NONPARAMETRIC INSTRUMENTAL REGRESSION WITH ERRORS IN VARIABLES

KARUN ADUSUMILLI AND TAISUKE OTSU

Abstract. This paper considers nonparametric instrumental variable regression when the endogenous variable is contaminated with classical measurement error. Existing methods are inconsistent in the presence of measurement error. We propose a wavelet deconvolution estimator for the structural function that modifies the generalized Fourier coefficients of the orthogonal series estimator to take into account the measurement error. We establish the convergence rates of our estimator for the cases of mildly/severely ill-posed models and ordinary/super smooth measurement errors. We characterize how the presence of measurement error slows down the convergence rates of the estimator. We also study the case where the measurement error density is unknown and needs to be estimated, and show that the estimation error of the measurement error density is negligible under mild conditions as far as the measurement error density is symmetric.

1. Introduction

This paper is concerned with estimation of the nonparametric instrumental regression function where the explanatory variable is measured with error

$$\begin{aligned} Y &= m(X^*) + U, \qquad E[U|W] = 0, \\ X &= X^* + \epsilon, \qquad \epsilon \perp (X^*, W). \end{aligned} \tag{1}$$

We wish to estimate the function m based on the observables of (Y, X, W), where Y is a response variable, X is a mismeasured version of the explanatory variable X^* due to the measurement error ϵ, and W is an observable instrumental variable. The variables (X^*, U, ϵ) are unobservable. The disturbance U may be correlated with the error-free but unobservable X^* so that $E[U|X^*]$ may not vanish (i.e., X^* may be endogenous). However, we can access an instrumental variable W that is observable and satisfies mean independence $E[U|W] = 0$. The measurement error ϵ enters additively and is independent from X^* and W, but it is allowed to be correlated with Y.

When X^* is observable, the structural function m can be identified by solving the integral equation $E[Y|X^* = \cdot] = m(\cdot)$ under certain conditions. This is typically an ill-posed inverse problem that calls for some regularization scheme to obtain a useful estimator for m. Several regularized estimators have been proposed in the literature, such as Newey and Powell (2003), Hall and Horowitz (2005), Blundell, Chen and Kristensen (2007), Darolles, Fan, Florens and Renault (2010), Horowitz (2011, 2012), and Gagliardini and Scaillet (2012). However, these

Otsu gratefully acknowledges financial support from the ERC Consolidator Grant (SNP 615882).

existing methods are generally invalid when the explanatory variable contains a measurement error.[1]

On the other hand, when the disturbance U satisfies mean independence $E[U|X^*] = 0$, then the estimation problem of m turns into one of nonparametric regression with errors in variables. This is another kind of ill-posed inverse problem and various deconvolution estimation methods are available in the literature, such as Fan and Truong (1993), Hall and Meister (2007), and Delaigle and Hall (2008).[2] However, these estimation methods are generally inconsistent when the explanatory variable is endogenous.

In reality these issues, i.e. endogeneity and measurement error in X^*, can occur at the same time. Currently there is no valid estimation method for m in the model (1), that is available in the literature. In this paper we propose an estimation method for m based on the orthogonal series estimation method (Horowitz, 2011, 2012) and the wavelet deconvolution technique (Pensky and Vidakovic, 1999, and Fan and Koo, 2002) to deal with endogeneity and measurement error, respectively. In particular, we propose a wavelet deconvolution estimator for the structural function m that modifies the generalized Fourier coefficients of the orthogonal series estimator to take into account the measurement error. An advantage of our approach is that a single tuning parameter is shown to be sufficient to characterize the mean squared error (MSE) risk under both endogeneity and measurement error. We establish the convergence rates of our estimator for the cases of mildly/severely ill-posed models and ordinary/super smooth measurement errors. Furthermore, we characterize how the presence of measurement error slows down the convergence rates of the estimator. Indeed we find that it does so in a fashion that is strikingly similar to the way ill-posedness of the instrumental regression model also affects the rates.

We also study the case where the measurement error density is unknown and needs to be estimated, and show that the estimation error of the measurement error density is negligible under mild conditions as far as the measurement error density is symmetric. In particular, we find that the convergence rates under known and unknown error distributions are equivalent if either the joint density of (X^*, W) is smoother than the error density, or the error density is supersmooth.

The rest of the paper is organized as follows. Section 2 introduces the setup and wavelet deconvolution estimator. Section 3 studies the asymptotic properties of the estimator when the measurement error density is known. Section 4 considers the case where the measurement error density is unknown and needs to be estimated. Appendix contains the proofs of our theoretical results.

Notation. Throughout the paper, let $|\cdot|$ be the Euclidean norm for the Euclidean space $\mathbb{R}^d$ and complex space $\mathbb{C}^d$, $\|f\|_2 = (\int |f(x)|^2 dx)^{1/2}$ be the L_2-norm of a function $f : \mathbb{R}^d \to \mathbb{C}$, $L_2(\mathbb{R}^d) = \{f : \|f\|_2 < \infty\}$ be the L_2-space, and $\langle f, g \rangle = \int f(x)\overline{g(x)}dx$ be the inner product in $L_2(\mathbb{R}^d)$, where $\bar{c}$ denotes the complex conjugate of $c \in \mathbb{C}$. Also, let $\mathrm{i} = \sqrt{-1}$ and $f^{\mathrm{ft}}(t) = \int e^{\mathrm{i}tx} f(x)dx$ be the Fourier transform of f.

[1]On the other hand, for the case where X^* is correctly measured but the instrumental variable W is mismeasured, we can apply the conventional estimation methods using noisy measurements of W as far as the measurement error in W is independent from other variables.

[2]See Meister (2009) for a review on deconvolution methods.

2. Estimator

2.1. **General construction.** Suppose we observe a random sample $\{Y_i, X_i, W_i\}_{i=1}^n$ of $(Y, X, W) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}$.[3] For simplicity we focus on the case where both X and W are scalar. See Section 2.4 below for the generalization to the vector case. Let f_{X^*W} be the joint density of (X^*, W), and define the integral operator $A : L_2(\mathbb{R}) \to L_2(\mathbb{R})$ as

$$(Ag)(w) = \int g(x) f_{X^*W}(x, w) dx. \tag{2}$$

Denote $r(w) = E[Y|W = w] f_W(w)$, where f_W is the density of W. If $m, r \in L_2(\mathbb{R})$ and $f_{X^*W} \in L_2(\mathbb{R}^2)$, then the relation in (1) may be described by the following integral equation (which is a Fredholm equation of the first kind)

$$Am = r. \tag{3}$$

In order to estimate the function m of interest that solves (3), we replace the operator A and the function r with some series estimators and then solve the empirical counterpart of (3).

More precisely, we choose a complete orthonormal basis $\{\psi_k\}_{k\in\mathcal{J}}$ of real-valued functions for $L_2(\mathbb{R})$. For standard orthonormal bases such as splines or cosines, we set $\mathcal{J} = \mathbb{N}$. In the next subsection, we argue that a wavelet basis is particularly suitable for deconvolution problems. Based on the basis $\{\psi_k\}_{k\in\mathcal{J}}$, we can expand

$$r = \sum_{k\in\mathcal{J}} a_k\psi_k, \qquad m = \sum_{l\in\mathcal{J}} b_l\psi_l, \qquad f_{X^*W} = \sum_{k,l\in\mathcal{J}} c_{k,l}\psi_k\psi_l, \tag{4}$$

where $a_k = \langle r, \psi_k\rangle$, $b_l = \langle m, \psi_l\rangle$, and $c_{k,l} = \langle f_{X^*W}, \psi_k\psi_l\rangle$. We note that $\{\psi_j\psi_k\}_{j,k\in\mathcal{J}}$ forms a complete orthonormal basis for $L_2(\mathbb{R}^2)$. Since the generalized Fourier coefficient a_k is written as $a_k = E[Y\psi_k(W)]$, we estimate it by the sample counterpart

$$\hat{a}_k = \frac{1}{n}\sum_{i=1}^n Y_i\psi_k(W_i). \tag{5}$$

On the other hand, the coefficient $c_{k,l}$ involves the joint density f_{X^*W} of the observable W and unobservable X^*. Therefore, its recovery requires a deconvolution technique. From the assumption $\epsilon \perp (X^*, W)$, we can see that $f_{XW}^{\text{ft}}(t, s) = f_{X^*W}^{\text{ft}}(t, s) f_\epsilon^{\text{ft}}(t)$ for all $t, s \in \mathbb{R}$, where f_ϵ^{ft} is the Fourier transform of f_ϵ, the density of ϵ. Assuming f_ϵ^{ft} does not vanish anywhere on the real line, $f_{X^*W}^{\text{ft}}$ may be identified by $f_{X^*W}^{\text{ft}} = f_{XW}^{\text{ft}}/f_\epsilon^{\text{ft}}$. Thus, using the Plancherel isometry (see, e.g., Meister, 2009, Theorem A.4), the coefficient $c_{j,k}$ can be expressed as

$$c_{k,l} = \frac{1}{(2\pi)^2}\left\langle \frac{f_{XW}^{\text{ft}}}{f_\epsilon^{\text{ft}}}, \psi_k^{\text{ft}}\psi_l^{\text{ft}}\right\rangle. \tag{6}$$

Suppose that f_ϵ^{ft} is known (the case of unknown f_ϵ^{ft} will be discussed in Section 4). Then by estimating $f_{XW}^{\text{ft}}(t, s)$ with the sample counterpart $\hat{f}_{XW}^{\text{ft}}(t, s) = \frac{1}{n}\sum_{i=1}^n e^{\mathrm{i}(tX_i+sW_i)}$, we can

[3] In the literature of nonparametric instrumental regression, several papers assumed that X and W are compactly supported (e.g., Hall and Horowitz, 2005, Horowitz, 2011, 2012, and Darolles, Fan, Florens and Renault, 2011). However, since X contains measurement error in our setup, the compact support assumption is restrictive.

estimate $c_{k,l}$ as follows

$$\hat{c}_{k,l} = \frac{1}{(2\pi)^2}\left\langle \frac{\hat{f}^{\mathrm{ft}}_{XW}}{f^{\mathrm{ft}}_{\epsilon}}, \psi^{\mathrm{ft}}_k \psi^{\mathrm{ft}}_l \right\rangle = \frac{1}{n}\sum_{i=1}^{n} \xi_k(X_i)\psi_l(W_i), \tag{7}$$

where $\xi_k(X) = \frac{1}{2\pi}\int e^{\mathrm{i}tX} \frac{\overline{\psi^{\mathrm{ft}}_k(t)}}{f^{\mathrm{ft}}_{\epsilon}(t)} dt$. Based on these estimators of the generalized Fourier coefficients, the function r and operator A can be estimated as

$$\hat{r}(w) = \sum_{k=1}^{J_n} \hat{a}_k \psi_k(w), \qquad (\hat{A}g)(w) = \sum_{k=1}^{J_n}\sum_{l=1}^{J_n} g_k \hat{c}_{k,l}\psi_l(w),$$

for $g \in L_2(\mathbb{R})$, where $g_k = \langle g, \psi_k\rangle$ and the integer J_n is a smoothing parameter satisfying $J_n \to \infty$ at a suitable rate. Then our estimator of m in the model (1) is obtained by solving the sample analog $\hat{A}\hat{m} = \hat{r}$ of (3) with respect to $\hat{m}$. In particular, the solution $\hat{m}$ may be explicitly written as

$$\hat{m}(x) = \sum_{k=1}^{J_n} \hat{b}_k \psi_k(x), \tag{8}$$

where the coefficients $\hat{\mathbf{b}} = (\hat{b}_1, \ldots, \hat{b}_{J_n})'$ are given by

$$\hat{\mathbf{b}} = (\mathbf{W}'\mathbf{X})^{-1}\mathbf{W}'\mathbf{Y},$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, and $\mathbf{W}$ and $\mathbf{X}$ are $n \times J_n$ matrices whose (i,k)-th elements are $\psi_k(W_i)$ and $\xi_k(X_i)$, respectively.

Note that the above formula for $\hat{\mathbf{b}}$ is identical to the conventional instrumental variable estimator. Our estimator takes the same form as the sieve estimator of Horowitz (2011, 2012) except for the matrix $\mathbf{X}$. If there is no measurement error, the (i,k)-th element of $\mathbf{X}$ becomes $\psi_k(X_i)$. To deal with the measurement error, we replace the elements of $\mathbf{X}$ with their deconvolution counterparts $\xi_k(X_i) = \frac{1}{2\pi}\int e^{\mathrm{i}tX_i} \frac{\overline{\psi^{\mathrm{ft}}_k(t)}}{f^{\mathrm{ft}}_{\epsilon}(t)} dt$.

2.2. **Wavelet deconvolution estimator.** In order to implement our series-based estimator in (8), we have to choose a basis for $L_2(\mathbb{R})$. In this paper, we suggest employing the wavelet basis. In particular, the band limited Meyer-type wavelet is suitable for deconvolution problems (see, Pensky and Vidakovic, 1990, and Fan and Koo, 2002).

Define the functions

$$\phi_{j;k}(x) = 2^{j/2}\phi(2^j x - k), \qquad \varphi_{j';k}(x) = 2^{j'/2}\varphi(2^{j'} x - k),$$

for $j, j' \in \mathbb{N}$ and $k \in \mathbb{Z}$, where $\phi, \varphi \in L_2(\mathbb{R})$ are chosen to satisfy some particular properties so as to make them father and mother wavelets, respectively. Take some $j \in \mathbb{N}$. The multi-resolution expansion theorem states that $\{\phi_{j;k}, \varphi_{j';k}\}_{j' \geq j, k \in \mathbb{Z}}$ form an orthonormal basis for $L_2(\mathbb{R})$ and thus any function $f \in L_2(\mathbb{R})$ can be represented as

$$f = \sum_{k \in \mathbb{Z}} c_{j,k}\phi_{j;k} + \sum_{j' \geq j}\sum_{k \in \mathbb{Z}} d_{j',k}\varphi_{j';k}.$$

Intuitively, the index j accounts for the resolution (fine scale structure) captured by the wavelets while k is simply a translation term.

We still have to choose the father and mother wavelets, ϕ and φ. Following Pensky and Vidakovic (1999), ϕ and φ are defined using their Fourier transforms as

$$\phi^{\mathrm{ft}}(t) = (P[t-\pi, t+\pi])^{1/2}, \qquad \varphi^{\mathrm{ft}}(t) = e^{-\mathrm{i}t/2}(P[|t|/2-\pi, |t|-\pi])^{1/2}, \tag{9}$$

for some probability measure P compactly supported on $[-\pi/3, \pi/3]$. We choose P such that its density is symmetric around 0. This ensures that ϕ, φ, and the orthonormal basis $\{\phi_{j;k}, \varphi_{j';k}\}_{j' \geq j, k \in \mathbb{Z}}$ are all real valued. In addition, we may take P smooth enough so that ϕ^{ft} and φ^{ft} are q times continuously differentiable. We call the wavelets satisfying the above properties as wavelets of order q.

In this paper we use only the linear part of the multi-resolution expansion. In other words, we employ the sieve space $\mathcal{L}_\phi^{(n)}$ spanned by $\{\phi_{j_n;k}\}_{|k| \leq L_n}$ for some resolution level j_n and series length L_n. Thus, our wavelet estimator of m is given by (8), where $\{\psi_j\}_{j \in J_n}$ is replaced with $\{\phi_{j_n;k}\}_{|k| \leq L_n}$.

The linear wavelet space $\mathcal{L}_\phi^{(n)}$ based on (9) is particularly convenient for deconvolution problems. Let $J_n = 4\pi 2^{j_n}/3$. Indeed the space $\mathcal{L}_\phi^{(n)}$ satisfies

$$\mathcal{L}_\phi^{(n)} \subseteq \{h \in L_2(\mathbb{R}) : h^{\mathrm{ft}} \text{ is supported on } \mathcal{C}_n \equiv [-J_n, J_n]\},$$

for any L_n (see, Meister, 2009, p. 17). Therefore, any function in $\mathcal{L}_\phi^{(n)}$ has a compactly supported Fourier transform (known as the band limited property) and J_n plays the role of a smoothing parameter. This property is important to control the estimation variance of the deconvolution estimators whose upper bound typically involves the term $\left\{\min_{|t| \leq J_n} |f_\epsilon^{\mathrm{ft}}(t)|\right\}^{-1}$ (see, Theorem 1 below). Thus, without the band limited property, it is not easy to control such term without additional smoothing, such as a ridge parameter.

We present two approximation properties of the linear wavelets, which are generalizations of Pensky and Vidakovic (1999, Lemma 2 and Theorem 3). These lemmas are used to establish the convergence rate of the wavelet deconvolution estimator for m. Let $P_{j_n} : L_2(\mathbb{R}) \to \mathcal{L}_\phi^{(n)}$ be the projection operator onto the space spanned by $\{\phi_{j_n;k}\}_{k \in \mathbb{Z}}$.

Lemma 1. *Suppose $f \in L_2(\mathbb{R}^d)$ satisfies $\int (1+|t|^2)^\alpha |f^{\mathrm{ft}}(t)|^2 e^{2\rho|t|^v} dt = C < \infty$ for some $\alpha, \rho, v \geq 0$. Then for some $c > 0$,*

$$\|f - P_{j_n} f\|_2 \leq cC 2^{-j_n \alpha} \exp\{-\rho(2\pi/3)^{dv} 2^{j_n v}\}.$$

Lemma 2. *Suppose $f \in L_2(\mathbb{R}^d)$ satisfies $\sup_{x \in \mathbb{R}^d} |x|^{(1+\eta)/2} |f(x)| < \infty$ for some $\eta > 0$. Then for wavelets $\{\phi_{j;k}, \varphi_{j';k}\}_{j' \geq j, k \in \mathbb{Z}}$ of order $q \geq (1+\eta)/2$, it holds*

$$\sum_{h=1}^{d} \sum_{|k_h| \geq L_n} |c_{j_n;k_1 \ldots, k_d}|^2 = O(2^{j_n d}/L_n^\eta).$$

Lemma 1 bounds the approximation error from leaving out the nonlinear part $\{\varphi_{j';k}\}_{j' \geq j, k \in \mathbb{Z}}$ from the multi-resolution expansion. This error depends on the smoothness of f characterized by the decay rate of its Fourier transform. Lemma 2 bounds the approximation error from truncating the linear wavelet estimator after a particular number of terms, L_n. In this case the approximation error depends on the decay rate of the function f in the tail. In the next section, we will see that to derive the convergence rate of our estimator for m, there is no constraint on the upper bound of the rate at which L_n is allowed to diverge to infinity. Therefore, theoretically we can choose L_n as large as possible. Only practical considerations prevent us from taking L_n very large. On the other hand, j_n (or equivalently J_n) plays the role of a standard smoothing parameter and should be chosen carefully to take into account the bias and variance trade-off.

2.3. **Other estimation methods.** The wavelet based estimator is not the only estimation method that is conceivable. The main advantage of the wavelet method presented in Section 2.2 is that it provides a unified and simple framework to tackle both endogeneity and measurement error issues with a single tuning parameter J_n. Here we provide a brief summary of other possible methods and their relative merits and drawbacks.

Hall and Meister (2007) provided a ridge parameter approach to density deconvolution. We may adapt this approach to the series estimation method of Section 2.1 to obtain an alternative estimate, $\breve{c}_{k,l}$, of $c_{k,l}$:

$$\breve{c}_{k,l} = \frac{1}{(2\pi)^2}\left\langle \frac{\hat{f}^{\text{ft}}_{XW}}{\max\{f^{\text{ft}}_{\epsilon}, h\}}, \psi^{\text{ft}}_k \psi^{\text{ft}}_l \right\rangle.$$

Here h is the ridge parameter function of the form $h(t) = n^{-\varsigma}|t|^a$, where $\varsigma > 0$ and $a \geq 0$ are tuning parameters. In contrast to the wavelet series approach of Section 2.2, this approach remains valid even if $f^{\text{ft}}_{\epsilon}(t) = 0$ at some t, and for arbitrary orthonormal bases including wavelets. However such an estimation method requires choosing two or more tuning parameters (i.e., the series length J_n and other ridge tuning parameters), which increases the complexity of the estimation.

An alternative approach to sieve based methods for nonparametric instrumental variable regression is using kernel methods (Hall and Horowitz, 2005, and Darolles, Fan, Florens and Renault, 2011). Supplementing them with the deconvolution kernel (e.g., Stefanski and Carroll, 1990, and Fan, 1991) can enable us to extend these methods to allow for measurement error. However the essential difficulty in this case seems to stem from the fact that the kernel methods suggested so far are based on the assumption of compact support for X^* and W. Extensions of the kernel methods to unbounded support, as we require here, would necessitate the presence of a trimming function in the kernel density.

2.4. **Vector case.** For the theoretical analysis in the next section, we concentrate on the case where both X and W are scalar for simplicity. However, it is possible to extend our estimation approach to the case where X and W are vector valued. For example, if X and W are bivariate, we need to prepare an orthonormal basis for $L_2(\mathbb{R}^2)$. Take some $j \in \mathbb{N}$. Using the basis $\{\phi_{j;k}, \varphi_{j';k}\}_{j' \geq j, k \in \mathbb{Z}}$ defined in the last subsection, a multi-resolution formula for $L_2(\mathbb{R}^2)$ is given by[4]

$$f(x_1, x_2) = \sum_{k,l \in \mathbb{Z}} c_{j;k,l}\phi_{j;k,l}(x_1, x_2) + \sum_{d=1}^{3}\sum_{j' \geq j}\sum_{k,l \in \mathbb{Z}} d^{(d)}_{j';k,l}\varphi^{(d)}_{j';k,l}(x_1, x_2),$$

[4]Indeed we employed such an expansion in (4) for f_{X^*W}.

where

$$\phi_{j;k,l}(x_1,x_2) = \phi_{j;k}(x_1)\phi_{j;l}(x_2), \qquad \varphi^{(1)}_{j;k,l}(x_1,x_2) = \phi_{j;k}(x_1)\psi_{j;l}(x_2),$$
$$\varphi^{(2)}_{j;k,l}(x_1,x_2) = \varphi_{j;k}(x_1)\phi_{j;l}(x_2), \qquad \varphi^{(3)}_{j;k,l}(x_1,x_2) = \varphi_{j;k}(x_1)\psi_{j;l}(x_2).$$

By using the linear wavelet space spanned by $\{\phi_{j_n;k,l}\}_{|k|,|l|\le L_n}$, we can also construct the series estimator of m in the bivariate case. An extension to the case of d-dimensional X and W follows in the same manner.

It is also possible to extend our estimation approach to the case where the model contains some additional exogenous explanatory variables Z, i.e.,

$$Y = m(X^*, Z) + U, \qquad E[U|W,Z] = 0.$$

In this case, similar to Horowitz (2011), we can modify the estimator $\hat{m}$ in (8) by introducing kernel weights. In particular, to estimate $m(\cdot,z)$ at a given z, we replace $\hat{a}_k$ in (5) and $\hat{c}_{k,l}$ in (7) with

$$\hat{a}_k(z) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{Z_i - z}{h_n}\right) Y_i\psi_k(W_i), \qquad \hat{c}_{k,l}(z) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{Z_i - z}{h_n}\right)\xi_k(X_i)\psi_l(W_i),$$

respectively, where K is a kernel function and h_n is a bandwidth. Then the estimator $\hat{m}(\cdot,z)$ is given by the same formula as in (8) for each z.

3. Asymptotic theory: Case of known f_ϵ

We now study the asymptotic properties of the estimator $\hat{m}$ in (8) using the linear wavelet space $\mathcal{L}^{(n)}_\phi$ spanned by $\{\phi_{j_n;k}\}_{|k|\le L_n}$. In this section, we consider the case where the density f_ϵ of the measurement error in X^* is known. The case of unknown f_ϵ will be studied in the next section.

Let us introduce some notation. Define the Sobolev space of order s as

$$\mathcal{S}(\mathbb{R}^d, s) = \begin{cases} \{h \in L_2(\mathbb{R}^d) : \left\|(1+|\cdot|^2)^{s/2}h^{\text{ft}}(\cdot)\right\|_2 < \infty\} & \text{for } s \in (-\infty,\infty) \\ \{h \in L_2(\mathbb{R}^d) : \int |h^{\text{ft}}(t)|^2 e^{c|t|^\gamma}dt < \infty \text{ for some } c,\gamma > 0\} & \text{for } s = \pm\infty \end{cases}$$

Let $f^{(n)}_{X^*W} = \sum_{|k|,|l|\le L_n} c_{k,l}\phi_{j_n;k}\phi_{j_n;l}$ and define the operator $A_n : L_2(\mathbb{R}) \to L_2(\mathbb{R})$ such that $(A_n g)(w) = \int g(x) f^{(n)}_{X^*W}(x,w)dx$. Also, denote the sieve measure of ill-posedness (Blundell, Chen and Kristensen, 2007, and Horowitz, 2012) to be

$$\rho_n = \sup_{h\in\mathcal{L}^{(n)}_\phi : h\neq 0} \frac{\|h\|_2}{\left\|(A^*A)^{1/2}h\right\|_2},$$

where A^* is the adjoint (or dual) operator of A defined in (2). Since A is assumed to be nonsingular, the object ρ_n is well-defined. Note that ρ_n depends on the smoothing parameter j_n (or equivalently $J_n = 4\pi 2^{j_n}/3$) through the sieve space $\mathcal{L}^{(n)}_\phi$. Let $\zeta_n = \left\{\min_{|t|\le J_n} |f^{\text{ft}}_\epsilon(t)|\right\}^{-1}$. Specifically, we note that ρ_n characterizes the degree of ill-posedness inherent in the integral equation (3), while ζ_n characterizes the same for the problem of deconvolution. We impose the following assumptions.

Assumption 1. *(i)* $f_{X^*W} \in \mathcal{S}(\mathbb{R}^2, s)$, $m \in \mathcal{S}(\mathbb{R}, s_1)$, *and* $r \in \mathcal{S}(\mathbb{R}, s+s_1)$ *for some* $s \in [2, \infty]$ *and* $s_1 \in [2, \infty)$. *(ii)* $\sup_{w\in\mathbb{R}} E[Y^2|W = w] \leq C < \infty$. *(iii)* $f_\epsilon^{\mathrm{ft}}(t) \neq 0$ *for all* $t \in \mathbb{R}$. *(iv) There exists some* $\eta > 0$ *such that* $\sup_{x,w\in\mathbb{R}} |x^2 + w^2|^{(1+\eta)/4}|f_{X^*W}(x, w)| < \infty$, $\sup_{x\in\mathbb{R}} |x|^{(1+\eta)/2}|m(x)| < \infty$, *and* $\sup_{x\in\mathbb{R}} |x|^{(1+\eta)/2}|r(x)| < \infty$.

Assumption 2. *(i) The operator* A *is nonsingular. (ii)* $\rho_n = O(J_n^s)$ *for* $s \in [2, \infty)$ *and* $\rho_n = O(e^{cJ_n^\gamma})$ *for* $s = \infty$. *(iii)* $\rho_n \sup_{\nu\in\mathcal{L}_\phi^{(n)}:\nu\neq 0} \frac{\|(A-A_n)\nu\|_2}{\|\nu\|_2} = O(J_n^{-s_1})$.

Assumption 1 gives a set of smoothness and boundedness conditions. In particular, Assumption 1 (i) requires that $r = Am$ should be much smoother than m. Assumption 1 (ii) is a mild condition on the conditional variance U given W that allows heteroskedasticity. Assumption 1 (iii) is a common requirement for deconvolution methods. Assumption 1 (iv) places some conditions on the decay rates of the functions $f_{X^*,W}$, m, and r in the tails (cf. Lemma 2). We also note that some of our assumptions on m are weaker than those in Horowitz (2011) which assumes compact support on data.

Assumption 2 collects conditions on the operator A in (2). Assumptions 2 (i) combined with Assumption 1 (iii) identifies the function m of interest from the model (1). Assumption 2 (ii) and (iii) are high level assumptions on the sieve measure of ill-posedness ρ_n that are commonly used in the literature (Blundell, Chen and Kristensen, 2007, and Horowitz, 2011).[5] For the case of $s \in [2, \infty)$, Assumption 2 (ii) is satisfied for wavelet series of order greater than s if we assume there exists some $c > 0$ such that $\|Ah\|_2 \geq c \left\|(1 + |\cdot|^2)^{-s/2}h^{\mathrm{ft}}(\cdot)\right\|_2$ for all $h \in L_2(\mathbb{R})$ (Blundell, Chen and Kristensen, 2007, Theorem 3). A sufficient condition for Assumption 2 (iii) is that A is a mapping from $\mathcal{S}(\mathbb{R}^2, s)$ to $\mathcal{S}(\mathbb{R}^2, s + s_1)$. Indeed, this follows by an application of Lemmas 1 and 2 after noting $\|(A - A_n)\nu\|_2 = \|(A - P_{j_n}A)\nu\|_2 + o(J_n^{-(s+s_1)})$ for some choice of L_n that is sufficiently large. Depending on the smoothness s of f_{X^*W} and associated sieve measure of ill-posedness ρ_n, we have two categories: (i) $s \in [2, \infty)$ and $\rho_n = O(J_n^s)$ (called mildly ill-posed case), and (ii) $s = \infty$ and $\rho_n = O(e^{cJ_n^\gamma})$ (called severely ill-posed case).

The following theorem establishes the convergence rates of our estimator $\hat{m}$ in (8) using the linear wavelets $\{\phi_{j_n;k}\}_{|k|\leq L_n}$ when the measurement error density f_ϵ is known.

Theorem 1. *Suppose that Assumptions 1 and 2 hold. Consider the estimator* $\hat{m}$ *in (8) using the linear wavelets* $\{\phi_{j_n;k}\}_{|k|\leq L_n}$ *of order greater than* s_1. *Furthermore, assume* $\rho_n\zeta_n(J_n/n)^{1/2} \to 0$ *and either (i)* $s \in [2, \infty)$ *and* $J_n^{2(s+s_1)+1}/L_n^\eta \to 0$ *or (ii)* $s = \infty$ *and* $n/L_n \to 0$ *as* $n \to \infty$. *Then*

$$\|\hat{m} - m\|_2 = O_p\left(J_n^{-s_1} + \rho_n\zeta_n(J_n/n)^{1/2}\right). \tag{10}$$

[5]Since we choose the order of wavelets $q \geq s_1$, it follows $\mathcal{L}_\phi^{(n)} \subseteq \mathcal{S}(\mathbb{R}, s_1)$. Thus Assumptions 2 (ii) and (iii) are equivalent to the corresponding assumption in Blundell, Chen and Kristensen (2007). Horowitz (2011) also makes a similar assumption but further restricts the space to satisfy $\|\nu\|_{\mathcal{S}(\mathbb{R},s_1)} = \left\|(1 + |\cdot|^2)^{s_1/2}\nu^{\mathrm{ft}}(\cdot)\right\|_2 < C_0$ for some constant $C_0 < \infty$. This is again equivalent to our definition: Indeed, since $\|\nu\|_{\mathcal{S}(\mathbb{R},s_1)} < \infty$ for any $\nu \in \mathcal{L}_\phi^{(n)}$, it follows

$$\sup_{\nu\in\mathcal{L}_\phi^{(n)}:\nu\neq 0} \frac{\left\|(\tilde{A} - \hat{A})\nu\right\|_2}{\|\nu\|_2} = \sup_{\nu\in\mathcal{L}_\phi^{(n)}:\nu\neq 0} \frac{\left\|(\tilde{A} - \hat{A})\nu/\|\nu\|_s\right\|_2}{\|\nu/\|\nu\|_s\|_2} = \sup_{\substack{\nu\in\mathcal{L}_\phi^{(n)}:\nu\neq 0,\\ \|\nu\|_{\mathcal{S}(\mathbb{R},s_1)}\leq 1}} \frac{\left\|(\tilde{A} - \hat{A})\nu\right\|_2}{\|\nu\|_2}.$$

A similar result also holds for assumption 2 (ii).

It should be noted that the L_2 convergence rate of the conventional series nonparametric regression estimator (i.e., X^* is exogenous and correctly measured) is $O_p\left(J_n^{-s_1} + (J_n/n)^{1/2}\right)$. Compared to this, it is clear that the additional component $\rho_n\zeta_n$ deteriorates the convergence rate of the estimator. Obviously the component ρ_n is associated with the ill-posedness of the integral equation in (3), and the component ζ_n is due to the existence of measurement error in X^*.

The condition on the order of the linear wavelets guarantees that its sieve space $\mathcal{L}_\phi^{(n)}$ is a subset of the Sobolev space $\mathcal{S}(\mathbb{R}, s_1)$. Note that the only requirement on the series length L_n is $J_n^{2(s+s_1)+1}/L_n^\eta \to 0$. This condition is fairly weak since we can let L_n grow arbitrarily fast. Only computational considerations prevent us from taking L_n too large.

Previous literature on nonparametric instrumental variable regression using the orthonormal series approach (e.g., Horowitz, 2011, 2012) has sometimes used the additional condition that the search space for $\hat{m}$ be restricted to a compact Sobolev ball $\{\nu : \|\nu\|_s \leq C\}$ for some $C < \infty$. Using a different method of proof than used previously, we are able to show that for the estimation method proposed here, this restriction may be dispensed with. Whether this relaxation can be extended to all possible orthonormal bases is however unclear since our proof is specific to wavelet series.

In order to characterize detailed properties of the convergence rate of $\hat{m}$, we consider some special cases categorized by the tail properties of the measurement error density f_ϵ.

3.1. **Ordinary smooth case.** Suppose that f_ϵ is ordinary smooth of order α, that is

$$C_1(1+|t|)^{-\alpha} \leq |f_\epsilon^{\text{ft}}(t)| \leq C_2(1+|t|)^{-\alpha} \quad \text{for all } t \in \mathbb{R}, \tag{11}$$

for some constants $C_2 > C_1 > 0$ and $\alpha > 1/2$. Typical examples of ordinary smooth densities are the Laplace and gamma densities. In this case, the component $\zeta_n = \left\{\min_{|t|\leq J_n} |f_\epsilon^{\text{ft}}(t)|\right\}^{-1}$ appearing in (10) is of order J_n^α.

For the mildly ill-posed case (i.e., $s \in [2, \infty)$ and $\rho_n = O(J_n^s)$), the convergence rate in (10) becomes $O_p\left(J_n^{-s_1} + (J_n^{2(\alpha+s)+1}/n)^{1/2}\right)$. Therefore, the optimal choice of J_n is given by $J_n = O(n^{1/(2(\alpha+s+s_1)+1)})$ and the optimal rate of convergence is

$$\|\hat{m} - m\|_2 = O_p(n^{-s_1/(2(\alpha+s+s_1)+1)}). \tag{12}$$

We can see that the presence of measurement error is equivalent to changing the smoothness of $f_{X^*,W}$ from s to $\alpha + s$. Intuitively, in the presence of measurement error, the degree of ill-posedness should be characterized not by the smoothness of $f_{X^*,W}$ for the unobservable X^*, but rather by that of $f_{X,W}$ for the observable X.

For the severely ill-posed case (i.e., $s = \infty$ and $\rho_n = O(e^{cJ_n^\gamma})$), the rate in (10) becomes $O_p\left(J_n^{-s_1} + e^{cJ_n^\gamma}(J_n^{2\alpha+1}/n)^{1/2}\right)$. The optimal choice of J_n is given by $J_n = (c_b \log n)^{1/\gamma}$ for some $c_b \in (0, 1/2c)$ and the optimal rate of convergence is

$$\|\hat{m} - m\|_2 = O_p((\log n)^{-s_1/\gamma}). \tag{13}$$

Therefore, in this case, the presence of measurement error no longer has any effect on the rate of convergence. The optimal choice of J_n is adaptive to unknown α and s_1.

3.2. **Supersmooth case.** Suppose now that f_ϵ is supersmooth, that is

$$C_1 \exp(-d_1|t|^\sigma) \leq |f_\epsilon^{\mathrm{ft}}(t)| \leq C_2 \exp(-d_2|t|^\sigma), \tag{14}$$

for some constants $C_2 > C_1 > 0$, $d_2 > d_1 > 0$, and $\sigma > 0$. Typical examples of supersmooth densities are the normal and Cauchy densities. In this case, the component ζ_n is of order $e^{d_1 J_n^\sigma}$.

For the mildly ill-posed case, the convergence rate in (10) becomes $O_p\left(J_n^{-s_1} + e^{d_1 J_n^\sigma}(J_n^{2s+1}/n)^{1/2}\right)$, and the optimal choice of J_n is $J_n = (c_b \log n)^{1/\sigma}$ for some $c_b \in (0, 1/2d_1)$, which yields the optimal convergence rate

$$\|\hat{m} - m\|_2 = O_p((\log n)^{-s_1/\sigma}). \tag{15}$$

In this case, the optimal choice of J_n is adaptive to unknown s and s_1.

For the severely ill-posed case, the rate in (10) becomes $O_p\left(J_n^{-s_1} + e^{cJ_n^\gamma + d_1 J_n^\sigma}(J_n^1/n)^{1/2}\right)$, and the optimal choice of J_n is $J_n = (c_b \log n)^{1/(\sigma\wedge\gamma)}$ for some $c_b \in (0, 1/(2d_1 + 2c))$, which yields the optimal convergence rate

$$\|\hat{m} - m\|_2 = O_p((\log n)^{-s_1/(\sigma\wedge\gamma)}). \tag{16}$$

Thus, measurement error affects the rate of convergence only if ζ_n (which may be taken as a measure of ill-posedness of the deconvolution problem) dominates ρ_n. In this case, the optimal choice of J_n is adaptive to unknown s_1.

4. Asymptotic theory: Case of unknown f_ϵ

The assumption of known measurement error density f_ϵ is unrealistic in many applications. Thus this section considers the situation where f_ϵ is unknown and needs to be estimated. In general, with single measurements, f_ϵ cannot be identified. Identification of f_ϵ can be restored however if we have two or more independent noisy measurements of the variable X^*. More specifically suppose that we observe

$$X_{i,j} = X_i^* + \epsilon_{i,j} \quad \text{for } j = 1, \ldots, N_i \text{ and } i = 1, \ldots, n,$$

where X_i^* is the 'true' observation and $\{\epsilon_{i,j}\}$ are independently distributed errors from the same error density f_ϵ. We thus have N_i repeated measurements of each variable X_i^*. We shall assume that the number of repeated observations is bounded above (i.e., $N_i \leq C < \infty$ for all i). The boundedness assumption is not critical for our theory but allows us to simplify the proofs considerably. Since in practice the number of repeated measurements is small anyway, we do not pursue the generalization to growing C.

We impose the following assumptions on f_ϵ.

Assumption 3. *(i) f_ϵ is symmetric around 0. (ii) There exist some $\delta \in (0, 1)$ and $M < \infty$ such that $P(|\epsilon| \geq L) \leq M(\log L)^{-\delta}$ for all $L > 0$.*

Assumption 3 (i) implies that f_ϵ^{ft} is real-valued for all $t \in \mathbb{R}$ and can be estimated as follows (Delaigle, Hall and Meister, 2008):

$$\hat{f}_\epsilon^{\text{ft}}(t) = \left| \frac{1}{N} \sum_{i=1}^{n} \sum_{j_1,j_2=1}^{N_i} \exp\{\mathrm{i}t(X_{i,j_1} - X_{i,j_2})\} \right|^{1/2},$$

where $N = \frac{1}{2}\sum_{i=1}^n N_i(N_i - 1)$ and we ignore all the observations with $N_i = 1$.[6] Assumption 3 (ii) is a mild condition on the tail decay of f_ϵ and is required for establishing uniform convergence of $\hat{f}_\epsilon^{\text{ft}}$ to f_ϵ^{ft} over some expanding interval. In particular, it is a much weaker condition than assuming bounded moments for ϵ.

Using the estimator $\hat{f}_\epsilon^{\text{ft}}$ in place of f_ϵ^{ft}, we can estimate the coefficients $c_{k,l}$ in (6) using the linear wavelet basis $\{\phi_{j_n;k}\}_{|k|\le L_n}$ as

$$\tilde{c}_{k,l} = \frac{1}{(2\pi)^2} \left\langle \frac{\hat{f}_{XW}^{\text{ft}}}{\hat{f}_\epsilon^{\text{ft}}}, \psi_k^{\text{ft}} \psi_l^{\text{ft}} \right\rangle, \tag{17}$$

where $\hat{f}_{XW}^{\text{ft}}(t,s) = \frac{1}{N}\sum_{i=1}^n \sum_{j=1}^{N_i} e^{\mathrm{i}(tX_{i,j}+sW_i)}$. Based on $\tilde{c}_{k,l}$, the estimator $\hat{m}$ in (8) is defined in the same manner as in the previous sections.

As in the last section, we consider both the cases where f_ϵ^{ft} is ordinary smooth and supersmooth. For the ordinary smooth case, we add the following assumption.

Assumption 4. f_ϵ^{ft} *satisfies (11) for some* $\alpha > 1/2$. *Also, assume* $s > \alpha$.

Note that in the context of (univariate) density deconvolution, Delaigle, Hall and Meister (2008) impose the condition $s > \alpha + 1/2$ in order to show that blind deconvolution with the optimal smoothing parameter achieves the minimax optimal rate of convergence under the pointwise mean-squared error criterion. Here we are able to relax the condition on s even further.[7] The following theorem shows that the estimation of f_ϵ does not change the convergence rate of the estimator $\hat{m}$ for the case of known f_ϵ.

Theorem 2. *Suppose that Assumptions 1-4 hold. Consider the estimator* $\hat{m}$ *in (8) using the linear wavelets* $\{\phi_{j_n;k}\}_{|k|\le L_n}$ *of order greater than* s_1 *and the estimated coefficients* $\tilde{c}_{k,l}$ *in (17). Then for* $s \in [2,\infty)$, *a choice of* J_n *satisfying* $J_n^{2(s+\alpha)+1}/n \to c < \infty$ *and* $J_n^{4\alpha}\log n/n \to 0$ *achieves the convergence rate in (12). Also, for* $s = \infty$, *a choice of* J_n *satisfying* $J_n = O(\log n)^\tau$ *for some* $\tau < \infty$ *achieves the convergence rate in (13).*

The condition $J_n^{2(s+\alpha)+1}/n \to c < \infty$ is not stringent since as shown in Section 3.1, the optimal choice of J_n for ordinary smooth f_ϵ with $s < \infty$ is $J_n = O(n^{1/(2(s+s_1+\alpha)+1)})$. Thus, Theorem 2

[6] For more general situations where f_ϵ could be asymmetric, we can still estimate the Fourier transform f_ϵ^{ft} via Kotlarski's identify as in Li and Vuong (1998). Li and Vuong's (1998) estimator takes a more complicated form and the asymptotic analysis using their estimator is beyond the scope of this paper.

[7] The intuition for this relaxation is as follows. For standard density deconvolution, to show that blind deconvolution does not affect the rate of convergence, we need $\int_{t_2\in\mathcal{C}_n}\int_{t_1}\left|\hat{f}_{XW}^{\text{ft}}\{\hat{\xi}^{-1/2} - \xi^{-1/2}\}\right|^2 dt_1 dt_2 = O_p(J_n^{1+2\alpha}/n)$. By contrast, in our case, as the proof of Theorem 2 makes clear, we only need to show $\int_{t_2\in\mathcal{C}_n}\left|\int_{t_1}\hat{f}_{XW}^{\text{ft}}\{\hat{\xi}^{-1/2} - \xi^{-1/2}\}\nu^{\text{ft}} dt_1\right|^2 dt_2 = O_p(J_n^{1+2\alpha}/n)$ for all $\nu \in \mathcal{L}_\phi^{(n)}$ satisfying $||\nu||_2 = 1$. The square integrability of ν essentially seems to provide an additional half a degree of 'smoothness' that allows us to relax the conditions further.

demonstrates that we can achieve the same rate of convergence using blind deconvolution under a few additional assumptions.

We now consider the case of supersmooth f_ϵ. Because of the slow rate of convergence, the variance term is dominated by the bias and consequently blind deconvolution does not affect the convergence rates. We formalize this in the next theorem.

Theorem 3. *Suppose that Assumptions 1-3 hold and f_ϵ satisfies (14). Consider the estimator $\hat{m}$ in (8) using the linear wavelets $\{\phi_{j_n;k}\}_{|k|\leq L_n}$ of order greater than s_1 and the estimated coefficients $\tilde{c}_{k,l}$ in (17). Then for $s \in [2,\infty)$, a choice of J_n satisfying $J_n = (c_b \log n)^{1/\sigma}$ for some $c_b \in (0, 1/4d_1)$ achieves the convergence rate in (15). Also, for $s = \infty$, a choice of J_n satisfying $J_n = (c_b \log n)^{1/(\sigma\wedge\gamma)}$ for some $c_b \in (0, 1/(2d_1 + 2(c \wedge d_1))$ achieves the convergence rate in (16).*

Compared to the results in Section 3.2, we note that Theorem 3 imposes more restrictions on the values J_n can take to achieve the same rate of convergence. For example, for the case of $s < \infty$, we require $J_n = (c_b \ln n)^{1/\sigma}$ with $c_b < 1/(4d_1)$ for blind deconvolution whereas we only need $c_b < 1/(2d_1)$ if f_ϵ were known. This distinction is of course not relevant if we are only interested in the rate of convergence, though it might matter in practical applications. Delaigle, Hall and Meister (2008) also impose the same restrictions on the values that c_b can take in order to arrive at the similar conclusion that blind deconvolution does not affect the rate of convergence if the error distribution is supersmooth.

Appendix A. Mathematical Appendix

Hereafter, "w.p.a.1" means "with probability approaching one". Also let $\|f\|_p = (\int |f(x)|^p dx)^{1/p}$ denote the L_p-norm of a function $f : \mathbb{R}^d \to \mathbb{C}$.

A.1. **Proof of Lemma 2.** The proof is a generalization of Pensky and Vidakovic (1999, Theorem 3). Denote $k = (k_1, \ldots, k_d)'$, $x = (x_1, \ldots, x_d)'$, and $\phi^{(d)}(x) = \prod_{h=1}^d \phi(x_h)$. Without loss of generality we may assume $\eta \le d$. Since we assume that ϕ is a father wavelet of order $q \ge (1+\eta)/2$, it follows $\sup_{x\in\mathbb{R}^d}\{|x|^{(1+\eta)/2}|\phi^{(d)}(z)|\} < \infty$. Thus we obtain

$$\begin{aligned} |k|^{(1+\eta)/2}|c_{j_n;k}| &\le 2^{j_n d/2}\int |(2^{j_n}x - k) - 2^{j_n}x|^{(1+\eta)/2}|\phi^{(d)}(2^{j_n}x-k)||f(x)|dx \\ &\le C_\eta 2^{j_n d/2}\sup_{x\in\mathbb{R}^d}\{|x|^{(1+\eta)/2}|\phi^{(d)}(x)|\}\int |f(x)|dx \\ &\quad + C_\eta 2^{j_n(1+\eta-d)/2}\sup_{x\in\mathbb{R}^d}\{|x|^{(1+\eta)/2}|f(x)|\}\,\|\phi\|_1^d \\ &\le C2^{j_n d/2}, \end{aligned}$$

where the second inequality follows from $|(2^m x - k) - 2^m x|^{(1+\eta)/2} \le C_\eta\{|2^m x - k|^{(1+\eta)/2} + |2^m x|^{(1+\eta)/2}\}$ for some constant $C_\eta > 0$ that depends only on η. Therefore, it holds

$$\sum_{h=1}^d \sum_{|k_h|\ge L_n} |c_{j_n;k}|^2 \le C^2 2^{j_n d}\sum_{h=1}^d \sum_{|k_h|\ge L_n} |k|^{-(1+\eta)} = O(2^{j_n d}/L_n^\eta),$$

where the equality follows from $|k|^{-1} \le d^{-1/2}\prod_{i=1}^d |k_i|^{-1/d}$.

A.2. **Proof of Theorem 1.** Since the proof is similar, we only show the statement for the mildly ill-posed case (i.e., $s \in [2,\infty)$ and $\rho_n = O(J_n^s)$). In this case, we show $\|\hat m - m\|_2 = O_p\left(J_n^{-s_1} + \zeta_n J_n^s (J_n/n)^{1/2}\right)$.

Define $m_n = \sum_{|k|\le L_n}\langle m, \phi_{j_n;k}\rangle \phi_{j_n;k}$ and $r_n = \sum_{|l|\le L_n}\langle r, \phi_{j_n;l}\rangle \phi_{j_n;l}$. At the end of this proof we shall show that

$$\sup_{\nu\in\mathcal{L}_\phi^{(n)}:\|\nu\|_2=1}\left\|(\hat A - A_n)\nu\right\|_2 = O_p\left(\zeta_n (J_n/n)^{1/2}\right). \tag{18}$$

Assumption 2 (ii) implies $\inf_{\nu\in\mathcal{L}_\phi^{(n)}:\|\nu\|_2=1}\|A\nu\|_2 \ge \rho_n^{-1}$ for all n large enough. Assumption 2 (iii) implies $\sup_{\nu\in\mathcal{L}_\phi^{(n)}:\|\nu\|_2=1}\|(A - A_n)\nu\|_2 = O_p(J_n^{-(s+s_1)})$. Thus, the condition $\rho_n = O(J_n^s)$ guarantees

$$\inf_{\nu\in\mathcal{L}_\phi^{(n)}:\|\nu\|_2=1}\|A_n\nu\|_2 \ge \rho_n^{-1}/2, \tag{19}$$

for all n large enough. Combining this with (18) and $\rho_n\zeta_n(J_n/n)^{1/2} \to 0$, we have

$$\inf_{\nu\in\mathcal{L}_\phi^{(n)}:\|\nu\|_2=1}\left\|\hat A\nu\right\|_2 \ge \rho_n^{-1}/4, \text{ w.p.a.1.} \tag{20}$$

Thus, the inverse operators $\hat A^{-1}$ and A_n^{-1} of $\hat A$ and A_n exist w.p.a.1 over the space $\mathcal{L}_\phi^{(n)}$, and this allows us to write $\hat m = \hat A^{-1}\hat r$ w.p.a.1.

By the triangle inequality,

$$\|\hat{m} - m_n\|_2 \leq \left\|A_n^{-1} r_n - m_n\right\|_2 + \left\|\hat{A}^{-1}\hat{r} - A_n^{-1} r_n\right\|_2 = \|T_1\|_2 + \|T_2\|_2 .$$

First, consider the term T_1. Note that

$$\begin{aligned} \|T_1\|_2 &\leq 2\rho_n \|r_n - A_n m_n\|_2 \\ &\leq 2\rho_n \left\{\|r_n - r\|_2 + \|A(m - m_n)\|_2 + \|(A - A_n) m_n\|_2\right\}, \end{aligned}$$

where the first inequality follows from (19) and the second inequality follows from the triangle inequality. Lemmas 1 and 2 with $r \in \mathcal{S}(\mathbb{R}, s + s_1)$ (Assumption 1 (i)), and the condition $J_n^{2(s+s_1)+1}/L_n^{\eta} \to 0$ imply

$$\|r - r_n\|_2 \leq \|r - P_{j_n} r\|_2 + \|P_{j_n} r - r_n\|_2 = O(J_n^{-(s+s_1)}).$$

Similarly, we have $\|m - m_n\|_2 = O(J_n^{-s_1})$ and by the Cauchy-Schwarz inequality,

$$\sup_{\|\nu\|_2 = 1} \|(A - A_n)\nu\|_2 \leq \left\|f_{X^*,W} - f^{(n)}_{X^*,W}\right\|_2 = O(J_n^{-s}).$$

Thus, we have

$$\begin{aligned} \|A(m - m_n)\|_2 &= \|(A - A_n)(m - m_n)\|_2 \\ &\leq \sup_{\|\nu\|_2 = 1} \|(A - A_n)\nu\|_2 \|m - m_n\|_2 = O(J_n^{-(s+s_1)}), \end{aligned}$$

where the first equality follows from $A_n(m - m_n) = 0$ (because $m - m_n$ belongs to the orthogonal space of $\mathcal{L}^{(n)}_{\phi}$). Also, Assumptions 2 (ii) and (iii) guarantee $\|(A - A_n) m_n\|_2 = O(J_n^{-(s+s_1)})$. Combining these results, we obtain $\|T_1\|_2 = O(J_n^{-s_1})$.

Next, consider the term T_2. By the triangle inequality and (20),

$$\begin{aligned} \|T_2\|_2 &\leq \left\|\hat{A}^{-1}(\hat{r} - r_n)\right\|_2 + \left\|\hat{A}^{-1}(A_n - \hat{A}) A_n^{-1} r_n\right\|_2 \\ &\leq 4\rho_n \left\{\|\hat{r} - r_n\|_2 + \left\|(A_n - \hat{A})(A_n^{-1} r_n)\right\|_2\right\}, \end{aligned}$$

w.p.a.1. By analogous arguments as in the proof of Meister (2009, Proposition 2.4) along with Assumption 1(ii), it follows $\|\hat{r} - r_n\|_2 = O_p((J_n/n)^{1/2})$. Also, by earlier arguments, $\left\|A_n^{-1} r_n\right\|_2 \leq \|m_n\|_2 + \|T_1\|_2 = O_p(1)$. Combining this with (18) and $\rho_n = O(J_n^s)$ (Assumption 2 (ii)), we have $\|T_2\|_2 = O_p\left(\zeta_n J_n (J_n/n)^{1/2}\right)$. Therefore, the conclusion follows.

It remains to show (18). Pick any $\nu \in \mathcal{L}^{(n)}_{\phi}$ satisfying $\|\nu\|_2 = 1$. Note that $\nu \in \mathcal{L}^{(n)}_{\phi}$ is expanded as $\nu = \sum_{|k| \leq L_n} \nu_k \phi_{j_n;k}$ with $\nu_k = \langle \nu, \phi_{j_n;k} \rangle$ and we can write

$$\left\|(\hat{A} - A_n)\nu\right\|_2^2 = \sum_{|l| \leq L_n} \left| \sum_{|k| \leq L_n} (\hat{c}_{k,l} - c_{k,l}) \nu_k \right|^2 .$$

Define $\hat{f}^{\rm ft}_{XW}(t_1,t_2) = \frac{1}{n}\sum_{i=1}^n e^{\mathrm{i}(t_1X_i+t_2W_i)}$. Then by the definition of $\hat{c}_{k,l}$,

$$\begin{aligned}\sum_{|k|\le L_n}(\hat{c}_{k,l}-c_{k,l})\nu_k &= \sum_{|k|\le L_n}\frac{1}{(2\pi)^2}\int\int\frac{\hat{f}^{\rm ft}_{XW}(t_1,t_2)-f^{\rm ft}_{XW}(t_1,t_2)}{f^{\rm ft}_\epsilon(t_1)}\overline{\phi^{\rm ft}_{j_n;k}(t_1)\phi^{\rm ft}_{j_n;l}(t_2)}\nu_k dt_1dt_2\\ &= \frac{1}{(2\pi)^2}\left\langle\int\frac{\hat{f}^{\rm ft}_{XW}(t_1,\cdot)-f^{\rm ft}_{XW}(t_1,\cdot)}{f^{\rm ft}_\epsilon(t_1)}\overline{\tilde{\nu}^{\rm ft}(t_1)}dt_1,\phi^{\rm ft}_{j_n;l}(\cdot)\right\rangle,\end{aligned}$$

where $\tilde{\nu}=\sum_{|k|\le L_n}\overline{\nu_k}\phi_{j_n;k}$. Using the facts $\phi^{\rm ft}_{j_n;l}$ is compactly supported on $\mathcal{C}_n\equiv[-J_n,J_n]$ for all $l\in\mathbb{Z}$, and the orthogonality of the wavelet series, it follows $\sum_{l\in\mathbb{Z}}\left\langle h,\phi^{\rm ft}_{j_n;l}\right\rangle^2=\sum_{l\in\mathbb{Z}}\left\langle h\mathbb{I}\{\cdot\in\mathcal{C}_n\},\phi^{\rm ft}_{j_n;l}\right\rangle^2\le(2\pi)\left\|h\mathbb{I}\{\cdot\in\mathcal{C}_n\}\right\|_2^2$, where the inequality follows from the fact that we left out the nonlinear part $\{\varphi_{j';k}\}_{j'\ge j,k\in\mathbb{Z}}$ of the wavelet basis. Thus, we have

$$\sum_{|l|\le L_n}\left|\sum_{|k|\le L_n}(\hat{c}_{k,l}-c_{k,l})\nu_k\right|^2\le\frac{1}{(2\pi)^3}\int_{t_2\in\mathcal{C}_n}\left|\int\frac{\hat{f}^{\rm ft}_{XW}(t_1,t_2)-f^{\rm ft}_{XW}(t_1,t_2)}{f^{\rm ft}_\epsilon(t_1)}\overline{\tilde{\nu}^{\rm ft}(t_1)}dt_1\right|^2dt_2.$$

By taking expectation,

$$E\left\|(\hat{A}-A_n)\nu\right\|_2^2\le E\left[\frac{2}{(2\pi)^3n}\int_{t_2\in\mathcal{C}_n}\left|\int e^{\mathrm{i}(t_1X+t_2W)}\frac{\overline{\tilde{\nu}^{\rm ft}(t_1)}}{f^{\rm ft}_\epsilon(t_1)}dt_1\right|^2dt_2\right]=O(\zeta_n^2J_n/n),$$

where the equality follows from $\|\tilde{\nu}\|_2=\|\nu\|_2=1$, and the fact that $\tilde{\nu}^{\rm ft}$ is compactly supported on $\mathcal{C}_n$ since $\tilde{\nu}\in\mathcal{L}^{(n)}_\phi$. This proves (18).

A.3. **Proof of Theorem 2.** For simplicity we restrict attention to the case $N_i=2$. For the more general situation where N_i is arbitrary but bounded above by C, the proof follows by similar arguments after accounting for the dependence structure in $\hat{f}^{\rm ft}_\epsilon(t)$. As before, we prove the case of $s\in[2,\infty)$. The proof for the case of $s=\infty$ follows along similar lines but is more straightforward. Recall the notation from the proof of Theorem 1. In addition, define $\xi(t)=(f^{\rm ft}_\epsilon(t))^2$ and $\hat{\xi}(t)=(\hat{f}^{\rm ft}_\epsilon(t))^2$. The claim follows from the proof of Theorem 1 if we show that for all $\nu\in\mathcal{L}^{(n)}_\phi$ satisfying $\|\nu\|_2=1$,

$$\sum_{|l|\le L_n}\left|\sum_{|k|\le L_n}(\tilde{c}_{k,l}-\hat{c}_{k,l})\nu_k\right|^2=O_p(J_n^{1+2\alpha}/n).$$

Now by similar arguments as in the proof of Theorem 1, it follows

$$\sum_{|l|\le L_n}\left|\sum_{|k|\le L_n}(\tilde{c}_{k,l}-\hat{c}_{k,l})\nu_k\right|^2\le(2\pi)^{-3}\int_{t_2\in\mathcal{C}_n}\left|\int h_n(t_1,t_2)\overline{\tilde{\nu}^{\rm ft}(t_1)}dt_1\right|^2dt_2,$$

where

$$h_n(t_1,t_2)=\hat{f}^{\rm ft}_{XW}(t_1,t_2)\{\hat{\xi}^{-1/2}(t_1)-\xi^{-1/2}(t_1)\}\mathbb{I}\{t_1,t_2\in\mathcal{C}_n\}.$$

Now, denoting $I_n=\int_{t_2\in\mathcal{C}_n}\left|\int h_n(t_1,t_2)\overline{\tilde{\nu}^{\rm ft}(t_1)}dt_1\right|^2dt_2$, the claim follows if we show that $I_n=O_p(J_n^{1+2\alpha}/n)$.

Notation: For the remainder of the proof we shall the drop the functional arguments t_1, t_2. These are to be inferred from the definition of the function and the context.

We now prove the following equalities which are used later in the proof:

$$\int_{\mathcal{C}_n} |\hat{\xi}^{1/2} - \xi^{1/2}|^{2q} = O(J_n^{1+2\alpha q}/n^q) \text{ for all } q \geq 1, \tag{21}$$

$$\sup_{|t|\leq J_n} \left| \frac{\xi(t)}{\hat{\xi}(t)} \right| = 1 + o_p(1) \tag{22}$$

First, we show (21). Expanding the expectations yields that $E\left[\int_{\mathcal{C}_n} |\hat{\xi} - \xi|^{2q}\right] = O(J_n/n^q)$. Thus, the elementary algebraic inequality $|\hat{\xi}^{1/2} - \xi^{1/2}| \leq \xi^{-1/2}|\hat{\xi} - \xi|$ and Assumption 4 imply

$$E\left[\int_{\mathcal{C}_n} |\hat{\xi}^{1/2} - \xi^{1/2}|^{2q}\right] \leq \left(\min_{t\in\mathcal{C}_n} |f_\epsilon^{\mathrm{ft}}(t)|\right)^{-2q} E\left[\int_{\mathcal{C}_n} |\hat{\xi} - \xi|^{2q}\right] = O(J_n^{1+2\alpha q}/n^q), \tag{23}$$

and the claim in (21) follows. Next, to show equation (22), we use Theorem 6.3 of Yukich (1987) which assures that for $J_n = O(n^{-c})$ for some $c > 0$ and under Assumption 3 (ii), $\sup_{|t|\leq J_n} \left|\hat{\xi}(t) - \xi(t)\right| = O_p(\sqrt{\log n/n})$. Combined with the rate condition $J_n^{4\alpha} \log n/n \to 0$, this ensures $\left(\min_{|t|\leq J_n} \left|\hat{\xi}(t)\right|\right)^{-1} = O_p(J_n^{2\alpha})$. Hence we obtain

$$\sup_{|t|\leq J_n} \left|\frac{\xi(t)}{\hat{\xi}(t)}\right| \leq 1 + \sup_{|t|\leq J_n} \left|\frac{\hat{\xi}(t) - \xi(t)}{\hat{\xi}(t)}\right| = 1 + O_p\left(\left(\frac{J_n^{4\alpha}\log n}{n}\right)^{1/2}\right) = 1 + o_p(1).$$

This proves the claim in (22).

We now show $I_n = O_p(J_n^{1+2\alpha}/n)$. Recalling $f_{XW}^{\mathrm{ft}} = f_{X^*W}^{\mathrm{ft}} f_\epsilon^{\mathrm{ft}}$, we may write

$$\begin{aligned}
I_n &\leq 2\int_{\mathcal{C}_n} \left|\int_{\mathcal{C}_n} (\hat{f}_{XW}^{\mathrm{ft}} - f_{XW}^{\mathrm{ft}})(\hat{\xi}^{-1/2} - \xi^{-1/2})\overline{\tilde{\nu}^{\mathrm{ft}}}\right|^2 + 2\int_{\mathcal{C}_n} \left|\int_{\mathcal{C}_n} \frac{f_{X^*W}^{\mathrm{ft}}}{\hat{\xi}^{1/2}}(\hat{\xi}^{1/2} - \xi^{1/2})\overline{\tilde{\nu}^{\mathrm{ft}}}\right|^2, \\
&\leq 2\int_{\mathcal{C}_n\times\mathcal{C}_n} \frac{|\hat{f}_{XW}^{\mathrm{ft}} - f_{XW}^{\mathrm{ft}}|^2}{\xi\hat{\xi}} |\hat{\xi}^{1/2} - \xi^{1/2}|^2 + 2\int_{\mathcal{C}_n\times\mathcal{C}_n} \left|\frac{f_{X^*W}^{\mathrm{ft}}}{\xi^{1/2}}\right|^2 \left|(\hat{\xi}^{1/2} - \xi^{1/2})\tilde{\nu}^{\mathrm{ft}}\right|^2 \frac{\xi}{\hat{\xi}} \\
&= I_{1n} + I_{2n}
\end{aligned}$$

where we used the Cauchy-Schwarz inequality and the fact that $\|\tilde{\nu}\|_2 = \|\nu\|_2 = 1$ to derive the second inequality. By Assumption 1 (i) and Young's inequality, we obtain

$$|f_{X^*W}^{\mathrm{ft}}| \leq C\{1 + (t_1^2 + t_2^2)\}^{-s/2} \leq C_1(1 \wedge |t_2|^\delta)^{-1}(1 \wedge |t_1|^{s-\delta})^{-1},$$

for some $C, C_1 > 0$ and δ satisfying $0 < \delta < s - \alpha$. Consequently by Assumption 3, $\sup_{t_1,t_2\in\mathbb{R}} \left|\frac{f_{X^*W}^{\mathrm{ft}}}{\xi^{1/2}}\right| = O(1)$. Furthermore, by analogous arguments as used to derive equation (21), it follows $\int_{\mathcal{C}_n} \left|(\hat{\xi}^{1/2} - \xi^{1/2})\tilde{\nu}^{\mathrm{ft}}\right|^2 = O_p(J_n^{2\alpha}/n)$. So using (22) assures $I_{2n} = O_p(J_n^{1+2\alpha}/n)$. Next consider the term I_{1n}, which can be written as

$$I_{1n} = 2\int_{\mathcal{C}_n\times\mathcal{C}_n} \left|\frac{f_{X^*W}^{\mathrm{ft}}}{\xi^{1/2}}\left(\frac{\hat{f}_{XW}^{\mathrm{ft}} - f_{XW}^{\mathrm{ft}}}{f_{XW}^{\mathrm{ft}}}\right)\right|^2 |\hat{\xi}^{1/2} - \xi^{1/2}|^2 \frac{\xi}{\hat{\xi}}$$

It follows after expanding the expectations that

$$E\left[\int_{\mathcal{C}_n\times\mathcal{C}_n} |\hat{f}_{XW}^{\mathrm{ft}} - f_{XW}^{\mathrm{ft}}|^4\right] = O(J_n^2/n^2).$$

Consequently we obtain

$$\int_{\mathcal{C}_n \times \mathcal{C}_n} \left| \frac{f^{\text{ft}}_{X^*W}}{\xi^{1/2}} \left(\frac{\hat{f}^{\text{ft}}_{XW} - f^{\text{ft}}_{XW}}{f^{\text{ft}}_{XW}} \right) \right|^4 = O_p\left(\frac{J_n^{4(s+\alpha)+2}}{n^2} \right) = O_p(1).$$

Combining this with (21), (22), and Assumption 4, Cauchy-Schwarz inequality implies $I_{1n} = O_p(J_n^{1+2\alpha}/n)$. Therefore, the conclusion follows.

A.4. **Proof of Theorem 3.** The proof is similar to that of Theorem 2 with a few changes. Instead of the bound in (21), we use the following

$$\int_{\mathcal{C}_n} \left| \frac{\hat{\xi}^{1/2} - \xi^{1/2}}{\xi^{1/2}} \right|^4 = O_p\left(\frac{J_n \exp\{4d_1 J_n^{\sigma}\}}{n} \right). \tag{24}$$

The result (24) follows by a similar argument for (21) after applying the elementary inequality $|\hat{\xi}^{1/2} - \xi^{1/2}|^2 \leq 2|\hat{\xi} - \xi|$. Next we note that the result in (22) is also applicable here by a similar reasoning under the rate condition $J_n = O(\log n)$. Furthermore, by proceeding as in the proof of Theorem 2, we obtain

$$E\left[\int_{\mathcal{C}_n \times \mathcal{C}_n} \left| \frac{\hat{f}^{\text{ft}}_{XW} - f^{\text{ft}}_{XW}}{\xi^{1/2}} \right|^4 \right] = O\left(\frac{J_n^2 \exp\{4d_1 J_n^{\sigma}\}}{n^2} \right). \tag{25}$$

So by Cauchy-Schwarz inequality and the fact $\|\tilde{\nu}\|_2 = 1$, the term I_n may be expanded as

$$I_n \leq 2\int_{\mathcal{C}_n \times \mathcal{C}_n} \left| \frac{\hat{f}^{\text{ft}}_{XW} - f^{\text{ft}}_{XW}}{\xi^{1/2}} \right|^2 \left| \frac{\hat{\xi}^{1/2} - \xi^{1/2}}{\xi^{1/2}} \right|^2 \frac{\xi}{\hat{\xi}} + 2\int_{\mathcal{C}_n \times \mathcal{C}_n} |f^{\text{ft}}_{X^*W}|^2 \left| \frac{\hat{\xi}^{1/2} - \xi^{1/2}}{\xi^{1/2}} \right|^2 \frac{\xi}{\hat{\xi}}.$$

From (24), (22) and (25), straightforward algebra enables us to show $I_n = O_p(n^{-\epsilon})$ for some $\epsilon > 0$. However, this term is clearly dominated by the bias term of order $J_n^{-s_1}$. Therefore, the conclusion follows.

References

[1] Blundell, R., Chen, X. and D. Kristensen (2007) Semi-nonparametric IV estimation of shape-invariant Engel curve, *Econometrica*, 75, 1613-1669.

[2] Darolles, S., Fan, Y., Florens, J. P. and E. Renault (2011) Nonparametric instrumental regression, *Econometrica*, 79, 1541-1565.

[3] 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.

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

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

[6] Fan, J. and J.-Y. Koo (2002) Wavelet deconvolution, *IEEE Transactions on Information Theory*, 48, 734-747.

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

[8] Gagliardini, P. and O. Scaillet (2012) Nonparametric instrumental variable estimation of structural quantile effect, *Econometrica*, 80, 1533-1562.

[9] Hall, P. and J. L. Horowitz (2005) Nonparametric methods for inference in the presence of instrumental variables, *Annals of Statistics*, 33, 2904-2929.

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

[11] Horowitz, J. L. (2011) Applied nonparametric instrumental variables estimation, *Econometrica*, 79, 347-394.

[12] Horowitz, J. L. (2012) Specification testing in nonparametric instrumental variable estimation, *Journal of Econometrics*, 167, 383-396.

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

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

[15] Newey, W. K. and Powell, J. L. (2003) Instrumental variable estimation of nonparametric model, *Econometrica*, 71, 1565-1578.

[16] Pensky, M. and B. Vidakovic (1999) Adaptive wavelet estimator for nonparametric density deconvolution, *Annals of Statistics*, 27, 2033-2053.

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

[18] Yukich, J. (1987) Some limit theorems for the empirical process indexed by functions, *Probability Theory and Related Fields*, 74, 71-90.

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

E-mail address: k.adusumilli@lse.ac.uk

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

E-mail address: t.otsu@lse.ac.uk