). For each $b = 1, \ldots, B$, the out-of-bag bootstrap counterpart of (5) can be obtained as

$$\tilde{R}_b(h) = \frac{1}{n_b^c h_{2,x}} \sum_{j\in\mathcal{I}_b^c} \int_{\mathcal{X}} \{Y_j - \hat{m}_b(x; h)\}^2 \mathbb{K}_{h_{2,x}}\left(\frac{x - W_j}{h_{2,x}}\right) dx,$$

where $h_{2,x}$ is an auxiliary bandwidth.

The out-of-bag bootstrap estimator for the mean squared prediction error $R(h)$ is then obtained by taking an average over the bootstrap samples:

$$\tilde{R}_{OOB}(h) = \frac{1}{B}\sum_{b=1}^{B} \tilde{R}_b(h), \tag{8}$$

and the optimal bandwidth, denoted h^*_{OOB}, is chosen as the minimiser of $\tilde{R}_{OOB}(h)$, i.e.,

$$h^*_{OOB} = \underset{h\in[L_{2,n},H_{2,n}]}{\operatorname{argmin}} \tilde{R}_{OOB}(h),$$

where $L_{2,n}$ and $H_{2,n}$ are deterministic sequences characterizing the upper and lower bounds of the region to search for h^*_{OOB}.

It is worth noting that an alternative approach of sample splitting is undesirable in this context. Such an approach proceeds by estimating m on some fraction of the data and using the remaining data to evaluate the estimator. This would result in an estimator of m using a sample size of less than n; hence, the bandwidth chosen is optimal for an estimator which does not use all observations. Of course, if the order of the optimal bandwidth is known, the selected bandwidth can be scaled down by the appropriate factor. However, the order of the optimal bandwidth typically depends on features of the underlying data, such as the smoothness of the measurement error, that are unlikely to be known in practice. Our out-of-bag approach avoids this issue, resulting in a bandwidth applicable for samples of size n.

To show the asymptotic validity of $\tilde{R}_{OOB}(h)$, we introduce following conditions on the auxiliary bandwidth and the bounds of the search region, which are slight relaxation of Assumptions (4) and (5).

Assumption.

(4'): $(nh_{2,x})^{-1}r^2_\epsilon(h_{2,x})\log(1/\sqrt{h_{2,x}}) \to 0$ *as* $n\to\infty$.

(5'): $(nL_{2,n})^{-1}r^2_\epsilon(L_{2,n})\log(1/\sqrt{L_{2,n}}) \to 0$, $H_{2,n}\to 0$, *and* $L_{2,n}\le h_{2,x}\le H_{2,n}$ *as* $n\to\infty$.

Theorem 2. *Under Assumptions (1)-(3), (4') and (5'), it holds*

$$R_n(h^*_{OOB}) \xrightarrow{p} 0.$$

It would be interesting to investigate whether our bandwidth selection procedures can achieve asymptotic optimality in an analogous sense to Härdle and Marron (1985) or Härdle, Hall and Marron (1988) for the error-free case. However, we leave this extension for future research.

2.3. **Unknown Measurement Error.** Throughout the preceding discussion, we have assumed that the characteristic function of the measurement error f^{ft}_ϵ is known to the researcher. However, this is neither realistic in practice nor necessary for our bandwidth selection procedures. Given additional auxiliary data, there are several potential methods to estimate this characteristic function. For example, with a second independent noisy measurement of the error contaminated regressor, the estimators of Delaigle, Hall and Meister (2008) or Li and Vuong (1998) can be used.

Given a consistent estimator of f^{ft}_ϵ, denoted $\hat{f}^{\text{ft}}_\epsilon$, the leave-one-out and out-of-bag criterion functions can be constructed as

$$\begin{aligned}
\check{R}_{LOO}(h) &= \frac{1}{nh_{1,x}}\sum_{j=1}^{n}\int_{\mathcal{X}}\{Y_j-\check{m}_j(x;h)\}^2\hat{\mathbb{K}}_{h_{1,x}}\left(\frac{x-W_j}{h_{1,x}}\right)dx,\\
\check{R}_{OOB}(h) &= \frac{1}{Bn^c_b h_{2,x}}\sum_{b=1}^{B}\sum_{j\in\mathcal{I}^c_b}\int_{\mathcal{X}}\{Y_j-\check{m}_b(x;h)\}^2\hat{\mathbb{K}}_{h_{2,x}}\left(\frac{x-W_j}{h_{2,x}}\right)dx,
\end{aligned}$$

where

$$\hat{\mathbb{K}}_h(x) = \frac{1}{2\pi}\int e^{-\mathrm{i}tx}\frac{K^{\mathrm{ft}}(t)}{\hat{f}_\epsilon^{\mathrm{ft}}(t/h)}dt, \qquad \check{m}(x;h) = \frac{\sum_{j=1}^n \hat{\mathbb{K}}_h\left(\frac{x-W_j}{h}\right)Y_j}{\sum_{j=1}^n \hat{\mathbb{K}}_h\left(\frac{x-W_j}{h}\right)}.$$

Note that, as in Delaigle, Hall and Meister (2008), it is not necessary to estimate leave-one-out or out-of-bag estimates of $\hat{f}_\epsilon^{\mathrm{ft}}$ in $\check{R}_{LOO}(h)$ or $\check{R}_{OOB}(h)$. While it is beyond the scope of this paper to prove the asymptotic validity of these criterion functions, we conjecture that using similar methods of proof to those of Theorems 1 and 2, similar results can be obtained. Indeed, our simulation results in Section 3 indicate that our methods continue to perform admirably when the error density is estimated.

2.4. **Multivariate Regression.** For ease of exposition, thus far we have restricted attention to a single mismeasured regressor; however, the methods proposed in this paper can easily be extended to multivariate settings where a mixture of correctly and incorrectly measured covariates are present.

Suppose the model of interest now takes the following form

$$Y = m(X,Z) + U, \qquad E[U|X,Z] = 0, \tag{9}$$

where $X \in \mathbb{R}^{d_X}$ is an error-free but unobservable set of covariates, and $Z \in \mathbb{R}^{d_Z}$ is an error-free observable set of covariates. Again, we denote $W \in \mathbb{R}^{d_X}$ as a set of observable noisy measurements of X contaminated with classical measurement error. Note that we now require a method to select $(d_X + d_Z)$ bandwidths.

The criterion functions for the selection of the optimal set of bandwidths take analogous forms to their univariate counterparts:

$$\begin{aligned}
\tilde{R}_{LOO}(h_X,h_Z) &= \frac{1}{n\left(\prod h_{1,x}\right)}\sum_{j=1}^n\int_\chi \{Y_j - \hat{m}_j(x,Z_j;h_X,h_Z)\}^2\mathbb{K}_{h_{1,x}}\left(\frac{x-W_j}{h_{1,x}}\right)dx,\\
\tilde{R}_{OOB}(h_X,h_Z) &= \frac{1}{B}\sum_{b=1}^B\frac{1}{n_b^c\left(\prod h_{2,x}\right)}\sum_{j\in\mathcal{I}_b^c}\int_\chi \{Y_j - \hat{m}_b(x,Z_j;h_X,h_Z)\}^2\mathbb{K}_{h_{2,x}}\left(\frac{x-W_j}{h_{2,x}}\right)dx,
\end{aligned}$$

where $h_X = (h_{X,1},\ldots,h_{X,d_X})$ and $h_Z = (h_{Z,1},\ldots,h_{Z,d_Z})$ are the bandwidths to be optimised, $h_{\iota,x} = (h_{\iota,x,1},\ldots,h_{\iota,x,d_X})$ for $\iota = 1,2$ is a vector of auxiliary bandwidths to estimate the density of f_{YX}, $\prod h_{\iota,x}$ is understood as the product of the elements of $h_{\iota,x}$, and $\frac{x-W_j}{h_{\iota,x}}$ is understood as the element-wise division for $\iota = 1,2$, $\mathbb{K}_h$ is a $\dim(h)$-dimensional deconvolution kernel function depending on bandwidth vector h as in, for example, Masry (1993), and $\hat{m}_j(x,z;h_X,h_Z)$ is the leave-j-out counterpart of the multivariate deconvolution kernel estimator

$$\hat{m}(x,z;h_X,h_Z) = \frac{\sum_{j=1}^n \mathbb{K}_{h_X}\left(\frac{x-W_j}{h_X}\right)K\left(\frac{z-Z_j}{h_Z}\right)Y_j}{\sum_{j=1}^n \mathbb{K}_{h_X}\left(\frac{x-W_j}{h_X}\right)K\left(\frac{z-Z_j}{h_Z}\right)},$$

with an ordinary kernel function K; $\hat{m}_b$ is the out-of-bag counterpart, which is defined analogously to $\hat{m}_j$.

To obtain the optimal bandwidth parameters, a grid search across all combinations of h_X and h_Z is required; this will be computationally demanding even if the dimensions of X and Z are small. Given the lower computational cost of the leave-one-out procedure, we suggest practitioners to use this approach rather than the out-of-bag method in this case.

3. Simulation

In this section, we evaluate the finite sample performance of the proposed bandwidth selection procedures using Monte Carlo simulation. The following data generating process is considered

$$Y = m(X) + U,$$

where U and X are drawn independently from $N(0,1)$ and four specifications of m are considered:

$$\begin{aligned}
&\text{DGP1} : m(x) = x,\\
&\text{DGP2} : m(x) = x - x^2,\\
&\text{DGP3} : m(x) = \cos(x),\\
&\text{DGP4} : m(x) = \sin(x).
\end{aligned}$$

Note that each function is further standardised by its respective standard deviation, $SD[m(X)]$, so that each regression function has the same explanatory power.

Although X is assumed unobservable, we observe $W = X + \epsilon$, where ϵ is independent of (X, U) and has a known distribution. We consider two cases for this distribution based on smoothness: an ordinary smooth setting where ϵ has a zero mean Laplace distribution, and a supersmooth setting where ϵ has a normal distribution with zero mean. We provide results for two levels of noise, $\sigma_\epsilon = 1/3$ and $\sigma_\epsilon = 1/2$, and two sample sizes, $n = 250$ and $n = 500$. All results are based on 1000 Monte Carlo replications.

Throughout the simulation study, we use the infinite-order flat-top kernel proposed by McMurry and Politis (2004), which is defined by its Fourier transform

$$K^{\text{ft}}(t) = \begin{cases} 1 & \text{if } |t| \leq 0.05, \\ \exp\left\{\frac{-\exp(-(|t|-0.05)^2)}{(|t|-1)^2}\right\} & \text{if } 0.05 < |t| < 1, \\ 0 & \text{if } |t| \geq 1. \end{cases}$$

We compare three methods of bandwidth selection. The out-of-bag method, the leave-one-out approach, and the SIMEX procedure of Delaigle and Hall (2008). The SIMEX procedure is based on a leave-one-out criterion. It involves estimating the optimal bandwidth for two simulated datasets with varying degrees of measurement error and deducing the implied optimal bandwidth for the original dataset based on these results. This is typically repeated B times, with the results averaged to get a final estimate for the optimal bandwidth. From a computational standpoint, this involves approximately $2B$ times more function evaluations than the leave-one-out approach of this paper. Following Delaigle and Hall (2008) we choose $B = 20$ and use the same number of bootstrap replications for the out-of-bag method. Benchmarking the computational cost of the

three methods, we found the leave-one-out approach to be 16.6 times faster than the out-of-bag method and 41.5 times faster than the SIMEX procedure.

All three methods require choosing a range of integration χ; we keep this fixed throughout Monte Carlo replications at $[-1.96, 1.96]$, which captures approximately 95% of the observations of X. Furthermore, all three methods use an initial bandwidth h_x. To choose this, we use the approach of Delaigle and Gijbels (2004b), and investigate the sensitivity of our results to this choice below.

Finally, it is necessary to choose a grid of potential bandwidths to search over. From preliminary investigations, results appear insensitive to this choice, providing that the choice-set is large enough to include the optimally chosen bandwidth; we used $[0.16, 0.44]$ for all three methods. While the assumptions in Section 2 suggest different search ranges for the leave-one-out and out-of-bag methods, we found that both methods selected very similar bandwidths irrespective of the size of these bounds. In practice, we recommend plotting the estimated mean squared prediction error over a wide range of bandwidths to visually inspect that a global minimum has been found in each case.

In Table 1, we give results for the MISE of $\hat{m}$ between the 2.5^{th} and 97.5^{th} percentile of X. To ease comparison, all results are multiplied by 10 and we highlight in bold the optimal method for each DGP, sample size, error distribution, and error variance combination.

Table 1: MISE Results

	DGP 1 (Linear)							
Error Standard Deviation	$\sigma_\epsilon = 1/3$				$\sigma_\epsilon = 1/2$			
Error Type	OS		SS		OS		SS	
Sample Size	250	500	250	500	250	500	250	500
SIMEX	0.65	0.33	0.78	0.33	0.84	0.54	1.14	0.61
OOB	**0.50**	**0.29**	**0.54**	**0.30**	**0.65**	0.52	**0.68**	**0.54**
LOO	0.55	0.30	0.58	0.31	0.75	**0.46**	0.80	0.55
	DGP 2 (Quadratic)							
SIMEX	1.04	0.47	1.07	0.53	1.37	0.86	1.58	1.24
OOB	**0.72**	**0.41**	**0.78**	**0.41**	1.10	**0.61**	1.43	**0.70**
LOO	0.74	0.42	0.84	0.42	**1.05**	0.65	**1.20**	0.74
	DGP 3 (Cos)							
SIMEX	1.47	0.71	1.51	0.74	2.60	1.36	2.70	1.79
OOB	1.07	**0.62**	**1.21**	0.64	**1.61**	**1.02**	1.89	**1.11**
LOO	**1.06**	0.63	1.22	**0.64**	1.66	1.04	**1.84**	1.18
	DGP 4 (Sin)							
SIMEX	0.97	0.36	1.19	0.38	2.89	0.68	2.99	1.02
OOB	0.83	0.34	0.88	0.37	1.27	0.68	**1.36**	0.89
LOO	**0.69**	**0.34**	**0.77**	**0.36**	**1.15**	**0.62**	1.42	**0.77**

While neither the out-of-bag nor leave-one-out method dominates the other, both are preferable to the SIMEX approach in all parameter settings; although, the gap between SIMEX and our methods narrows with a larger sample size. As expected, the results for all three approaches improve as the sample size increases and as the measurement error noise decreases. Furthermore, each method shows better performance under ordinary smooth error relative to supersmooth. This stands in agreement with the theoretical literature which shows that the convergence rate of deconvolution based estimators deteriorates in the face of supersmooth error relative to ordinary smooth.

In Table 2, we give analogous results when the measurement error density is unknown. For this setting, a second noisy measure of X is generated as $W^r = X + \epsilon^r$, where ϵ^r is distributed identically to and independently of ϵ. The leave-one-out and out-of-bag methods are carried out as outlined in Section 2.3, with the characteristic function of ϵ estimated via the method of Delaigle, Hall and Meister (2008). The SIMEX approach is constructed as discussed in Section 3.4 of Delaigle and Hall (2008).

Table 2: MISE Results (Estimated Error Density)

	DGP 1 (Linear)							
Error Standard Deviation	$\sigma_\epsilon = 1/3$				$\sigma_\epsilon = 1/2$			
Error Type	OS		SS		OS		SS	
Sample Size	250	500	250	500	250	500	250	500
SIMEX	0.76	0.36	1.12	0.40	0.85	0.60	1.59	0.71
OOB	**0.50**	**0.28**	**0.51**	**0.31**	**0.63**	**0.47**	**0.66**	**0.49**
LOO	0.54	0.29	0.56	0.32	0.72	0.51	0.76	0.55
	DGP 2 (Quadratic)							
SIMEX	0.87	0.50	1.02	0.56	1.40	0.90	1.47	1.26
OOB	**0.70**	**0.38**	**0.74**	**0.43**	1.04	**0.59**	1.17	**0.72**
LOO	0.73	0.39	0.78	0.44	**1.04**	0.62	**1.12**	0.80
	DGP 3 (Cos)							
SIMEX	1.91	0.74	1.89	0.76	2.68	1.34	2.72	1.84
OOB	1.28	**0.57**	**1.31**	**0.66**	1.62	1.01	1.91	**1.18**
LOO	**1.07**	0.58	1.33	0.67	**1.57**	**1.01**	**1.73**	1.27
	DGP 4 (Sin)							
SIMEX	1.15	0.36	1.67	0.39	3.05	0.70	3.09	1.29
OOB	0.78	0.34	0.80	**0.36**	1.36	0.65	1.50	0.97
LOO	**0.67**	**0.34**	**0.71**	0.36	**1.14**	**0.61**	**1.30**	**0.86**

It is encouraging to see that when the measurement error density must be estimated, the MISE is virtually unaffected. Indeed, in several cases, the MISE is, in fact, lower when the density is estimated in comparison to the known density setting. Consequently, the findings are

very similar to those found in the known density case: the out-of-bag and leave-one-out methods generally provide comparable results, and both dominate the SIMEX approach.

To determine the sensitivity of the results to the pilot bandwidth, we proceed as follows. Denote by $h_{x,r}$ the pilot bandwidth selected using Delaigle and Gijbels (2004b) in the r^{th} Monte Carlo replication, and $h_r^*(h_{x,r})$ the optimal bandwidth selected using the pilot bandwidth $h_{x,r}$ in the r^{th} Monte Carlo replication. For each of the three considered methods, we calculate the sensitivity of the bandwidth choice to a smaller pilot bandwidth as $\frac{1}{r}\sum_{j=1}^{r}|MISE(h_r^*(h_x)) - MISE(h^*(0.5h_x))|$. To measure the sensitivity to a larger pilot bandwidth, we calculate $\frac{1}{r}\sum_{j=1}^{r} MISE(h_r^*(h_x)) - MISE(h^*(1.5h_x))|$. The results of this experiment are given in Tables 3 and 4 for the case when the measurement error density is estimated (where all results are again multiplied by 10 for ease of comparison).[6]

Table 3: Pilot Bandwidth Sensitivity (Smaller)

	DGP 1 (Linear)							
Error Standard Deviation	$\sigma_\epsilon = 1/3$				$\sigma_\epsilon = 1/2$			
Error Type	OS		SS		OS		SS	
Sample Size	250	500	250	500	250	500	250	500
SIMEX	0.31	0.24	0.62	0.51	0.24	0.25	0.39	0.31
OOB	0.11	0.07	0.19	0.15	0.21	0.13	0.30	0.21
LOO	0.31	0.19	0.41	0.32	0.50	0.31	0.60	0.51
	DGP 2 (Quadratic)							
SIMEX	0.51	0.34	0.72	0.58	0.78	0.43	0.49	0.46
OOB	0.21	0.15	0.39	0.27	0.38	0.26	0.58	0.49
LOO	0.53	0.26	0.77	0.57	1.04	0.58	1.13	1.00
	DGP 3 (Cos)							
SIMEX	1.53	0.44	0.71	0.55	1.58	0.46	0.75	0.75
OOB	0.30	0.19	0.55	0.39	0.57	0.37	0.83	0.69
LOO	0.81	0.45	1.20	0.61	1.46	0.94	1.69	1.41
	DGP 4 (Sin)							
SIMEX	0.35	0.26	0.33	0.25	0.23	0.29	0.29	0.18
OOB	0.18	0.13	0.28	0.22	0.25	0.22	0.31	0.23
LOO	0.39	0.24	0.55	0.31	0.61	0.45	0.69	0.53

Table 4: Pilot Bandwidth Sensitivity (Larger)

	DGP 1 (Linear)	
Error Standard Deviation	$\sigma_\epsilon = 1/3$	$\sigma_\epsilon = 1/2$

[6]Results for the known measurement error case were qualitatively similar.

Error Type	OS		SS		OS		SS	
Sample Size	250	500	250	500	250	500	250	500
SIMEX	0.19	0.11	0.57	0.15	0.28	0.20	0.28	0.22
OOB	0.11	0.08	0.12	0.10	0.14	0.10	0.21	0.15
LOO	0.17	0.07	0.18	0.11	0.25	0.14	0.28	0.22
	DGP 2 (Quadratic)							
SIMEX	0.30	0.16	0.36	0.18	0.43	0.30	0.63	0.50
OOB	0.33	0.18	0.36	0.19	0.54	0.33	0.74	0.57
LOO	0.28	0.13	0.34	0.15	0.46	0.28	0.67	0.49
	DGP 3 (Cos)							
SIMEX	0.44	0.27	0.61	0.50	0.96	0.48	1.01	1.04
OOB	0.68	0.39	0.77	0.53	1.06	0.67	1.36	1.17
LOO	0.50	0.26	0.65	0.48	0.87	0.52	1.13	0.94
	DGP 4 (Sin)							
SIMEX	0.37	0.21	0.55	0.40	0.30	0.25	0.45	0.30
OOB	0.44	0.40	0.45	0.42	0.48	0.51	0.29	0.45
LOO	0.36	0.19	0.48	0.40	0.30	0.29	0.47	0.31

In general, the out-of-bag approach shows less sensitivity to a smaller pilot bandwidth than either of the other two methods, with the leave-one-out approach showing the most sensitivity, particularly when the error variance is high. For a larger pilot bandwidth choice and a linear model, the out-of-bag method again shows the lowest sensitivity. However, in each of the other models, the out-of-bag approach shows the greatest sensitivity for a larger pilot bandwidth, while the SIMEX and out-of-bag approaches display a similar level of sensitivity across the board. Overall, it is encouraging to see that the methods proposed in this paper show a similar level of sensitivity to the initial bandwidth choice compared to the current state-of-the-art, while generally providing lower MISE.

4. Empirical Application

In this section, we apply our bandwidth selection procedures to data from the 2012-2013 and 2013-2014 waves of the National Health and Nutrition Examination Survey (NHANES). In particular, we estimate the relationship between systolic blood pressure (SBP) and cognitive ability. Recent studies (see, for example, Peters *et al.*, 2008, and Novak and Hajjar, 2010) have shown that a reduction in cognitive performance is not just a consequence of ageing but is also linked to hypertension (excessively high blood pressure) - a condition which generally increases with age. Hypertension is a widespread illness, affecting more than a third of the world's population (Pereira *et al.*, 2009) and is particularly prevalent in older individuals.

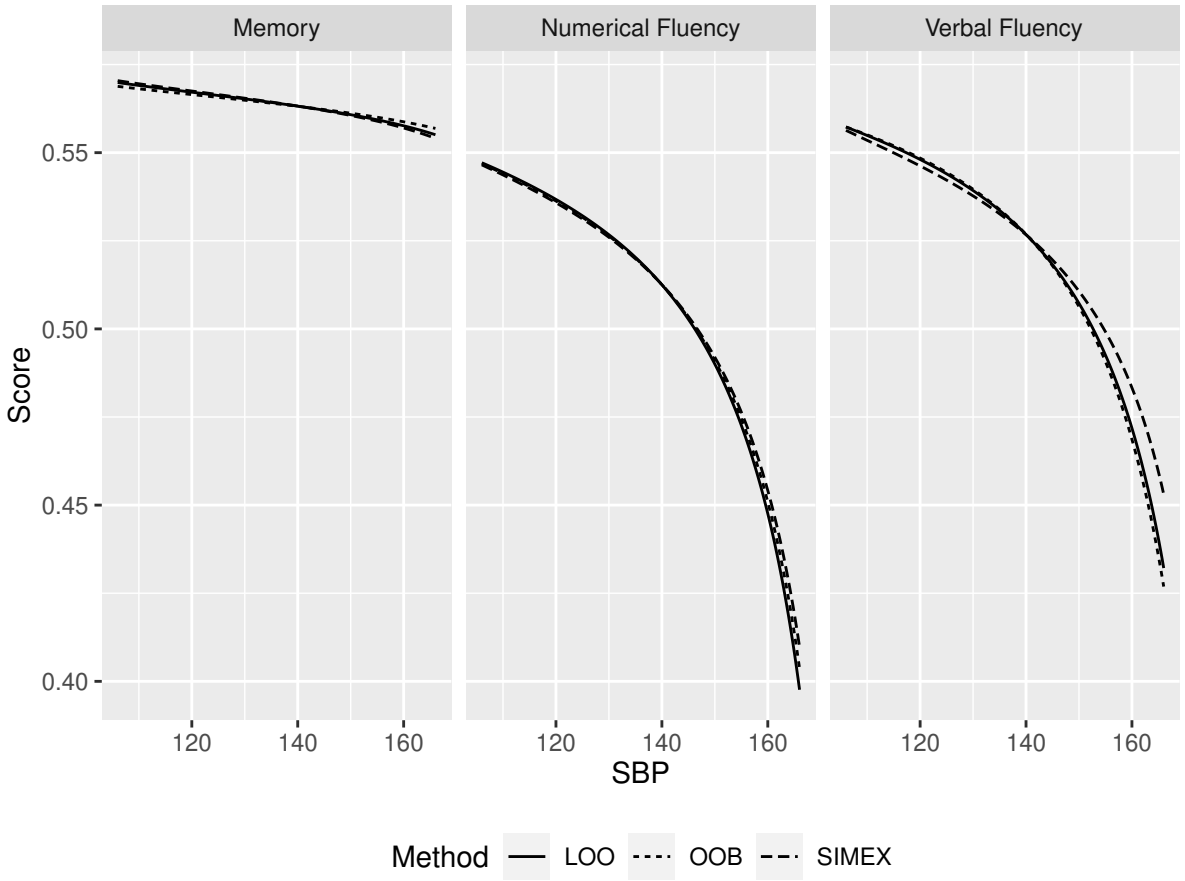
Previous research has also found a link between hypotension (excessively low blood pressure) and cognitive function (see, for example, Sabayan and Westendorp, 2015). These findings suggest that the effect of SBP on cognitive ability is nonlinear; hence, nonparametric regression estimation is likely to be appropriate. Furthermore, it is well-known that SBP measurements are prone to noise. This variation is due to, among other things, the time of day when the test is taken, the food recently eaten by the individual, and the individual's recent activity. As such, it is routine for measurements of SBP to be repeated. We use this repeated measure to estimate the measurement error characteristic function using the method of Delaigle, Hall and Meister (2008) and proceed with the bandwidth selection procedures outlined in Section 2.3.

The NHANES also has information on three measures of cognitive ability. The CERAD Word Learning Test is a standard tool to measure the ability to memorise verbal information. The Animal Fluency Test is used to examine verbal fluency and is commonly used to distinguish between those with normal cognitive function and those with mild or severe cognitive impairment. Finally, the Digit Symbol Substitution Test requires speed of thought, sustained concentration, and numerical fluency. Each test is designed to be culture-free and is administered in the language of the subject. We restrict attention to males between the ages of 60 and 80, giving a sample size of 1324, and standardise all variables to have unit variance.

To estimate the regression functions for each of the three test score outcomes, we use the deconvolution kernel estimator from Delaigle, Hall and Meister (2008). The bandwidth is chosen using the two methods of this paper and the SIMEX approach of Delaigle and Hall (2008) to provide a comparison. All parameter settings are the same as those used in Section 3. In Figure 1, we plot the estimated regression functions for the Memory Test, the Numerical Fluency Test, and the Verbal Fluency Test as a function of SBP.

In each case, the bandwidths chosen by each method are very close. The largest discrepancy is seen in the Memory Test regression, with the leave-one-out method choosing 0.64, the out-of-bag method choosing 0.71, and the SIMEX approach selecting 0.61. However, these bandwidth appear to make little difference in the function estimation, with each producing a relatively flat linear function. In the other two regressions, a more nonlinear shape is found, and all three methods give very similar bandwidth choices. Interestingly, the inverted "U" shape suggested by the previous literature is not apparent.

FIGURE 1. Estimated Regression Functions



Notes: The left pane plots the estimated regression functions from a regression of the Memory Test score (measured as a percentile in the sample) on SBP using the three different bandwidth selection mechanisms. Leave-one-out selected 0.64, out-of-bag selected 0.71, and SIMEX selected 0.61. The middle pane plots analogous regression functions for the Numerical Fluency Test. Leave-one-out selected 0.35, out-of-bag selected 0.36, and SIMEX selected 0.37. The right pane plots analogous regression functions for the Verbal Fluency Test. Leave-one-out selected 0.35, out-of-bag selected 0.34, and SIMEX selected 0.39.

Appendix A. Mathematical proofs

A.1. **Proof of Theorem 1.** Let $\hat{f}(x;a) = \frac{1}{na}\sum_{l=1}^{n}\mathbb{K}_a\left(\frac{x-W_l}{a}\right)$ denote the deconvolution kernel density estimator of $f(x)$, $\hat{f}_j(\cdot;a) = \frac{1}{(n-1)a}\sum_{l\neq j}\mathbb{K}_a\left(\frac{x-W_l}{a}\right)$ be the leave-j-out counterpart of $\hat{f}(\cdot;a)$, and $r_n(a) = (na)^{-1/2}r_\epsilon(a)\sqrt{\log(1/\sqrt{a})} + a^p$. For $k_1, k_2, k_3 \in \{1,2\}$, let

$$\hat{g}_{k_1,k_2,k_3}(x;h_{1,x},h) = \frac{1}{(nh_{1,x})^{k_2}((n-1)h)^{k_3}}\sum_{j=1}^{n}\{Y_j - m(x)\}^{k_1}\mathbb{K}_{h_{1,x}}^{k_2}\left(\frac{x-W_j}{h_{1,x}}\right)\mathbb{K}_h^{k_3}\left(\frac{x-W_j}{h}\right),$$

and define

$$\bar{R}_{LOO}(h) = \frac{1}{nh_{1,x}}\sum_{j=1}^{n}\int_{\mathcal{X}}\left\{\begin{array}{c}\{\hat{m}_j(x;h)-m(x)\}^2\\ -2\{Y_j-m(x)\}\{\hat{m}_j(x;h)-m(x)\}\end{array}\right\}\mathbb{K}_{h_{1,x}}\left(\frac{x-W_j}{h_{1,x}}\right)dx. \quad (10)$$

Then $h^*_{LOO} = \arg\min_{h\in[L_{1,n},H_{1,n}]}\bar{R}_{LOO}(h)$. Letting $\Delta_n(a) = R_n(a) - \bar{R}_{LOO}(a)$, the optimality of h^*_{LOO} implies

$$\begin{aligned} R_n(h^*_{LOO}) &= \bar{R}_{LOO}(h^*_{LOO}) + \Delta_n(h^*_{LOO}) \\ &\leq \bar{R}_{LOO}(h_{1,x}) + \Delta_n(h^*_{LOO}) = R_n(h_{1,x}) - \Delta_n(h_{1,x}) + \Delta_n(h^*_{LOO}). \quad (11)\end{aligned}$$

The conclusion then would follow if

$$R_n(h_{1,x}) \xrightarrow{p} 0, \quad (12)$$

$$\sup_{h\in[L_{1,n},H_{1,n}]}|\Delta_n(h)| \xrightarrow{p} 0. \quad (13)$$

For (12), as we can write

$$R_n(h_{1,x}) = \int_{\mathcal{X}}\frac{\{m(x)\hat{f}(x;h_{1,x}) - \hat{m}(x;h_{1,x})\hat{f}(x;h_{1,x})\}^2 f(x)}{\{\hat{f}(x;h_{1,x})\}^2}dx,$$

under Assumption (2) (f is bounded away from zero over $\mathcal{X}$), it is sufficient to show

$$\sup_{x\in\mathbb{R}}|\hat{m}(x;h_{1,x})\hat{f}(x;h_{1,x}) - m(x)\hat{f}(x;h_{1,x})| = o_p(1), \quad \sup_{x\in\mathbb{R}}|\hat{f}(x;h_{1,x}) - f(x)| = o_p(1),$$

which follow by Lemma 6, Assumption (2) (m is bounded), and Assumption (4) ($r_n(h_{1,x}) \to 0$).

For (13), first note that

$$\hat{m}(x;h) - m(x) = \frac{\{\hat{m}(x;h)-m(x)\}\hat{f}(x;h)}{f(x)} + \frac{\{\hat{m}(x;h)-m(x)\}\{f(x)-\hat{f}(x;h)\}}{f(x)}, \quad (14)$$

$$\hat{m}_j(x;h) - m(x) = \frac{\{\hat{m}_j(x;h)-m(x)\}\hat{f}_j(x;h)}{f(x)} + \frac{\{\hat{m}_j(x;h)-m(x)\}\{f(x)-\hat{f}_j(x;h)\}}{f(x)}, \quad (15)$$

and Lemma 6 and Assumption (5) guarantee that, in each case, the second term is negligible compared to the first term uniformly over $h \in [L_{1,n}, H_{1,n}]$. Thus, (14) and (15) imply that it is sufficient for (13) to show that $\sup_{h\in[L_{1,n},H_{1,n}]}|\Delta^*_n(h)| \xrightarrow{p} 0$, where

$$\Delta^*_n(h) = \int_{\mathcal{X}}\frac{1}{f^2(x)}\left\{\begin{array}{c} f(x)\{\hat{m}(x;h)\hat{f}(x;h) - m(x)\hat{f}(x;h)\}^2 \\ -\frac{1}{nh_{1,x}}\sum_{j=1}^{n}\{\hat{m}_j(x;h)\hat{f}_j(x;h) - m(x)\hat{f}_j(x;h)\}^2\mathbb{K}_{h_{1,x}}\left(\frac{x-W_j}{h_{1,x}}\right) \\ +\frac{2f(x)}{nh_{1,x}}\sum_{j=1}^{n}\{Y_j-m(x)\}\{\hat{m}_j(x;h)\hat{f}_j(x;h) - m(x)\hat{f}_j(x;h)\}\mathbb{K}_{h_{1,x}}\left(\frac{x-W_j}{h_{1,x}}\right)dx\end{array}\right\}.$$

Observe that

$$\begin{aligned}\hat{m}_j(x;h)\hat{f}_j(x;h) &= \frac{n}{n-1}\hat{m}(x;h)\hat{f}(x;h) - \frac{1}{(n-1)h}Y_j\mathbb{K}_h\left(\frac{x-W_j}{h}\right),\\ \hat{f}_j(x;h) &= \frac{n}{n-1}\hat{f}(x;h) - \frac{1}{(n-1)h}\mathbb{K}_h\left(\frac{x-W_j}{h}\right),\end{aligned}$$

which imply

$$\begin{aligned}\hat{m}_j(x;h)\hat{f}_j(x;h) - m(x)\hat{f}_j(x;h) &= \frac{n}{n-1}\{\hat{m}(x;h)\hat{f}(x;h) - m(x)\hat{f}(x;h)\}\\ &\quad -\frac{1}{(n-1)h}\{Y_j - m(x)\}\mathbb{K}_h\left(\frac{x-W_j}{h}\right). \qquad (16)\end{aligned}$$

By using (16), we can decompose $\Delta_n^*(h) = \sum_{\iota=1}^5 \Delta_{n,\iota}^*(h)$, where

$$\begin{aligned}\Delta_{n,1}^*(h) &= \frac{1-2n}{(n-1)^2}\int_{\mathcal{X}}\frac{\{\hat{m}(x;h)\hat{f}(x;h) - m(x)\hat{f}(x;h)\}^2}{f(x)}dx,\\ \Delta_{n,2}^*(h) &= -\frac{n^2}{(n-1)^2}\int_{\mathcal{X}}\frac{\{\hat{m}(x;h)\hat{f}(x;h) - m(x)\hat{f}(x;h)\}^2\{\hat{f}(x;h_{1,x}) - f(x)\}}{f^2(x)}dx,\\ \Delta_{n,3}^*(h) &= \frac{2n}{(n-1)}\int_{\mathcal{X}}\frac{\{\hat{m}(x;h)\hat{f}(x;h) - m(x)\hat{f}(x;h)\}\{\hat{m}(x;h_{1,x})\hat{f}(x;h_{1,x}) - m(x)\hat{f}(x;h_{1,x})\}}{f(x)}dx,\\ \Delta_{n,4}^*(h) &= \frac{2n}{(n-1)}\int_{\mathcal{X}}\frac{\{\hat{m}(x;h)\hat{f}(x;h) - m(x)\hat{f}(x;h)\}\hat{g}_{1,1,1}(x;h_{1,x},h)}{f^2(x)}dx,\\ \Delta_{n,5}^*(h) &= -\int_{\mathcal{X}}\left\{\frac{\hat{g}_{2,1,2}(x;h_{1,x},h)}{f^2(x)} + \frac{2\hat{g}_{2,1,1}(x;h_{1,x},h)}{f(x)}\right\}dx.\end{aligned}$$

For $\Delta_{n,1}^*(h)$, by Lemma 6 and Assumption (2) (f is bounded away from zero over $\mathcal{X}$), we have

$$\Delta_{n,1}^*(h) = O_p(n^{-1}r_n^2(h)), \qquad (17)$$

uniformly over $h \in [L_{1,n}, H_{1,n}]$. By similar arguments, we can show

$$\Delta_{n,2}^*(h) = O_p(r_n^2(h)r_n(h_{1,x})), \qquad (18)$$

$$\Delta_{n,3}^*(h) = O_p(r_n(h)r_n(h_{1,x})), \qquad (19)$$

uniformly over $h \in [L_{1,n}, H_{1,n}]$.

For $\Delta_{n,4}^*(h)$, note that

$$\begin{aligned}\sup_{x\in\mathbb{R}}|\hat{g}_{1,1,1}(x;h_{1,x},h)| &= O_p\left(\frac{1}{n^{7/8}h_{1,x}h}\int\frac{|K^{\mathrm{ft}}(t)|}{|f_\epsilon^{\mathrm{ft}}(t/h_{1,x})|}dt\int\frac{|K^{\mathrm{ft}}(t)|}{|f_\epsilon^{\mathrm{ft}}(t/h)|}dt\right)\\ &= O_p\left((n^{7/8}h_{1,x}h)^{-1}r_\epsilon(h_{1,x})r_\epsilon(h)\right),\end{aligned}$$

uniformly over $h \in [L_{1,n}, H_{1,n}]$, where the first equality follows by Assumption (2) (m is bounded) and $\max_{1\le j\le n}|Y_j| = O_p(n^{1/8})$ (by Lemma 3 and $E[Y^8] < \infty$ in Assumption (1)), and the second equality follows by Assumption (3) ($K^{\mathrm{ft}}(t)$ is supported on $[-1,1]$). Thus, by Assumption (2) (f is bounded away from zero over $\mathcal{X}$), Lemma 6 implies

$$\Delta_{n,4}^*(h) = O_p\left((n^{7/8}h_{1,x}h)^{-1}r_\epsilon(h_{1,x})r_\epsilon(h)r_n(h)\right), \qquad (20)$$

uniformly over $h \in [L_{1,n}, H_{1,n}]$.

By similar arguments, using $\max_{1\leq j\leq n}|Y_j|^2 = O_p(n^{1/4})$ under Assumption (2) ($E[Y^8]<\infty$), we can show

$$\Delta^*_{n,5}(h) = O_p\left((n^{7/4}h_{1,x}h^2)^{-1}r_\epsilon(h_{1,x})r^2_\epsilon(h) + (n^{3/4}h_{1,x}h)^{-1}r_\epsilon(h_{1,x})r_\epsilon(h)\right), \tag{21}$$

uniformly over $h\in[L_{1,n},H_{1,n}]$.

The statement in (13) then follows by (17)-(21), Assumptions (4)-(5), and

$$\begin{aligned}\sup_{h\in[L_{1,n},H_{1,n}]} r_\epsilon(h) &= r_\epsilon(L_{1,n}),\\ \sup_{h\in[L_{1,n},H_{1,n}]}|r_n(h)| &\leq (nL_{1,n})^{-1/2}r_\epsilon(L_{1,n})\sqrt{\log\left(1/\sqrt{L_{1,n}}\right)} + H^p_{1,n}.\end{aligned}$$

A.2. **Proof of Theorem 2.** Let $\hat{f}_{-b}(x;a) = \frac{1}{n^c_b a}\sum_{l\in\mathcal{I}^c_b}\mathbb{K}_a\left(\frac{x-W_l}{a}\right)$ denote the deconvolution kernel density estimator of $f(x)$ by leaving the bootstrap sample b out, $\hat{m}_{-b}(x;a) = \frac{\sum_{l\in\mathcal{I}^c_b}Y_l\mathbb{K}_a\left(\frac{x-W_l}{a}\right)}{\sum_{l\in\mathcal{I}^c_b}\mathbb{K}_a\left(\frac{x-W_l}{a}\right)}$ denote the deconvolution kernel regression estimator of $m(x)$ by leaving the bootstrap sample b out, and E^* and Var^* denote the conditional expectation and conditional variance, respectively, for the bootstrap resample given the original sample $\{Y_j,W_j\}_{j=1}^n$. Define

$$\bar{R}_{OOB}(h) = E^*\left[\frac{1}{n^c_b h_{2,x}}\sum_{j\in\mathcal{I}^c_b}\int_{\mathcal{X}}\left\{\begin{array}{c}\{\hat{m}_b(x;h)-m(x)\}^2\\ -2\{Y_j-m(x)\}\{\hat{m}_b(x;h)-m(x)\}\end{array}\right\}\mathbb{K}_{h_{2,x}}\left(\frac{x-W_j}{h_{2,x}}\right)dx\right].$$

Then $h^*_{OOB} = \arg\min_{h\in[L_{2,n},H_{2,n}]}\bar{R}_{LOO}(h)$. Letting $\Xi_n(a) = R_n(a) - \bar{R}_{OOB}(a)$, following similar arguments as in (11), the optimality of h^*_{OOB} implies

$$R_n(h^*_{OOB}) \leq R_n(h_{2,x}) - \Xi_n(h_{2,x}) + \Xi_n(h^*_{OOB}). \tag{22}$$

The conclusion would then follow by (22) and (12) if

$$\sup_{h\in[L_{2,n},H_{2,n}]}|\Xi_n(h)| \xrightarrow{p} 0. \tag{23}$$

First note that

$$\hat{m}_b(x;h) - m(x) = \frac{\{\hat{m}_b(x;h)-m(x)\}\hat{f}_b(x;h)}{f(x)} + \frac{\{\hat{m}_b(x;h)-m(x)\}\{f(x)-\hat{f}_b(x;h)\}}{f(x)}, \tag{24}$$

where the second term is dominated by the first. Thus, (14) and (24) imply that it is sufficient for (23) to show that $\sup_{h\in[L_{2,n},H_{2,n}]}|\Xi^*_n(h)|\xrightarrow{p}0$, where

$$\Xi^*_n(h) = E^*\left[\int_{\mathcal{X}}\frac{1}{f^2(x)}\left\{\begin{array}{c}f(x)\{\hat{m}(x;h)\hat{f}(x;h)-m(x)\hat{f}(x;h)\}^2\\ -\{\hat{m}_b(x;h)\hat{f}_b(x;h)-m(x)\hat{f}_b(x;h)\}^2\hat{f}_{-b}(x;h_{2,x})\\ +2f(x)\left\{\begin{array}{c}\{\hat{m}_b(x;h)\hat{f}_b(x;h)-m(x)\hat{f}_b(x;h)\}\\ \times\{\hat{m}_{-b}(x;h_{2,x})\hat{f}_{-b}(x;h_{2,x})-m(x)\hat{f}_{-b}(x;h_{2,x})\}\end{array}\right\}\end{array}\right\}dx\right].$$

Also note that $E^*[\hat{f}_b(x;h)] = \hat{f}(x;h)$ and $E^*[\hat{m}_b(x;h)\hat{f}_b(x;h)] = \hat{m}(x;h)\hat{f}(x;h)$, which allows us to decompose $\Xi^*_n(h) = \sum_{\iota=1}^5 \Xi^*_{n,\iota}(h)$, where

$$\Xi^*_{n,1}(h) = E^*\left[\int_{\mathcal{X}} \frac{\{\hat{m}(x;h)\hat{f}(x;h) - m(x)\hat{f}(x;h)\}^2\{f(x) - \hat{f}_{-b}(x;h_{2,x})\}}{f^2(x)}dx\right],$$

$$\Xi^*_{n,2}(h) = -E^*\left[\int_{\mathcal{X}} \frac{\{\hat{m}_b(x;h)\hat{f}_b(x;h) - \hat{m}(x;h)\hat{f}(x;h)\}^2\hat{f}_{-b}(x;h_{2,x})}{f^2(x)}dx\right],$$

$$\Xi^*_{n,3}(h) = -E^*\left[\int_{\mathcal{X}} \frac{\{m(x)\hat{f}(x;h) - m(x)\hat{f}_b(x;h)\}^2\hat{f}_{-b}(x;h_{2,x})}{f^2(x)}dx\right]$$

$$\Xi^*_{n,4}(h) = -2E^*\left[\int_{\mathcal{X}} \left\{ \begin{array}{c} \{\hat{m}_b(x;h)\hat{f}_b(x;h) - \hat{m}(x;h)\hat{f}(x;h)\} \\ \times\{m(x)\hat{f}(x;h) - m(x)\hat{f}_b(x;h)\} \end{array} \right\} \frac{\hat{f}_{-b}(x;h_{2,x})}{f^2(x)}dx\right],$$

$$\Xi^*_{n,5}(h) = 2E^*\left[\int_{\mathcal{X}} \left\{ \begin{array}{c} \{\hat{m}_b(x;h)\hat{f}_b(x;h) - m(x)\hat{f}_b(x;h)\} \\ \times\{\hat{m}_{-b}(x;h_{2,x})\hat{f}_{-b}(x;h_{2,x}) - m(x)\hat{f}_{-b}(x;h_{2,x})\} \end{array} \right\} \frac{1}{f(x)}dx\right].$$

Let $\hat{f}_{\tilde{n}}(x;h)$ denote the deconvolution density kernel estimator using sample size $\tilde{n} = \exp(-1)n$ (this represents the average number of observations in the out-of-bag sample, as shown in Breiman, 2001), then for $\Xi^*_{n,4}(h)$, uniformly over $h \in [L_{2,n}, H_{2,n}]$, we have

$$\begin{aligned} \left|\Xi^*_{n,4}(h)\right| &= \left|2\int_{\mathcal{X}} \frac{\hat{f}_{\tilde{n}}(x;h_{2,x})}{f^2(x)} m(x) E^*\left[\left(\hat{m}_b(x;h)\hat{f}_b(x;h) - \hat{m}(x;h)\hat{f}(x;h)\right)\left(\hat{f}(x;h) - \hat{f}_b(x;h)\right)\right]dx\right| \\ &\leq \left|2\int_{\mathcal{X}} \frac{\hat{f}_{\tilde{n}}(x;h_{2,x})}{f^2(x)} m(x)\sqrt{Var^*(\hat{m}_b(x;h)\hat{f}_b(x;h))}\sqrt{Var^*(\hat{f}_b(x;h))}dx\right| \\ &= O_p\left((nh)^{-1}r^2_\epsilon(h)\right), \end{aligned} \tag{25}$$

where the second step uses the Cauchy-Schwartz inequality, and the third follows from Lemma 6, Lemma 7, and Assumption (2) (f is bounded away from zero over $\mathcal{X}$).

For $\Xi^*_{n,1}(h)$, we write

$$\begin{aligned} \left|\Xi^*_{n,1}(h)\right| &= \left|\int_{\mathcal{X}} \{\hat{m}(x;h)\hat{f}(x;h) - m(x)\hat{f}(x;h)\}^2 \frac{\{f(x) - \hat{f}_{\tilde{n}}(x;h_{2,x})\}}{f^2(x)}dx\right| \\ &= O_p\left(r^2_n(h)r_{\tilde{n}}(h_{2,x})\right) = O_p\left(r^2_n(h)r_n(h_{2,x})\right), \end{aligned} \tag{26}$$

uniformly over $h \in [L_{2,n}, H_{2,n}]$, where the second step follows from Lemma 6 and Assumption (2), and the final step follows from $\tilde{n}$ being a constant multiple of n.

Turning to $\Xi^*_{n,3}(h)$, we have

$$\begin{aligned} \left|\Xi^*_{n,3}(h)\right| &= \left|\int_{\mathcal{X}} \frac{\hat{f}_{\tilde{n}}(x;h_{2,x})}{f^2(x)} m^2(x) E^*\left[\left(\hat{f}_b(x;h) - E^*\left[\hat{f}_b(x;h)\right]\right)^2\right]dx\right| \\ &= \int_{\mathcal{X}} \frac{\hat{f}_{\tilde{n}}(x;h_{2,x})}{f^2(x)} m^2(x) Var^*(\hat{f}_b(x;h))dx = O_p\left((nh)^{-1}r^2_\epsilon(h)\right), \end{aligned} \tag{27}$$

uniformly over $h \in [L_{2,n}, H_{2,n}]$, where the final equality follows from Lemma 6, Lemma 7, and Assumption (2). In a similar manner, we can write

$$\left|\Xi^*_{n,2}(h)\right| = \left| \int_{\mathcal{X}} \frac{\hat{f}_{\tilde{n}}(x; h_{2,x})}{f^2(x)} Var^*(\hat{m}_b(x;h)\hat{f}_b(x;h))dx \right| = O_p\left((nh)^{-1} r^2_\epsilon(h)\right), \tag{28}$$

uniformly over $h \in [L_{2,n}, H_{2,n}]$.

Finally, we characterise the bound for $\Xi^*_{n,5}(h)$. Using similar arguments to those before, we can write

$$\begin{aligned} \left|\Xi^*_{n,5}(h)\right| &= \left| 2\int_{\mathcal{X}} \frac{1}{f(x)} \{\hat{m}_{\tilde{n}}(x; h_{2,x})\hat{f}_{\tilde{n}}(x; h_{2,x}) - m(x)\hat{f}_{\tilde{n}}(x; h_{2,x})\}\{\hat{m}(x;h)\hat{f}(x;h) - m(x)\hat{f}(x;h)\}dx \right| \\ &= O_p\left(r_{\tilde{n}}(h_{2,x}) r_n(h)\right) = O_p\left(r_n(h_{2,x}) r_n(h)\right), \end{aligned} \tag{29}$$

uniformly over $h \in [L_{2,n}, H_{2,n}]$.

In summary, combining (25)-(29), the conclusion of the theorem follows by Assumptions (4') and (5').

A.3. **Lemmas.** For a probability measure Q on a measurable space $(S, \mathcal{S})$, let $L^2(Q)$ denote the space of all measurable functions $f : S \to \mathbb{R}$ such that $\|f\|_{Q,2} = \sqrt{\int |f|^2 dQ} < \infty$. For a class of measurable functions $\mathcal{F}$ on S such that $\mathcal{F} \subset L^2(Q)$, let $N(\mathcal{F}, \|\cdot\|_{Q,2}, \delta)$ denote the δ-covering number for $\mathcal{F}$ with respect to $\|\cdot\|_{Q,2}$. The class $\mathcal{F}$ is said to be pointwise measurable if there exists a countable subclass $\mathcal{G} \subset \mathcal{F}$ such that for every $f \in \mathcal{F}$ there exists a sequence $g_m \in \mathcal{G}$ with $g_m \to f$ pointwise. A function $F : S \to [0, \infty)$ is said to be an envelope for $\mathcal{F}$ if $F(x) \geq \sup_{f \in \mathcal{F}} |f(x)|$ for all $x \in S$.

Lemma 1. *[Chernozhukov* et al.*, 2014, Corollary 5.1] Let $X, X_1, \ldots, X_n$ be i.i.d. random variables taking values in a measurable space $(S, \mathcal{S})$, and let $\mathcal{F}$ be a pointwise measurable class of (measurable) real-valued functions on $\mathcal{S}$ with measurable envelope F. Suppose that there exist constants $A \geq e$ and $\nu \geq 1$ such that*

$$\sup_Q N(\mathcal{F}, \|\cdot\|_{Q,2}, \delta\|F\|_{Q,2}) \leq (A/\delta)^\nu,$$

for all $\delta \in (0, 1]$, where $\sup_Q$ is taken over all finitely discrete distributions on S. Furthermore, suppose that $0 < E[F^2(X)] < \infty$, and let $\sigma^2 > 0$ be any positive constant such that $\sup_{f \in \mathcal{F}} E[f^2(X)] \leq \sigma^2 \leq E[F^2(X)]$. Then, it holds

$$\begin{aligned} & E\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{n}} \sum_{j=1}^n \{f(X_j) - E[f(X)]\} \right| \right] \\ \leq\ & C\left\{ \sqrt{\nu\sigma^2 \log\left(\frac{A\sqrt{E[F^2(X)]}}{\sigma} \right)} + \frac{\nu B_n}{\sqrt{n}} \log\left(\frac{A\sqrt{E[F^2(X)]}}{\sigma} \right) \right\}, \end{aligned}$$

where $B_n = \sqrt{E[\max_{1 \leq j \leq n} F^2(X_j)]}$ and $C > 0$ is a universal constant.

Lemma 2. *Under Assumptions (1) - (3), for $s = 0, 1, 2$, it holds*

$$\sup_{x\in\mathbb{R}} E\left[\left|Y^s \mathbb{K}_a\left(\frac{x-W}{a}\right)\right|^2\right] = O\left(a r_\epsilon^2(a)\right).$$

Proof. Note that

$$\begin{aligned} E\left[\left|Y^s \mathbb{K}_a\left(\frac{x-W}{a}\right)\right|^2\right] &= \iint \left|\mathbb{K}_a\left(\frac{x-u-v}{a}\right)\right|^2 E[Y^{2s}|X=u] f(u) f_\epsilon(v) du dv \\ &= a \iint |\mathbb{K}_a(\tilde{u})|^2 E[Y^{2s}|X = x-v-a\tilde{u}] f(x-v-a\tilde{u}) f_\epsilon(v) d\tilde{u} dv \\ &= O\left(a \int |\mathbb{K}_a(\tilde{u})|^2 d\tilde{u}\right), \end{aligned}$$

where the first equality follows by Assumption (1) (ϵ and (Y, X) are independent), the second equality follows by the change of variables $\tilde{u} = (x-u-v)/a$, and the last equality follows by Assumption (2) ($E[Y^{2s}|X]$ and f are bounded). The conclusion then follows by Parseval's identity and Assumption (3) (K^{ft} is supported on $[-1, 1]$). □

Lemma 3. *[Kato and Sasaki, 2019, Lemma A.2] Let $\zeta_1, \ldots, \zeta_n$ be random variables such that $E[|\zeta_j|^r] < \infty$ for all $j = 1, \cdots, n$ and some $r \geq 1$. Then*

$$E\left[\max_{1\leq j\leq n} |\zeta_j|\right] \leq n^{1/r} \max_{1\leq j\leq n} (E[|\zeta_j|^r])^{1/r}.$$

Consider the class of functions

$$\mathcal{F}_n^{(s)} = \left\{(y, w) \mapsto y^s \mathbb{K}_a\left(\frac{x-w}{a}\right) : x \in \mathbb{R}\right\},$$

for $s = 0, 1, 2$. Let $F_n^{(s)}(y, w) = c_s |y|^s r_\epsilon(a)$ for some positive constant c_s. Then $F_n^{(s)}$ is an envelope function for $\mathcal{F}_n^{(s)}$ if c_s is large enough for each $s = 0, 1, 2$.

Lemma 4. *If K^{ft} is supported on $[-1, 1]$ and f_ϵ^{ft} does not vanish on $\mathbb{R}$, there exist constants A, $\nu \geq e$ independent of n such that*

$$\sup_Q N\left(\mathcal{F}_n^{(s)}, \|\cdot\|_{Q,2}, \delta \|F_n^{(s)}\|_{Q,2}\right) \leq (A/\delta)^\nu,$$

for all $\delta \in (0, 1]$ and $s = 0, 1, 2$, where $\sup_Q$ is taken over all finitely discrete distributions on $\mathbb{R}^2$.

Proof. Consider two classes of functions

$$\mathcal{K}_n = \left\{w \mapsto \mathbb{K}_a\left(\frac{x-w}{a}\right) : x \in \mathbb{R}\right\}, \quad \mathcal{Y}^{(s)} = \left\{y \mapsto y^s : x \in \mathbb{R}\right\}.$$

Let $K_n(w) = \kappa r_\epsilon(a)$ for some positive constant κ and $Y^{(s)}(y) = (c_s/\kappa)|y|^s$. Then $K_n(w)$ and $Y^{(s)}(y)$ are envelope functions of $\mathcal{K}_n$ and $\mathcal{Y}^{(s)}$ respectively, if κ is large enough. It is clear that $\mathcal{F}_n^{(s)} = \mathcal{Y}^{(s)} \cdot \mathcal{K}_n$, where $\mathcal{Y}^{(s)} \cdot \mathcal{K}_n = \left\{(y, w) \mapsto f_1(y) f_2(w) : f_1 \in \mathcal{Y}^{(s)}, f_2 \in \mathcal{K}_n\right\}$ is the pointwise product of $\mathcal{Y}^{(s)}$ and $\mathcal{K}_n$, and $F_n^{(s)}(y, w) = Y^{(s)}(y) K_n(w)$.

Since Corollary A.1 of Chernozhukov *et al.* (2014) implies $\sup_Q N(\mathcal{Y}^{(s)}, \|\cdot\|_{Q,2}, d) = 1$ for all $d > 0$, it is sufficient to show that there exist $A, \nu \geq e$ such that

$$\sup_Q N\left(\mathcal{K}_n, \|\cdot\|_{Q,2}, \delta\|K_n\|_{Q,2}\right) \leq (A/\delta)^\nu,$$

for all $\delta \in (0,1]$, which follows by Lemma 1 of Kato and Sasaki (2018). □

Lemma 5. *Under Assumption (1) - (3), for $s = 0, 1, 2$, it holds*

$$\sup_{x\in\mathbb{R}} \left| a^{-1} E\left[Y^s \mathbb{K}_a\left(\frac{x - W}{a}\right)\right] - E[Y^s|X = x]f(x)\right| = O(a^p).$$

Proof. Note that

$$\begin{aligned} E\left[Y^s \mathbb{K}_a\left(\frac{x-W}{a}\right)\right] &= \frac{1}{2\pi}\int e^{-\mathrm{i}tx/a}\{E[Y^s|X]f\}^{\mathrm{ft}}(t/a)K^{\mathrm{ft}}(t)dt \\ &= \frac{1}{2\pi}\int e^{-\mathrm{i}\tilde{t}x}\{E[Y^s|X]f\}^{\mathrm{ft}}(\tilde{t})\{aK^{\mathrm{ft}}(\tilde{t}a)\}d\tilde{t} \\ &= E\left[Y^s K\left(\frac{x-X}{a}\right)\right], \end{aligned}$$

where the first equality follows by the definition of $\mathbb{K}$ and Assumption (1) (ϵ and (Y, X) are independent), the second equality follows by the change of variables $\tilde{t} = t/a$, and the last equality follows by the convolution theorem and $aK^{\mathrm{ft}}(ta) = \{K(\cdot/a)\}^{\mathrm{ft}}(t)$.

Also note that

$$\begin{aligned} a^{-1}E\left[Y^s K\left(\frac{x-X}{a}\right)\right] &= a^{-1}\int E[Y^s|X = u]f(u)K\left(\frac{u-x}{a}\right)du \\ &= \int E[Y^s|X = x + a\tilde{u}]f(x + a\tilde{u})K(\tilde{u})d\tilde{u} \\ &= E[Y^s|X = x]f(x) + a^p\int \{E[Y^s|X]f\}^{(p)}(\bar{x}_{s,x,a\tilde{u}})\tilde{u}^p K(\tilde{u})d\tilde{u}, \end{aligned}$$

for some $\bar{x}_{s,x,a\tilde{u}} \in [x \pm a\tilde{u}]$, where the first equality follows by Assumption (3) (K is symmetric around zero), the second equality follows by the change of variables $\tilde{u} = (u - x)/a$, and the last equality follows by Assumption (2) ($E[Y^s|X]$ and f are p-time continuously differentiable) and Assumption (3) ($\int u^q K(u)du = 0$ for $q = 1, 2, \ldots, p-1$). So, the conclusion follows by Assumption (2) ($\{E[Y^s|X]$ and f have bounded derivatives up to p-th order) and Assumption (3) ($\int u^p K(u)du \neq 0$) and the conclusion follows. □

Lemma 6. *Suppose that Assumptions (1)-(3) hold true, $a \to 0$, and $(n^{1/2}a)^{-1}\log(1/\sqrt{a}) \to 0$ as $n \to \infty$. Then, for $s = 0, 1, 2$, it holds*

$$\sup_{x\in\mathbb{R}} \left| \frac{1}{na}\sum_{j=1}^n Y_j^s \mathbb{K}_a\left(\frac{x - W_j}{a}\right) - E[Y^s|X = x]f(x)\right| = O_p(r_n(a)).$$

Proof. First, we apply Lemma 1 to the class of functions $\mathcal{F}_n^{(s)}$ for $s = 0, 1, 2$. In particular, note that Lemma 2 implies $\sup_{f\in\mathcal{F}_n^{(s)}} E[f^2(Y, W)] = O(ar_\epsilon^2(a))$ and Lemma 3 and Assumption (1) ($E[Y^8] < \infty$) implies $\max_{1\leq j\leq n} Y_j^{2s} = O_p(n^{s/4})$, which gives $\sqrt{E[\max_{1\leq j\leq n}\{F_n^{(s)}(Y_j, W_j)\}^2]} =$

$O(n^{s/8} r_\epsilon(a))$. Thus, Lemma 4 and $(n^{1/2}a)^{-1}\log(1/\sqrt{a}) \to 0$ implies

$$\sup_{x\in\mathbb{R}} \left| \frac{1}{na} \sum_{j=1}^{n} Y_j^s \mathbb{K}_a\left(\frac{x-W_j}{a}\right) - a^{-1} E\left[Y^s \mathbb{K}_a\left(\frac{x-W}{a}\right)\right] \right| = O_p\left((na)^{-1/2} r_\epsilon(a) \sqrt{\log(1/\sqrt{a})} \right).$$

Therefore, the conclusion follows by Lemma 5. □

Lemma 7. *[Delaigle and Gijbels, 2004b, Proposition 4.2] Suppose that Assumptions (1)-(3) hold. Then*

$$\begin{aligned} \int_{\mathcal{X}} Var^*(\hat{f}_b(x;h))dx &= O_p((nh)^{-1} r_\epsilon^2(h)), \\ \int_{\mathcal{X}} Var^*(\hat{m}_b(x;h)\hat{f}_b(x;h))dx &= O_p((nh)^{-1} r_\epsilon^2(h)). \end{aligned}$$

References

[1] Bartalotti, O., Brummet, Q. and S. Dieterle (2020) A correction for regression discontinuity designs with group-specific mismeasurement of the running variable, forthcoming in *Journal of Business & Economics Statistics*.

[2] Breiman, L. (2001) Random forests, *Machine Learning*, 45, 5-32.

[3] Blattman, C., Jamison, J., Koroknay-Palicz, T., Rodrigues, K. and M. Sheridan (2016) Measuring the measurement error: A method to qualitatively validate survey data, *Journal of Development Economics*, 120, 99-112.

[4] Calonico, S., Cattaneo, M. D. and M. H. Farrell (2018) On the effect of bias estimation on coverage accuracy in nonparametric inference, *Journal of the American Statistical Association*, 113, 767-779.

[5] Calonico, S., Cattaneo, M. D. and M. H. Farrell (2020) Coverage error optimal confidence intervals for local polynomial regression, Working paper.

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

[7] Chernozhukov, V., Chetverikov, D. and K. Kato (2014) Gaussian approximation of suprema of empirical processes, *Annals of Statistics*, 42, 1564-1597.

[8] Chichignoud, M., Hoang, V. H., Ngoc, T. M. P. and V. Rivoirard (2017) Adaptive wavelet multivariate regression with errors in variables, *Electronic Journal of Statistics*, 11, 682-724.

[9] Davezies, L. and T. L. Barbanchon (2017) Regression discontinuity design with continuous measurement error in the running variable, *Journal of Econometrics*, 200, 260-281.

[10] Delaigle, A. and I. Gijbels (2004a) Practical bandwidth selection in deconvolution kernel density estimation, *Computational Statistics & Data Analysis*, 45, 249-267.

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

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

[13] Delaigle, A., Hall, P. and F. Jamshidi (2015) Confidence bands in non-parametric errors-in-variables regression, *Journal of the Royal Statistical Society: Series B*, 149-169.

[14] Delaigle, A., Hall, P. and A. Meister (2008) On deconvolution with repeated measurements, *Annals of Statistics*, 36, 665-685.

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

[16] Fan, J. (1991) On the optimal rates of convergence for nonparametric deconvolution problems, *Annals of Statistics*, 19, 1257-1272.

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

[18] Hall, P. and A. Meister (2007) A ridge-parameter approach to deconvolution, *Annals of Statistics*, 35, 1535-1558.

[19] Härdle, W., Hall, P. and J. S. Marron (1988) How far are automatically chosen regression smoothing parameters from their optimum? *Journal of the American Statistical Association*, 83, 86-95.

[20] Härdle, W. and J. S. Marron (1985) Optimal bandwidth selection in nonparametric regression function estimation, *Annals of Statistics*, 13, 1465-1481.

[21] Kato, K. and Y. Sasaki (2018) Uniform confidence bands in deconvolution with unknown error distribution, *Journal of Econometrics*, 207, 129-161.

[22] Kato, K. and Y. Sasaki (2019) Uniform confidence bands for nonparametric errors-in-variables regression, *Journal of Econometrics*, 213(2), 516-555.

[23] Li, T. and Q. Vuong (1998) Nonparametric estimation of the measurement error model using multiple indicators, *Journal of Multivariate Analysis*, 65, 139-165.

[24] Masry, E. (1993) Asymptotic normality for deconvolution estimators of multivariate densities of stationary processes, *Journal of Multivariate Analysis*, 44, 47-68.

[25] McMurry, T. L. and D. N. Politis (2004) Nonparametric regression with infinite order flat-top kernels, *Journal of Nonparametric Statistics*, 16, 549-562.

[26] Meister, A. (2009) *Deconvolution Problems in Nonparametric Statistics*, Springer.

[27] Novak, V. and I. Hajjar (2010) The relationship between blood pressure and cognitive function, *Nature Reviews Cardiology*, 7, 686-698.

[28] Pereira, M., Lunet, N., Azevedo, A. and H. Barros (2009) Differences in prevalence, awareness, treatment and control of hypertension between developing and developed countries, *Journal of Hypertension*, 27, 963-975.

[29] Peters, R., Beckett, N., Forette, F., Tuomilehto, J., Clarke, R., Ritchie, C., Waldman, A., Walton, I., Poulter, R., Ma, S. and M. Comsa (2008) Incident dementia and blood pressure lowering in the Hypertension in the Very Elderly Trial cognitive function assessment (HYVET-COG): a double-blind, placebo controlled trial, *Lancet Neurology*, 7, 683-689.

[30] Sabayan, B. and R. G. Westendorp (2015) Blood pressure control and cognitive impairment—why low is not always better, *JAMA Internal Medicine*, 175, 586-587.

[31] Schennach, S. M. (2016) Recent advances in the measurement error literature, *Annual Review of Economics*, 8, 341-377.

[32] Stefanski, L. A. and R. J. Carroll (1990) Deconvolving kernel density estimators, *Statistics*, 21, 169-184.

[33] Wong, W. H. (1983) On the consistency of cross-validation in kernel nonparametric regression, *Annals of Statistics*, 11, 1136-1141.

Department of Economics, Southern Methodist University, 3300 Dyer Street, Dallas, TX 75275, US.

Email address: haod@smu.edu

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.

Email address: t.otsu@lse.ac.uk

Department of Economics and Business Economics, Fuglesangs Allé 4 building 2631, 12 8210 Aarhus V, Denmark

Email address: lntaylor@econ.au.dk