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 errors-in-variables regression model. 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 the approaches to the SIMEX method of Delaigle and Hall (2008) in a Monte Carlo study. As well as enjoying considerable advantages in terms of computational cost, the out-of-bag procedure is shown to produce the most stable results and generally leads to lower MISE in finite samples.

1. Introduction

Measurement error is rife in social science 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, bias can be introduced in estimation, masking the true relationship between variables and rendering testing procedures invalid. Moreover, measurement error can be particularly troublesome when using nonparametric methods, which have become commonplace in applied work as a result of increases in computing power and data availability. A vital concern when using any nonparametric technique is the choice of bandwidth, which has led to high demand for robust, data-driven methods to select this parameter.

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

To this end, the current paper considers 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

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

and $\epsilon \in \mathbb{R}$ is a classical measurement error independent of (Y, X) with 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 (Fan and Truong, 1993)

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

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

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

with an (ordinary) kernel function $K : \mathbb{R} \to \mathbb{R}$ and bandwidth parameter h.

In this paper, we develop two novel bandwidth selection procedures for the deconvolution regression estimator. Each method is based on evaluating the prediction errors

from the regression using a second (density) deconvolution. The first 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 leave-one-out method and the existing SIMEX approach of Delaigle and Hall (2008), the out-of-bag procedure can dramatically reduce computational cost in larger samples. This is particularly pertinent for nonparametric deconvolution estimators which suffer from slow rates of convergence and thus require more data. Moreover, as suggested by our simulation results, the out-of-bag approach also provides better - and more stable - results than either of the other procedures.

Deconvolution methods within statistics have predominantly focused on density estimation and regression models 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 different 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 error is unknown (for example, Delaigle, Hall and Meister, 2008). For a detailed review on this ever-growing literature 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 the 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 mean integrated squared error (MISE), and Delaigle and Gijbels (2004a) developed a bootstrap

method. However, there has been much 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 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 systolic 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. 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 (OOB) 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

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

where K_y is an ordinary kernel function and (h_y, h_x) are bandwidth parameters for this density estimation. Since Y is error-free, we apply the deconvolution kernel $\mathbb{K}$ 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

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

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

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

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

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

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

$$h^*_{LOO} = \underset{h\in[L_n,H_n]}{\operatorname{argmin}}\, \tilde{R}_{LOO}(h),$$

where L_n and H_n are deterministic sequences characterising the upper and lower bounds of the region to search for h^*, respectively.

In order to implement $\tilde{R}_{LOO}(h)$, an auxiliary bandwidth h_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 a regression bandwidth selector that plays a similar role to the pilot bandwidth in our case. As in these papers, we find that our method is relatively insensitive to this initial choice, providing that it is selected in a sensible manner. 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.

To establish the asymptotic validity of $\tilde{R}_{LOO}(h)$, based on Wong (1983), we focus on the following integrated squared error loss[1]

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

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), $E[Y^8] < \infty$, 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): *$(n^{1/2}h_x)^{-1} \log(1/\sqrt{h_x}) \max\{n^{-1/2} r^2_\epsilon(h_x), 1\} \to 0$ and $h_x \to 0$ as $n \to \infty$.*

[1]For the error-free case, Wong (1983) considered the average squared error loss $n^{-1}\sum_{j=1}^n \{\hat{m}(X_j) - m(X_j)\}^2$ 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 $R_n(h)$.

(5): $(n^{1/2}L_n)^{-1}\log(1/\sqrt{L_n})\max\{n^{-1/2}r_\epsilon^2(L_n),1\}\to 0$, $(n^{3/4}h_xL_n)^{-1}r_\epsilon(h_x)r_\epsilon(L_n)\to 0$, *and* $H_n\to 0$ *as* $n\to\infty$.

Assumption (1) requires random sampling and some regularity conditions. In particular, $E[Y^8]<\infty$ is used in Lemma 3 in Appendix to control the order of $\max_{1\le j\le n}|Y_j|^4$. The non-vanishing condition for f_ϵ^{ft} is commonly employed in kernel-based deconvolution methods and is satisfied for many distributions. Our method may be extended to the case where f_ϵ^{ft} 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, as well as 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 for deconvolution problems, the Fourier transform of K is further required to be compactly supported. Assumption (4) imposes restrictions on the auxiliary bandwidth h_x.[2] In particular, to establish the uniform rate of convergence for the estimands based on h_x, we need Lemma 6 in Appendix, which requires $(n^{1/2}h_x)^{-1}\log(1/\sqrt{h_x})\to 0$. The other conditions are 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_n, we only require $H_n=o(1)$. However, the conditions on the lower bound L_n are more complicated, as it depends on both the

[2]In the ordinary smooth case when f_ϵ is of order α and K is of order β, the MSE optimal bandwidth $h_x^*\sim n^{-1/(1+2\alpha+2\beta)}$ and $r_\epsilon(h_x^*)\sim h_x^{*^{-\alpha}}=n^{\alpha/(1+2\alpha+2\beta)}$. Then Assumption (4) can be satisfied by h_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_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 result can be obtained for the supersmooth case.

choice of the auxiliary bandwidth h_x and the smoothness of the error distribution, which is reflected by $r_\epsilon(\cdot)$.

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 $R_n(h^*_{LOO}) \leq R_n(h_x) + O\left(\sup_{h\in[L_n,H_n]} |\tau_n(h)|\right)$ (see eq. (2) in Appendix A.1), where $\tau_n(\cdot)$ depends on the auxiliary bandwidth, h_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 Appendix, $R_n(h_x) = O_p(r_n^2(h_x))$, where $r_n(h_x) = (nh_x)^{-1/2} r_\epsilon(h_x)\sqrt{\log(1/\sqrt{h_x})}$. Thus, if we further impose $\sup_{h\in[L_n,H_n]} |\tau_n(h)| r_n^{-2}(h_x) \to 0$, then $R_n(h^*_{LOO}) \leq R_n(h_x)(1 + o(1))$, which shows that h^*_{LOO} asymptotically improves the auxiliary bandwidth h_x since it leads to a smaller value for R_n.

2.2. **Out-Of-Bag Approach.** While the leave-one-out approach presented in the last subsection is theoretically reasonable, it is practically less so. With a relatively large sample size, and a sensibly sized grid of candidate bandwidths, the number of leave-j-estimates $\hat{m}_j(\cdot; h)$ to be calculated can become overwhelming and the computational cost excessive; this is particularly true for deconvolution kernel estimators which are more computationally expensive to estimate than ordinary kernel estimators.

As an alternative which circumvents this problem, we suggest the following 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 make up approximately 36.8% of the total sample size (Breiman, 2001). For each $b = 1, \ldots, B$, the out-of-bag bootstrap counterpart of (5) can be obtained as

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

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.,

$$h_{OOB}^* = \underset{h \in [L_n, H_n]}{\operatorname{argmin}} \tilde{R}_{OOB}(h).$$

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. This is undesirable. 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, which 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 the following slight relaxation of Assumptions (4) and (5).

Assumption.

(4'): $(nh_x)^{-1}r_\epsilon^2(h_x)\log(1/\sqrt{h_x}) \to 0$ *and* $h_x \to 0$ *as* $n \to \infty$.

(5'): $(nL_n)^{-1}r_\epsilon^2(L_n)\log(1/\sqrt{L_n}) \to 0$ *and* $H_n \to 0$ *as* $n \to \infty$.

Theorem 2. *Under Assumptions (1)-(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_ϵ^{ft} 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_ϵ^{ft}, denoted $\hat{f}_\epsilon^{\text{ft}}$, we can construct both the leave-one-out and out-of-bag criterion functions as

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

where

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

Note that, as in Delaigle, Hall and Meister (2008), it is not necessary to estimate leave-one-out or out-of-bag estimates of $\hat{f}_\epsilon^{\text{ft}}$ in $\hat{\bar{R}}_{LOO}(h)$ or $\hat{\bar{R}}_{OOB}(h)$. 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.

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:

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

where $h_x = (h_{x,1}, \ldots, h_{x,d_X})$ is a vector of auxiliary bandwidths to estimate the density of f_{YX}, $h_X = (h_{X,1}, \ldots, h_{X,d_X})$ and $h_Z = (h_{Z,1}, \ldots, h_{Z,d_Z})$ are the bandwidths to be optimised, $\prod h_x$ is understood as the product of the elements of h_x, $\mathbb{K}_{d_X}$ is a d_X-dimensional deconvolution kernel function 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

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

with an ordinary kernel function K; $\hat{m}_b$ is 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 very computationally demanding even if the dimensions

of X and Z are small. Given the considerably lower computational cost of the out-of-bag procedure, we suggest practitioners to use this approach rather than the leave-one-out method.

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:

$$\text{DGP1} : m(x) = x,$$

$$\text{DGP2} : m(x) = x - x^2,$$

$$\text{DGP3} : m(x) = \cos(x),$$

$$\text{DGP4} : m(x) = \sin(x).$$

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 error 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 500 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

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

A trimming term is used in the denominator of the regression estimator to ensure stability in the tails of the distribution. Specifically, all values of the estimated density of X (the denominator of the estimator) below 0.01 are set to 0.01.

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 several times with the results averaged to get a final estimate for the optimal bandwidth. As well as being computationally demanding, requiring the calculation of $2nBH$ (where B is the number of repetitions and H is the number of candidate bandwidths) deconvolution kernel regression estimators, it also requires knowledge of the distribution of the measurement error. If auxiliary data is available, this distribution can be estimated via deconvolution methods, requiring the choice of an additional bandwidth parameter. In contrast, the OOB method we propose requires estimating only BH estimators and both the LOO and OOB procedures are easily extended to the case of unknown error without the need for further deconvolution estimation.

All three methods require choosing a range of integration χ; we set this between the 5^{th} and 95^{th} percentile of W. Furthermore, all three methods use an initial bandwidth

h_x. To choose this, we use the approach of Delaigle and Gijbels (2004b), and discuss the sensitivity of our results to this choice below. Finally, it is necessary to choose a grid of potential bandwidths from which to select the optimal choice $[L_n, H_n]$. Results are insensitive to this choice, providing that the choice-set is large enough to include the optimally chosen bandwidth.

We found that the performance of the OOB procedure converged with a relatively small number of bootstrap resamples; thus, only 50 were used in these simulations (this was also the number used by Efron and Tibshirani, 1997, in a similar bootstrap cross-validation procedure). While for the SIMEX approach, convergence was achieved with only 20 resamples. In Table 1, we give results for the median integrated squared error (MedISE) of m between the 5^{th} and 95^{th} percentile of X. Results for the mean integrated squared error (MISE) are given in Table 2. To ease comparison, all results are multiplied by 100 and we highlight in bold the optimal method for each DGP, sample size, error distribution, and error variance combination.

Table 1: Simulation Results - Median Integrated Squared Error

	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	0.34	0.23	0.28	0.15	0.55	0.45	0.38	0.23
OOB	**0.28**	**0.20**	**0.24**	0.14	**0.41**	**0.34**	**0.31**	**0.19**
LOO	0.32	0.21	0.26	**0.13**	0.52	0.43	0.31	0.19

	DGP 2 (Quadratic)							
SIMEX	0.45	0.28	0.35	0.18	0.72	0.58	0.58	0.33
OOB	0.36	**0.24**	**0.31**	0.17	**0.57**	**0.40**	0.54	0.28
LOO	**0.35**	0.26	0.32	**0.16**	0.59	0.53	**0.43**	**0.28**
	DGP 3 (Cos)							
SIMEX	0.57	0.46	0.43	0.23	1.04	0.93	0.82	0.46
OOB	**0.51**	**0.36**	0.44	0.22	0.82	**0.60**	0.84	0.43
LOO	0.52	0.45	**0.39**	**0.22**	**0.80**	0.90	**0.64**	**0.41**
	DGP 4 (Sin)							
SIMEX	0.63	0.44	0.37	0.21	1.11	0.93	0.60	0.37
OOB	0.62	0.43	0.39	0.21	**0.89**	**0.87**	0.57	0.34
LOO	**0.60**	**0.41**	**0.35**	**0.19**	1.11	0.96	**0.52**	**0.31**

Table 2: Simulation Results - Mean Integrated Squared Error

	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	3.20	0.33	1.22	0.21	351	30.7	31.2	7.84
OOB	**0.39**	**0.27**	**0.29**	**0.17**	194	**0.59**	**0.44**	**0.24**
LOO	5.51	0.92	0.32	0.17	**133**	17.0	36.9	0.26

	DGP 2 (Quadratic)							
SIMEX	70.2	0.42	0.50	0.25	790	87.2	150	18.9
OOB	**10.3**	**0.32**	**0.41**	0.22	**170**	**0.65**	**0.67**	**0.36**
LOO	10.8	0.37	0.42	**0.21**	275	1.43	105	1.16
	DGP 3 (Cos)							
SIMEX	85.6	0.67	0.75	0.32	571	99.1	99.3	87.5
OOB	**0.72**	**0.49**	0.55	0.29	**485**	**0.99**	**1.04**	0.56
LOO	85.5	0.64	**0.54**	**0.28**	591	50.5	6.85	**0.52**
	DGP 4 (Sin)							
SIMEX	24.2	0.53	1.07	0.25	446	12.7	16.2	1.62
OOB	**0.75**	**0.51**	**0.44**	0.24	197	**1.21**	**0.66**	**0.38**
LOO	21.4	0.56	0.45	**0.23**	**133**	11.8	19.2	0.40

While neither the OOB nor LOO method dominates the other, both are preferable to the SIMEX approach. However, as can be seen in Table 2, the OOB method is considerably more stable than either of the other approaches. In many cases, particularly when the sample size is small or the measurement error noise is large, the SIMEX and LOO methods have very large MISE, indicating some erratic results; this is rarely seen for the OOB approach.

As is expected, the results for all methods improve with the sample size and as the measurement error noise decreases. However, it is somewhat surprising to see that each of the three methods show better performance in the case of supersmooth error in comparison to ordinary smooth. This contradicts the theoretical literature which shows the convergence

rate of deconvolution based estimators deteriorates in the face of supersmooth error relative to ordinary smooth. A possible explanation for this finding is that most bandwidth selection procedures tend to select a bandwidth that oversmooths slightly. If this is the case for the three methods presented here, since the optimal bandwidth for supersmooth error is larger than for ordinary smooth, this tendency to oversmooth may have a less detrimental effect in the supersmooth case.

To determine the sensitivity of the results to the pilot bandwidth choice, 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. Also, let $h_r^*(h_{x,r})$ denote the optimal bandwidth selected in the r^{th} Monte Carlo replication using the pilot bandwidth $h_{x,r}$ For each of the three considered methods, we calculate the sensitivity of the bandwidth choice to a smaller pilot bandwidth as $\frac{1}{r}\sum_{j=1}^{r}\{h_r^*(h_x) - h_{LOO}^*(0.5h_x)\}^2$. To measure the sensitivity to a larger pilot bandwidth, we calculate $\frac{1}{r}\sum_{j=1}^{r}\{h_r^*(h_x) - h_{LOO}^*(1.5h_x)\}^2$. The results are given in Tables 3 and 4 below.

Table 3: Simulation Results - 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	0.66	0.67	1.11	1.17	1.06	0.85	2.11	1.82
OOB	0.40	0.30	0.53	0.42	0.82	0.52	1.73	1.40
LOO	0.44	0.37	0.69	0.53	0.57	0.52	1.48	1.34

	DGP 2 (Quadratic)							
SIMEX	0.65	0.64	1.22	1.11	1.20	0.91	2.04	1.90
OOB	0.34	0.24	0.55	0.43	0.68	0.41	1.51	1.29
LOO	0.35	0.29	0.76	0.76	0.47	0.36	1.30	1.17
	DGP 3 (Cos)							
SIMEX	0.44	0.43	1.03	0.90	0.78	0.57	1.61	1.71
OOB	0.20	0.13	0.34	0.27	0.47	0.26	1.14	1.10
LOO	0.23	0.17	0.48	0.48	0.30	0.22	1.07	1.13
	DGP 4 (Sin)							
SIMEX	0.50	0.44	0.88	0.88	1.02	0.64	1.82	1.48
OOB	0.32	0.22	0.36	0.23	0.61	0.34	1.30	0.98
LOO	0.26	0.23	0.43	0.24	0.38	0.32	1.06	0.86

Table 4: Simulation Results - Pilot Bandwidth Sensitivity (Larger)

	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	0.22	0.25	0.30	0.29	0.31	0.31	0.42	0.41
OOB	0.24	0.20	0.18	0.17	0.34	0.26	0.39	0.33
LOO	0.20	0.13	0.20	0.20	0.24	0.18	0.29	0.26

	DGP 2 (Quadratic)							
SIMEX	0.30	0.19	0.32	0.23	0.50	0.34	0.55	0.51
OOB	0.44	0.29	0.37	0.28	0.59	0.43	0.63	0.53
LOO	0.23	0.17	0.25	0.19	0.38	0.24	0.51	0.37
	DGP 3 (Cos)							
SIMEX	0.37	0.24	0.35	0.25	0.50	0.39	0.67	0.56
OOB	0.55	0.40	0.55	0.37	0.71	0.55	0.85	0.69
LOO	0.32	0.24	0.33	0.24	0.47	0.35	0.63	0.50
	DGP 4 (Sin)							
SIMEX	0.35	0.33	0.31	0.29	0.40	0.39	0.51	0.49
OOB	0.50	0.44	0.46	0.38	0.53	0.46	0.53	0.53
LOO	0.34	0.24	0.28	0.25	0.37	0.31	0.45	0.43

First, it should be noted that all three methods show somewhat similar levels of sensitivity to the pilot bandwidth choice. Nevertheless, when comparing the sensitivity of the three approaches to a smaller pilot bandwidth choice, the OOB method is the least sensitive for the ordinary smooth case, while the LOO method is the least sensitive for the supersmooth setting; the SIMEX approach shows the most sensitivity across both settings.

For a larger pilot bandwidth choice, the LOO approach shows the least sensitivity across all settings. In general, the SIMEX approach is less sensitive to a larger pilot bandwidth than the OOB method. However, when the regression function is linear, the OOB method shows less sensitivity than the SIMEX procedure.

4. Empirical Application

In this section, we apply our bandwidth selection procedures to data from the 2013-2014 wave 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 and follows an inverted "U" shape; hence, nonparametric estimation should be used. 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 systolic blood pressure to be repeated. We use this repeated measurement to help determine the nature of the measurement error.

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 recall new 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 memory. Each test is designed to be culture-free

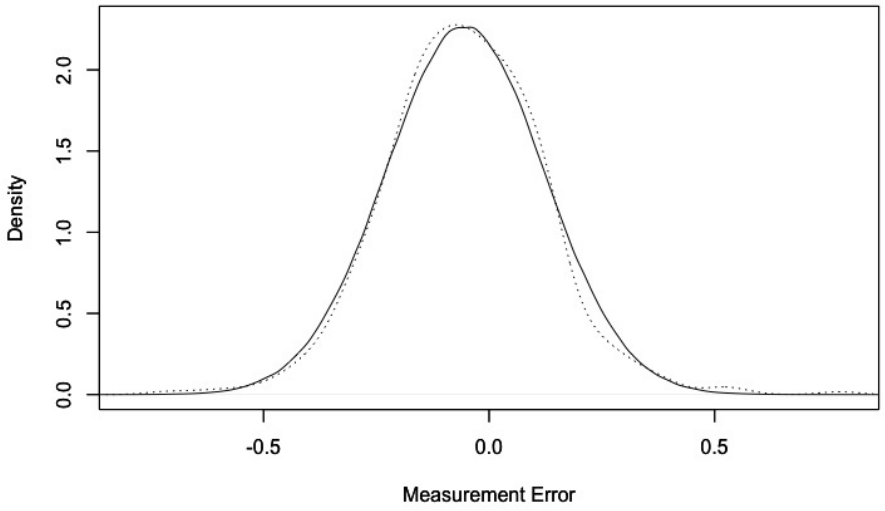
and is administered in the subject's language. We restrict attention to males between the ages of 60 and 80, giving a sample size of 531, and standardise all variables to have unit variance.

Denote by $W^{(1)}$ and $W^{(2)}$ two noisy measures of SBP. Also, let

$$\begin{aligned} W_j^{(1)} &= X_j + \epsilon_j^{(1)}, \\ W_j^{(2)} &= X_j + \epsilon_j^{(2)}, \end{aligned}$$

where X_j is the true (long term) SBP of individual j and $\epsilon_j^{(1)}$ and $\epsilon_j^{(2)}$ are zero mean independent and symmetrically distributed classical measurement errors. As the noisy measurement of SBP, we use $W = (W^{(1)} + W^{(2)})/2$. This implies that the measurement error is $\epsilon = (\epsilon^{(1)} + \epsilon^{(2)})/2$ which has the same distribution as $(W^{(1)} - W^{(2)})/2$. In Figure 1, we plot this distribution and compare it with a Gaussian density. The mean is equal to -0.05 (which cannot be rejected as being equal to 0 by a conventional t-test at the 5% significance level) and the standard deviation is 0.18. Although a Shapiro-Wilk test rejects the null hypothesis of Normality, it is clear that the distributions are close. Thus, we assume our measurement error follows a Gaussian distribution.

FIGURE 1. Distribution of Measurement Error



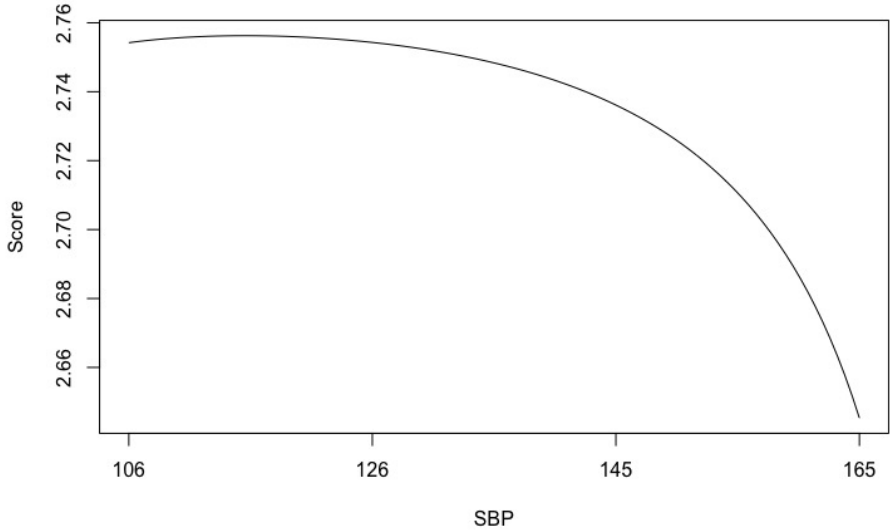
Notes: This figure plots the estimated distribution of the measurement error (dashed line), together with a Gaussian distribution with mean of -0.05 and standard deviation of 0.18 (solid line).

To estimate the regression functions for each of the three test score outcomes, we use the deconvolution kernel estimator and choose the bandwidth using the OOB put forward in this paper.[3] All parameter settings are the same as those used in Section 3. In Figure 2, we plot the estimated regression for the Word Learning Test score as a function of SBP. The bandwidth selected is 0.39. There is a clear nonlinear relationship, however, the effect is small; moving from an SBP reading of 106 to 165 (from the 5^{th} percentile to the 95^{th}) decreases the test score by 0.1 of a standard deviation. Furthermore, the inverted "U" shape found in the previous literature is not apparent in this context.

In Figures 3 and 4, analogous regression estimates are plotted for the Fluency Test score and the Symbol Substitution Test score. In the former, the selected bandwidth is 0.31, and in the latter 0.435. Both regressions follow the same general shape as the Word Learning Test; however, the effect is much larger in each of these cases.

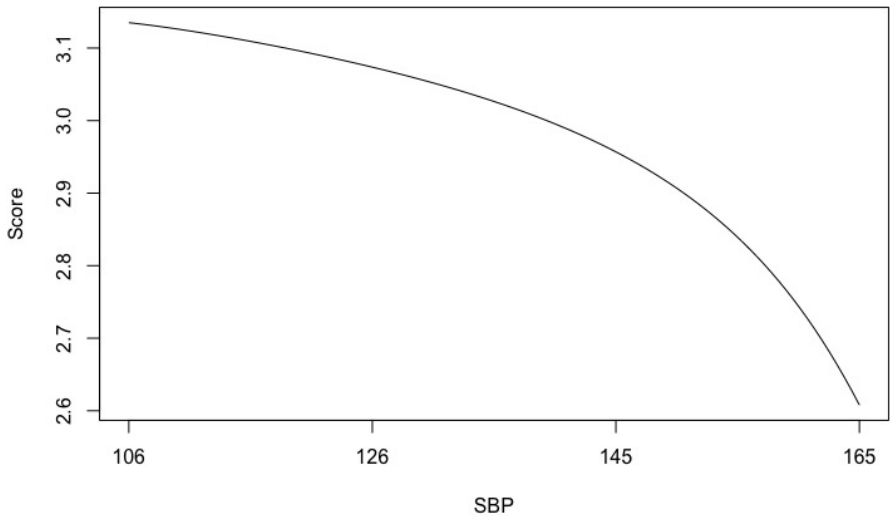
[3]The LOO approach selected similar bandwidths in each case.

FIGURE 2. Word Learning Test Regression



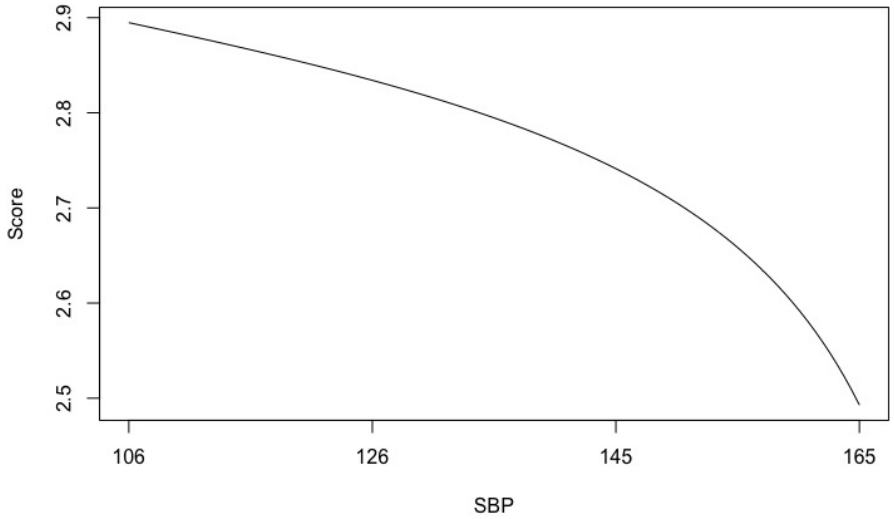
Notes: This figure plots the estimated regression function from a regression of the Word Learning Test score on SBP using the deconvolution kernel estimator. The bandwidth used is equal to 0.39.

FIGURE 3. Fluency Test Regression



Notes: This figure plots the estimated regression function from a regression of the Fluency Test score on SBP using the deconvolution kernel estimator. The bandwidth used is equal to 0.435.

FIGURE 4. Symbol Substitution Regression



Notes: This figure plots the estimated regression function from a regression of the Symbol Substitution Test score on SBP using the deconvolution kernel estimator. The bandwidth used is equal to 0.31.

References

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

[2] 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.

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

[4] 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.

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

[6] 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.

[7] 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.

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

[9] 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.

[10] Efron, B. and R. Tibshirani (1997) Improvements on cross-validation: the 632+ bootstrap method, *Journal of the American Statistical Association*, 92, 548-560.

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

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

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

[14] 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.

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

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

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

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

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

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

[21] 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.

[22] 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.

[23] 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.

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

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

[26] 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