IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 28, 2019, accepted July 15, 2019, date of publication July 17, 2019, date of current version August 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2929571

Semiparametric Wavelet-Based JPEG IV Estimator for Endogenously Truncated Data

NIR BILLFELD AND MOSHE KIM
Department of Economics, University of Haifa, Haifa 3498838, Israel
Corresponding author: Moshe Kim (kim@econ.haifa.ac.il)

This work was supported by the Research Authority, the University of Haifa.

ABSTRACT A new and an enriched JPEG algorithm is provided for identifying redundancies in a sequence of irregular noisy data points which also accommodates a reference-free criterion function. Our main contribution is by formulating analytically (instead of approximating) the inverse of the transpose of JPEG-wavelet transform without involving matrices which are computationally cumbersome. The algorithm is suitable for the widely-spread situations where the original data distribution is unobservable such as in cases where there is deficient representation of the entire population in the training data (in machine learning) and thus the covariate shift assumption is violated. The proposed estimator corrects for both biases, the one generated by endogenous truncation and the one generated by endogenous covariates. Results from utilizing 2 000 000 different distribution functions verify the applicability and high accuracy of our procedure to cases in which the disturbances are neither jointly nor marginally normally distributed.

INDEX TERMS JPEG, semiparametric, biorthogonal wavelet, causality, proximal gradient-descent, lifting scheme, denoising, covariate shift, training data, reference-free.

I. INTRODUCTION

Scientists routinely try to model and extract causal relations among covariates, rather than merely their correlations.[1] In practice however, the presence of endogenous covariates in the model challenges the causal inference due to comovement of the random disturbance with these covariates. We distinguish between population induced comovement and training data (as in machine learning) comovement without having to rely on a covariate shift assumption since the behavioral (causal) model embedded in the training data does not necessarily describing the behavior in the entire population (see discussion in [2]).

The common way to overcome the aforementioned, is to generate a variation in the endogenous covariate without introducing variation in the random disturbance. This idea is achieved by employing a proper instrumental variable (IV).[2]

Application of a proper instrumental variable generates variation in the endogenous covariate without introducing variation in the random disturbance and hence is orthogonal to it. Thus, IV should contribute to exogeneity and therefore has been extremely popular in empirical work.

Once we have analytically shown that the IV estimator is no longer valid in an *endogenously truncated* environment, we offer a truncation-proof estimator, which is a semiparametric wavelet-based JPEG-IV denoising algorithm.[3] This denoising algorithm decomposes the random disturbance into a noise and a systemic bias part, enabling the elimination of the truncation bias. The magnitude of the this bias is captured by the size of the wavelet coefficients which quantify and measure the degree of redundancy hidden in a sequence of data points. Consequently, this algorithm nests the conventional IV estimator as a special case due to the fact that in the absence of systemic endogenous truncation, the wavelet coefficients approach zero except for the intercept which describe the coarse level of the function.[4]

Our main contribution to the biorthogonal wavelet estimator is by formulating analytically (instead of approxi-

The associate editor coordinating the review of this manuscript and approving it for publication was Ramakrishnan Srinivasan.

[1]For a specific form of causality due to treatment effect see [1] definition.

[2]Note that we deal with endogenously *truncated* sample selection model to differentiate from *censored* sample selection models [3]–[5], where there exists information pertaining to the non-participants.

[3]A wavelet is a bandwidth-free estimator that is based on a multi-scale representation of the data. It is a widely used denoising technique [6].

[4]Unlike Fourier transform, the wavelet estimator preservers not only the data average (coarse) behavior but also its local behavior capturing deviations (details) from the average. This fact renders our denoising suitable for irregular-spaced data which largely depend on local behavior and play an important role in the denoising.

mating) the inverse of the transpose of wavelet transform without involving matrices which are computationally cumbersome [7][5] as well as the management of irregular-spaced data.[6] Additionally, our proposed methodology enables the combination of several penalty functions in the estimation procedure which are resolution-dependent.[7]

Wavelets are useful in denoising data. Several image quality assessment (IQA) measures have been introduced to choose the optimal level of denoising. These approaches can be classified to full-reference (FR) in cases the original image (noise-free) is observed; reduced-reference (RR) in cases where there is a partial information about the reference; reference-free (RF) in cases where the original image is not accessible [10], [11]. As we deal with truncated distributions, the source (the complete non-truncated distribution) is intrinsically unobservable and thus, we cannot assess the success of the denoising by comparing it to the original non-truncated distribution. Therefore, we select both the thresholding (tuning) parameter as well as the penalty function using a reference-free criterion function.

The proposed JPEG IV is biorthogonal, thus preserving both the symmetry (the original shape of the data) and compact support (small number of coefficients) properties of the data.[8] Importantly, the proposed methodology is easy to compute by precluding the need to find an optimal bandwidth as conventionally done.[9] These properties make it suitable for denoising by alleviating both the problem of coefficient expansion as well as border discontinuities [15]. The proposed algorithm corrects for both sources of bias: the endogeneity of covariates as well as the endogenous self-selection biases.

We run Monte Carlo simulations to measure the magnitude of the potential bias in the parameters' estimates under endogenous truncation, obtained by employing a conventional IV to eliminate the endogeneity bias. Our empirical implementation shows that even under mild correlation between the random disturbances, the resulting bias in the estimated parameter of the endogenous covariate in the substantive equation can amount to almost tenfold the true parameter value. Further, for sake of generality of the offered estimator, we subject it to various distributions in which the disturbances are neither jointly nor marginally normally distributed. These disturbances are constructed as realizations of non-symmetric and non-unimodal distribution functions.[10]

The rest of this paper is organized as follows. The methodology is presented is section II. Section III prepares the ground for the biorthogonal wavelet. Section IV presents our proposed JPEG algorithm. In section V we employ Monte Carlo simulations to validate our estimator performance. Section VI concludes.

II. METHODOLOGY

As discussed above, the IV is based on the following basic requirements: it is correlated with the endogenous covariate, as well as orthogonal to the random disturbance. Additionally, it must satisfy the exclusion restriction, such that in the presence of the endogenous covariate, the IV must be excluded from the regression. The IV is allowed to affect the dependent variable only through its effect on the endogenous covariate. However, the orthogonality condition is rarely satisfied in the presence of endogenous truncation, which is very frequently the nature of data used in empirical research, and therefore the IV will not provide a solution for the endogeneity problem. In what follows, we demonstrate the shortcoming of the conventional IV estimator, as well as potential bias generated in an environment of endogenously truncated data.

Suppose that there is a population random variable $\omega = (\mathrm{z}; \mathrm{x}_1, \mathbf{x_{-1}}; \mathbf{w})$ and that there is an independent and identically distributed sample $\left\{z_i, x_{1i}, \boldsymbol{x_{-1_i}}, \boldsymbol{w_i}\right\}_{i=1}^{N}$ drawn from this population, referred to as the complete data set consisting of N observations.[11] The instrumental variable is z, the endogenous variable is x_1 and the exogenous random variables are $(\mathbf{x_{-1}}\boldsymbol{\mathcal{I}}\,\mathbf{w})$, and where $\mathbf{w} \in \mathbb{R}^l$ is a covariate vector.

Let ξ_{1i}, ξ_{2i} and v_i be jointly dependent random disturbances with the respective marginal distribution functions F_{ξ_1}, F_{ξ_2} and F_{v}. Their joint distribution function is $F_{\xi_1,\xi_2,\mathrm{v}}$. The model is semiparametric, as neither the marginals nor the joint distribution function are required to be specified by the researcher.

The underlying model is composed of two parts. The first part consists of a selection equation, while the second part consists of the substantive (of interest) equation.

The population (non-truncated) selection equation is defined as:

$$y_{2i}^{*} = \boldsymbol{w}_i^T \boldsymbol{\gamma} + \xi_{2i} \tag{1}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^l$ and $\boldsymbol{w_i} \in \mathbb{R}^l$ are the selection equation's coefficients and covariates vector, respectively. The selection equation's random disturbance is denoted by ξ_{2i}.

[5] "The implemented routine for the inverse transpose transform is approximate." [7], p.285.

[6] In the orthogonal wavelets design various interpolation methods are used to alleviate these irregularities [8] and specific methodologies can be used to extend the Haar wavelet transform to the unequally spaced case [9].

[7] It is known that soft thresholding provides smoother results relative to the hard thresholding because it is continuous. The latter, however, provides better edge preservation in comparison with the former.

[8] Our JPEG IV is a biorthogonal wavelet as it requires two sets of vectors, which are the *dual basis* and the *series expansion* sets, to obtain a denoised representation of the data. The elements in the former set are orthogonal to the corresponding elements in the latter set. See [12] for a formal definition of biorthogonality.

[9] Kernel estimation involves computational burden due to the necessity of finding the optimal bandwidth [13]. Unlike the nonparametric case, in the semiparametric context there is no "protocol" for finding the optimal bandwidth, as the traditional bandwidth choice methods might lead to bias estimates due to improper bandwidth choice [14].

[10] Unlike the practice in some other studies applying only normally distributed disturbances.

[11] Capital letters indicate random variables; lower case letters indicate realizations of these random variables.

The substantive equation and the endogenous variable equation are defined as a system of equations:

$$\begin{bmatrix} y_{1i}^* \\ x_{1i}^* \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_i^T \\ [\boldsymbol{z}_i^T, \boldsymbol{x}_{-1_i}^T] \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\delta} \end{bmatrix} + \begin{bmatrix} \xi_{1i} \\ \mathrm{v}_i \end{bmatrix} \quad \begin{matrix} \text{the substantive equation} \\ \text{the endogenous variable equation} \end{matrix} \tag{2}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{p_1}$ and $\boldsymbol{\delta} \in \mathbb{R}^{p_2}$ are covariates vectors, x_{1i} is an endogenous variable included in vector $\boldsymbol{x}_i \in \mathbb{R}^{p_1}$, and the exogenous variables are denoted by $\boldsymbol{x}_{-1_i}^T$. The substantive equation's random disturbances are denoted by ξ_{1i} and v_{1i}.

However the variables $y_{1i}^*, y_{2i}^*, x_{1i}^*$ are latent in the truncated environment and their respective observed realizations are denoted by y_{1i}, y_{2i}, x_{1i}, defined in (3) and (4) to follow.

The variable y_{2i}^* is latent, while y_{2i} is observed and defined as:

$$y_{2i} = \begin{cases} 1 & \text{if } y_{2i}^* \geq 0 \\ \text{Unobserved} & \text{if } y_{2i}^* < 0, \end{cases} \quad \text{the selection equation} \tag{3}$$

$$\begin{bmatrix} y_{1i} \\ x_{1i} \end{bmatrix} = \begin{cases} \begin{bmatrix} y_{1i}^* \\ x_{1i}^* \end{bmatrix} & \text{if } y_{2i}^* \geq 0 \\ \text{Unobserved} & \text{if } y_{2i}^* < 0, \end{cases} \quad \text{the substantive equations} \tag{4}$$

In the next section we reformulate the substantive equation as a partially linear single index model.

A. SEMIPARAMETRIC SELECTIVITY BIAS CORRECTION

The key difference between censored and truncated sample selection models is that in the former the entire covariate set (including the non-participants) and the selection variable are fully observed. In the latter, the entire data are truncated. Nevertheless, in both cases, the substantive equation can be represented as a partially linear regression, in which the dependent variable is observed only for the participants, as we are about to show. Following [16], the conditional expectation of the substantive equation in semiparametric (censored)[12] sample selection models is some generally unknown function $\mathcal{M}_1(.)$ (to be estimated) of the selection equation's covariates variables $\boldsymbol{w_i}$:

$$\mathbb{E}\left[\xi_{1i} | \xi_{2i} > -\boldsymbol{w}_i^T \boldsymbol{\gamma}\right] = \mathcal{M}_1(\boldsymbol{w}_i^T \boldsymbol{\gamma}) \tag{5}$$

such that $\boldsymbol{\gamma}$ is the selection equation's coefficient vector. Since y_{1i} is observed only if i is a participant, the substantive equation's dependent variable obtains the following functional form:

$$y_{1i} = \boldsymbol{x}_i^T \boldsymbol{\beta} + \underbrace{\mathcal{M}_1(\boldsymbol{w}_i^T \boldsymbol{\gamma})}_{\text{the bias term}} + \underbrace{\tilde{\epsilon}_{1i}}_{\text{white noise}} \tag{6}$$

The regression equation in (6) is referred to as a semiparametric partially linear regression (SP-NLS), in which the non-linear part is the bias term function. This regression can be estimated semiparametrically in cases of a truncated sample selection model using a non-linear least squares procedure as suggested by [13].

Both [13] and [16] models involve a kernel function estimation. However, kernel estimates' accuracy is sensitive to the bandwidth selected. This entails a potential problem of finding the optimal bandwidth resulting in computational complexity.[13] Due to the lack of applicability of the traditional bandwidth selection methods in the semiparametric context, informal methods are being used, that may lead to a non-ignorable bias in the estimates [18].[14] In order to avoid the problems involved with kernel estimation, our methodology relies on a (thresholding-propagated) nonlinear wavelet-based JPEG IV estimator to approximate the bias term (in (6)).

The substantive equation depicted in (6) deals with endogenous truncation bias, assuming that the random disturbance and the covariates are not jointly dependent. However, in cases where this random disturbance is jointly dependent with one (or more) of the covariates there will emerge two bias terms: the first one is propagated by the endogenous truncation and the second one is propagated by the endogenous covariate. Next we present a decomposition Theorem 1, which enables reformulating the substantive equations as a partially linear single index model in the presence of an endogenous covariate.

B. DECOMPOSITION OF THE SUBSTANTIVE EQUATIONS

Theorem 1: Let the underlying model be as depicted in (3) *and* (4). *Denote the random disturbances* ε_i *and* ϵ_{1i} *which are constructed as:* $\varepsilon_i = y_{1i}^* - \mathbb{E}[y_{1i}^* | \boldsymbol{x}_i]$ *and* $\epsilon_{1i} = y_{1i} - \mathbb{E}[y_{1i}^* | y_{2i} = 1]$, *respectively. The following requirements must hold:*

(i) $\mathbb{E}[y_{1i}^* | y_{2i} = 1] = \mathbb{E}[\boldsymbol{x}_i^T \boldsymbol{\beta} \,|\, \boldsymbol{x}_i] + \mathbb{E}[\xi_{1i} \,|\, \boldsymbol{x}_i] + \mathbb{E}[\varepsilon_i | y_{2i} = 1]$ $\forall i \in \{1, ..., N\}$;

(ii) $y_{1i} = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_{1i}^*, \quad \mathbb{E}[\epsilon_{1i} \,|\, y_{2i} = 1] = 0,$
$\epsilon_{1i}^* \equiv \epsilon_{1i} + \mathbb{E}[\xi_{1i} \,|\, \boldsymbol{x}_i] + \mathcal{M}(\boldsymbol{w_i}^T \boldsymbol{\gamma}), \quad \forall i \in \{i | y_{2i} = 1\}.$

Proof: By construction $\varepsilon_i = y_{1i}^* - \mathbb{E}[y_{1i}^* | \boldsymbol{x}_i]$, it follows that

$$y_{1i}^* = \mathbb{E}[\boldsymbol{x}_i^T \boldsymbol{\beta} \,|\, \boldsymbol{x}_i] + \mathbb{E}[\xi_{1i} \,|\, \boldsymbol{x}_i] + \varepsilon_i \tag{7}$$

[12] His approach is a generalization of the well-known inverse-mills ratio estimator introduced by [3] for the substantive equation's bias term $\mathbb{E}\left[\xi_{1i} | \xi_{2i} > -\boldsymbol{w}_i^T \boldsymbol{\gamma}\right]$ in the case of a censored sample selection model. Note the difference between censored data and truncated data, which is the case we deal with.

[13] There is an open question whether there is a way to choose a bandwidth sequence that is optimal for the estimation of the parameters [17].

[14] "The well known bandwidth selection rules used in non-parametric estimation, such as cross validation, are not generally applicable to semiparametric settings." [18, p. 191]

Using (7) we get:

$$\begin{aligned}&\mathbb{E}[y_{1i}^*|y_{2i}=1]\\&=\mathbb{E}[\varepsilon_i|y_{2i}=1]\\&\quad+\mathbb{E}\left\{\mathbb{E}[\boldsymbol{x}_i^T\boldsymbol{\beta}\,|\,\boldsymbol{x}_i]|y_{2i}=1\right\}+\mathbb{E}\left\{\mathbb{E}[\xi_{1i}\,|\,\boldsymbol{x}_i]\right\}|y_{2i}=1\}\end{aligned}\tag{8}$$

which is simplified to:

$$\mathbb{E}[y_{1i}^*|y_{2i}=1]=\mathbb{E}[\boldsymbol{x}_i^T\boldsymbol{\beta}\,|\,\boldsymbol{x}_i]+\mathbb{E}[\xi_{1i}\,|\,\boldsymbol{x}_i]+\mathbb{E}[\varepsilon_i|y_{2i}=1]\tag{9}$$

In order to obtain the substantive equation in the truncated environment, we construct $\epsilon_{1i}=y_{1i}-\mathbb{E}[y_{1i}^*|y_{2i}=1]$ where $\mathbb{E}[y_{1i}^*|y_{2i}=1]$ is obtained from (9).[15] Following [17], the conditional expectation of ε_i, given participation is expressed by some unknown function $\mathcal{M}_1(\cdot)$ as $\mathbb{E}[\varepsilon_i|y_{2i}=1]=\mathcal{M}_1(\boldsymbol{w}_i^T\boldsymbol{\gamma})$. Thus, we obtain:

$$y_{1i}=\underbrace{\boldsymbol{x}_i^T\boldsymbol{\beta}}_{\substack{\text{substantive}\\\text{covariates}}}+\overbrace{\underbrace{\mathcal{M}_1(\boldsymbol{w}_i^T\boldsymbol{\gamma})}_{\substack{\text{selection bias}\\\text{term}}}+\underbrace{\mathbb{E}[\xi_{1i}\,|\,\boldsymbol{x}_i]}_{\substack{\text{endogeneity bias}\\\text{term}}}}^{\epsilon_{1i}^*}+\underbrace{\epsilon_{1i}}_{\substack{\text{white}\\\text{noise}}}\tag{10}$$

For sake of brevity we present equation (10), which is a decomposition of the substantive equation into its components, such as the substantive equation's covariates, selection bias term, endogeneity bias term and a stochastic white noise term. It is easy to see that the conventional IV cannot be sufficient in eliminating the endogeneity bias $\mathbb{E}[\xi_{1i}\,|\,\boldsymbol{x}_i]$ in (10), since under truncation the endogeneity bias term is actually $\mathbb{E}[\xi_{1i}\,|\,\boldsymbol{x}_i, y_{2i}=1]$.

Similarly, we construct $\epsilon_{2i}=x_{1i}-\mathbb{E}[x_{1i}^*|y_{2i}=1]$ where $\mathbb{E}[x_{1i}^*|y_{2i}=1]$ satisfies:

$$\mathbb{E}[x_{1i}^*|y_{2i}=1]=\mathbb{E}[\mathrm{v}_i|y_{2i}=1]+\mathbb{E}[[\boldsymbol{z}_i^T,\boldsymbol{x}_{-1_i}^T]\boldsymbol{\delta}\,|\,y_{2i}=1]\tag{11}$$

to get:

$$\epsilon_{2i}=x_{1i}-\mathbb{E}[\mathrm{v}_i|y_{2i}=1]-\mathbb{E}[[\boldsymbol{z}_i^T,\boldsymbol{x}_{-1_i}^T]\boldsymbol{\delta}\,|\,y_{2i}=1]\tag{12}$$

We express $\mathbb{E}[\mathrm{v}_i|y_{2i}=1]$ in (12) as $\mathbb{E}[\mathrm{v}_i|y_{2i}=1]=\mathcal{M}_2(\boldsymbol{w}_i^T\boldsymbol{\gamma})$ where $\mathcal{M}_2(\cdot)$ is some unknown function and obtain (see Theorem 4 to follow):

$$x_{1i}=\underbrace{[\boldsymbol{z}_i^T,\boldsymbol{x}_{-1_i}^T]\boldsymbol{\delta}}_{\substack{\text{substantive}\\\text{covariates}}}+\overbrace{\underbrace{\mathcal{M}_2(\boldsymbol{w}_i^T\boldsymbol{\gamma})}_{\substack{\text{selection bias}\\\text{term}}}+\underbrace{\epsilon_{2i}}_{\substack{\text{white}\\\text{noise}}}}^{\epsilon_{2i}^*}\tag{13}$$

It is easy to see the joint dependence of ϵ_{2i}^* and ϵ_{1i}^* through the selection bias terms in (10) and (13).

■

[15]By construction of y_{1i}, the equality $\mathbb{E}[y_{1i}|y_{2i}=1]=\mathbb{E}[y_{1i}^*|y_{2i}=1]$ must be satisfied. It implies that $\mathbb{E}[\epsilon_{1i}|y_{2i}=1]=\mathbb{E}[y_{1i}|y_{2i}=1]-\mathbb{E}\left\{\mathbb{E}[y_{1i}^*|y_{2i}=1]|y_{2i}=1\right\}=\mathbb{E}[y_{1i}|y_{2i}=1]-\mathbb{E}[y_{1i}^*|y_{2i}=1]=0$.

Next we formulate the relationship between the covariates and dependent variables in the equations to be estimated, in the presence of an endogenous covariate in the substantive equation under truncation.

C. TRUNCATED SAMPLE SELECTION MODEL WITH AN ENDOGENOUS COVARIATE

In cases where the substantive equation's dependent variable is a function of an endogenous covariate x_{1i}, both x_{1i} as well as y_{1i} (as in (4)) are truncated, we face a truncated sample selection model with an endogenous covariate.

Thus, the semiparametric partially linear index model in a truncated environment consists of the following system of equations:

$$\begin{bmatrix}y_{1i}\\x_{1i}\end{bmatrix}=\begin{cases}\boldsymbol{x}_i^T\boldsymbol{\beta}+\mathcal{M}_1(\boldsymbol{w}_i^T\boldsymbol{\gamma})+\overbrace{\mathbb{E}[\xi_{1i}|\boldsymbol{x}_i]+\underbrace{\epsilon_{1i}}_{\text{white noise}}}^{\epsilon_{1i}^{**}}\\[2ex]\underbrace{[\boldsymbol{z}_i^T,\boldsymbol{x}_{-1_i}^T]\boldsymbol{\delta}+\mathcal{M}_2(\boldsymbol{w}_i^T\boldsymbol{\gamma})+\epsilon_{2i}}_{\text{white noise}}\end{cases}\tag{14}$$

where ϵ_{1i}^{**} and ϵ_{2i} are two jointly dependent random disturbances,[16] and by construction are independent of the random variables vector $\mathbf{w}$.[17] The intrinsic endogeneity in the model is captured by the joint dependence of ϵ_{1i}^{**} and the covariates.[18] The presence of the function $\mathcal{M}_2(.)$ implies that we allow for a dependence between $\mathrm{v_i}$ (the endogenous part of x_i) and the selection equation's random disturbance ξ_{2i} (in (1), the complete, non-truncated, sample selection equation).

Our primary interest is to show that the instrumental variable and the random disturbance might be correlated in a truncated environment as will be depicted in Theorem 2 to follow. By doing this, we denote a truncated environment using the indicator (selection variable) $\mathrm{s}=I(\xi_{2i}>-\mathbf{w}^T\boldsymbol{\gamma})$ and postulate the following assumptions:

Assumption 1: The instrumental variable z *is jointly distributed with all covariates in the data:* $\mathbb{E}[\mathrm{z}|\mathbf{w}\ \mathcal{D}\ \boldsymbol{w}]=\mathcal{G}(\mathbf{w})$ *where* $\mathcal{G}(\cdot)$ *is some function of* $\mathbf{w}$.

Assumption 2: Conditioning the instrumental variable z *both on random variable* $\mathbf{w}$ *and a stochastic function of* $\mathbf{w}$ *denoted by* $\mathcal{F}(\mathbf{w},\varepsilon)$ *(given that the stochastic component* ε *is an i.i.d white noise which is independent of* z*), would be the same as conditioning it only on* $\mathbf{w}$. *Formally:* $\mathbb{E}[\mathrm{z}|\mathbf{w}\ \mathcal{D}\ \boldsymbol{w},\mathcal{F}(\mathbf{w},\varepsilon)]=\mathbb{E}[\mathrm{z}|\mathbf{w}\ \mathcal{D}\ \boldsymbol{w}]$.

These two assumptions implies that the conditional expectation of the instrumental variable, given the selection variable, is a function of the random variable vector $\mathbf{w}$, as the following proposition argues:

[16]There is dependence of these two random disturbances due to the dependence between $\mathrm{v_i}$ and $\xi_{1\mathrm{i}}$ (as in (2)) in the complete (non-truncated) data.

[17]Not to be confused with its realization $\boldsymbol{w_i}$.

[18]The intrinsic model's endogeneity is related to the joint dependence of the random disturbance and the covariates *in the population*, unlike a conditional joint dependence of the random disturbance and the covariates given participation in the sample.

Proposition 1: Given assumptions 1 and 2, the conditional expectation of the instrumental variable given the selection variable is a function of $\mathbf{w}$, *rather* $\mathbb{E}[\mathrm{z}|\mathrm{s}=s]=\int_{\boldsymbol{w}}\mathcal{G}(\boldsymbol{w})f_{\mathbf{w}|\mathrm{s}=s}(\boldsymbol{w}|\mathrm{s}=s)d\boldsymbol{w}\ \forall s\in\{0,1\}$, *where* $f_{\mathbf{w}|\mathrm{s}=1}(\boldsymbol{w}|\mathrm{s}=1)$ *and* $f_{\mathbf{w}|\mathrm{s}=0}(\boldsymbol{w}|\mathrm{s}=0)$ *are the conditional density functions of vector* $\mathbf{w}$ *given participation and non-participation, respectively.*

Proof: It easy to see that $\mathbf{w}$ mediates between z and s using the Tower property of conditional expectation [19][19]:

$$\mathbb{E}[\mathrm{z}|\mathrm{s}=s]=\mathbb{E}_{\mathbf{w}}\left[\mathbb{E}[\mathrm{z}|\mathbf{w},\mathrm{s}]|\mathrm{s}=s\right],\quad s\in\{0,1\}$$

The indicator variable s is a stochastic function of $\mathbf{w}$, thus, it follows from assumption 2 that

$$\mathbb{E}_{\mathbf{w}}\left[\mathbb{E}[\mathrm{z}|\mathbf{w},\mathrm{s}]|\mathrm{s}=s\right]=\mathbb{E}_{\mathbf{w}}[\mathbb{E}\left[\mathrm{z}|\mathbf{w}\right]|\mathrm{s}=s],\quad s\in\{0,1\}$$

Following assumption 1 we get:

$$\begin{aligned}\mathbb{E}[\mathrm{z}|\mathrm{s}=s]&=\mathbb{E}_{\mathbf{w}}[\mathbb{E}\left[\mathrm{z}|\mathbf{w}\right]|\mathrm{s}=s]+\mathbb{E}_{\mathbf{w}}[\mathcal{G}|\mathrm{s}=s]\\&=\int_{\boldsymbol{w}}\mathcal{G}(\boldsymbol{w})f_{\mathbf{w}|\mathrm{s}=s}(\boldsymbol{w}|\mathrm{s}=s)d\boldsymbol{w},\quad s\in\{0,1\}.\end{aligned}\tag{15}$$

■

In Theorem 2 to follow we use proposition 1 and present our primary argument: in truncated sample selection models, the orthogonality condition of the instrumental variable with respect to the random disturbance might be violated. This violation stems from a dependency between the instrumental variables and the selection equation's covariates.

Theorem 2 (Lack of orthogonality): Let ξ_1 *and* ξ_2 *be two jointly distributed random disturbances, and let* z *be a valid instrumental variable satisfying* $\mathbb{E}[\mathrm{z}\cdot\xi_1]=0$. *Denote a random variables vector* $\mathbf{w}\in\mathbb{R}^l$, *a parameters vector* $\boldsymbol{\gamma}\in\mathbb{R}^l$ *and a truncated environment using the indicator variable* $\mathrm{s}=I(\xi_2>-\mathbf{w}'\boldsymbol{\gamma})$. *Suppose that the following conditions are satisfied:*
(i) assumptions 1 and 2 hold;
(ii) $\mathbb{E}[\xi_1|\mathrm{s}=s,\mathbf{w}\ \boldsymbol{\mathcal{D}}\ \boldsymbol{w}]=\mathcal{M}(\boldsymbol{w}^T\boldsymbol{\gamma})$; *(iii)* z *and* ξ_1 *are conditionally independent given* $\mathbf{w}$ *and* s; *(iv)* $\mathcal{G}$ *and* $\mathcal{M}$ *are linearly dependent in the truncated environment (given* s*).*[20]

Under conditions (i)-(iv) above, z *is not orthogonal to the random disturbance* ξ_1 *given* s.

Proof: Using the Tower property, the following must hold:

$$\begin{aligned}&\mathbb{E}[\mathrm{z}\xi_1|\mathrm{s}=s]\\&=\mathbb{E}_{\mathbf{w}}\left[\mathbb{E}[\mathrm{z}\xi_1|\mathbf{w},\mathrm{s}]|\mathrm{s}=s\right]\\&=\underbrace{\mathbb{E}_{\mathbf{w}}\left[\mathbb{E}_{\mathrm{z}}[\mathrm{z}|\mathbf{w},\mathrm{s}]\mathbb{E}_{\xi_1}[\xi_1|\mathbf{w},\mathrm{s}]|\mathrm{s}=s\right]}_{\substack{\text{by conditional independence of}\\ \text{z and }\xi_1\text{ given }\mathbf{w}\text{ and s}}}\\&=\mathbb{E}_{\mathbf{w}}[\mathcal{G}\mathcal{M}|\mathrm{s}=s]=\int_{\boldsymbol{w}}\mathcal{G}(\boldsymbol{w})\mathcal{M}(\boldsymbol{w}^T\boldsymbol{\gamma})f_{\mathbf{w}|\mathrm{s}=s}(\boldsymbol{w}|\mathrm{s}=s)d\boldsymbol{w}\end{aligned}$$

[19]The Tower property is referred interchangeability to the law of iterated expectations. For formal proof see [19].

[20]The conditional linear dependence between $\mathcal{G}$ and $\mathcal{M}$ given the indicator (selection) variable implies that $\mathbb{E}[\mathcal{G}\mathcal{M}|\mathrm{s}=s]\neq\mathbb{E}[\mathcal{G}|\mathrm{s}=s]\,\mathbb{E}[\mathcal{M}|\mathrm{s}=s]$. Since $\mathcal{G}$ and $\mathcal{M}$ are both functions of the random variable $\mathbf{w}$, this inequality implies that $\int\mathcal{G}(\boldsymbol{w})\mathcal{M}(\boldsymbol{w}^T\boldsymbol{\gamma})f_{\mathbf{w}|\mathrm{s}=s}(\boldsymbol{w})d\boldsymbol{w}\neq\int\mathcal{G}(\boldsymbol{w})f_{\mathbf{w}|\mathrm{s}=s}(\boldsymbol{w})d\boldsymbol{w}\int\mathcal{M}(\boldsymbol{w}^T\boldsymbol{\gamma})f_{\mathbf{w}|\mathrm{s}=s}(\boldsymbol{w})d\boldsymbol{w}$.

Similarly (using proposition 1),

$$\mathbb{E}[\mathrm{z}|\mathrm{s}=s]=\int_{\boldsymbol{w}}\mathcal{G}(\boldsymbol{w})f_{\mathbf{w}|\mathrm{s}=s}(\boldsymbol{w}|\mathrm{s}=s)d\boldsymbol{w}$$

and

$$\begin{aligned}&\mathbb{E}[\xi_1|\mathrm{s}=s]\\&=\mathbb{E}_{\mathbf{w}}\left[\mathbb{E}[\xi_1|\mathbf{w},\mathrm{s}]|\mathrm{s}=s\right]\\&=\mathbb{E}_{\mathbf{w}}[\mathcal{M}|\mathrm{s}=s]=\int_{\boldsymbol{w}}\mathcal{M}(\boldsymbol{w}^T\boldsymbol{\gamma})f_{\mathbf{w}|\mathrm{s}=s}(\boldsymbol{w}|\mathrm{s}=s)d\boldsymbol{w}\end{aligned}\tag{16}$$

As $\mathcal{G}$ and $\mathcal{M}$ are conditionally linearly dependent random variables in the truncated environment (given s), implies:

$$\begin{aligned}&\int\mathcal{G}(\boldsymbol{w})\mathcal{M}(\boldsymbol{w}^T\boldsymbol{\gamma})f_{\mathbf{w}|\mathrm{s}=s}(\boldsymbol{w})d\boldsymbol{w}\\&\qquad\neq\int\mathcal{G}(\boldsymbol{w})f_{\mathbf{w}|\mathrm{s}=s}(\boldsymbol{w})d\boldsymbol{w}\int\mathcal{M}(\boldsymbol{w}^T\boldsymbol{\gamma})f_{\mathbf{w}|\mathrm{s}=s}(\boldsymbol{w})d\boldsymbol{w}\end{aligned}$$

and consequently:

$$\begin{aligned}&\mathbb{E}[\mathrm{z}\xi_1|\mathrm{s}=s]\\&\neq\mathbb{E}[\mathrm{z}|\mathrm{s}=s]\mathbb{E}[\xi_1|\mathrm{s}=s]\Rightarrow\mathbb{COV}[\mathrm{z},\xi_1|\mathrm{s}=s]\neq0\end{aligned}\tag{17}$$

Therefore, z is not orthogonal to ξ_1 given s (in the truncated environment). ■

However, the orthogonality condition can be satisfied by removing the contamination factor, which is the covariate generating the comovement between the random disturbance and the instrumental variable, as shown in the following Theorem 3.

Theorem 3 (Bias removal): Let ξ_1 *and* ξ_2 *be two jointly distributed random disturbances, and let* z *be a valid instrumental variable satisfying* $\mathbb{E}[\mathrm{z}\cdot\xi_1]=0$. *Denote a random variables vector* $\mathbf{w}\in\mathbb{R}^l$, *a parameters vector* $\boldsymbol{\gamma}\in\mathbb{R}^l$ *and a truncated environment using the indicator variable* $\mathrm{s}=I(\xi_2>-\mathbf{w}'\boldsymbol{\gamma})$. *Suppose that the following conditions are satisfied: (i) assumptions 1 and 2 hold;*
(ii) $\mathbb{E}[\xi_1|\mathrm{s}=s,\mathbf{w}\ \boldsymbol{\mathcal{D}}\ \boldsymbol{w}]=\mathcal{M}(\boldsymbol{w}^T\boldsymbol{\gamma})$.

Under conditions (i) and (ii) above, removing the contamination factor (the bias term) from the residual in the truncated environment leads to orthogonality of the instrumental variable to the substantive equation's disturbance, such that: $\mathbb{E}[\mathrm{z}\left[\xi_1-\mathcal{M}(\mathbf{w}^{\mathrm{T}}\boldsymbol{\gamma})\right]|\mathrm{s}=s]=0$.

Proof: Express $\mathbb{E}[\mathrm{z}\left[\xi_1-\mathcal{M}(\mathbf{w}^{\mathrm{T}}\boldsymbol{\gamma})\right]|\mathrm{s}=s]$ as a difference of two conditional expectations:

$$\begin{aligned}&\mathbb{E}[\mathrm{z}\left[\xi_1-\mathcal{M}(\mathbf{w}^{\mathrm{T}}\boldsymbol{\gamma})\right]|\mathrm{s}=s]\\&\qquad=\mathbb{E}[\mathrm{z}\xi_1|\mathrm{s}=s]-\mathbb{E}[\mathrm{z}\mathcal{M}(\mathbf{w}^{\mathrm{T}}\boldsymbol{\gamma})|\mathrm{s}=s]\end{aligned}$$

Using the Tower property, to get:

$$\begin{aligned}&\mathbb{E}[\mathrm{z}\mathcal{M}(\mathbf{w}^{\mathrm{T}}\boldsymbol{\gamma})|\mathrm{s}=s]\\&=\mathbb{E}_{\mathbf{w}}\left[\mathbb{E}[\mathrm{z}\mathcal{M}(\mathbf{w}^{\mathrm{T}}\boldsymbol{\gamma})|\mathbf{w},\mathrm{s}]|\mathrm{s}=s\right]\\&=\underbrace{\mathbb{E}_{\mathbf{w}}\left[\mathbb{E}_{\mathrm{z}}[\mathrm{z}|\mathbf{w},\mathrm{s}]\mathbb{E}[\mathcal{M}(\mathbf{w}^{\mathrm{T}}\boldsymbol{\gamma})|\mathbf{w},\mathrm{s}]|\mathrm{s}=s\right]}_{\substack{\text{by conditional independence of}\\ \text{z and }\mathcal{M}(\mathbf{w}^T\boldsymbol{\gamma})\text{ given }\mathbf{w}\text{ and s}}}\\&=\mathbb{E}[\mathcal{G}(\mathbf{w})\mathcal{M}(\mathbf{w}^{\mathrm{T}}\boldsymbol{\gamma})|\mathrm{s}=s]\end{aligned}$$

As $\mathbb{E}[z\xi_1|s=s]=\mathbb{E}[\mathcal{G}(\mathbf{w})\mathcal{M}(\mathbf{w}^{\mathrm{T}}\boldsymbol{\gamma})|s=s]$ (proof of Theorem 2), which implies that $\mathbb{E}[z\left[\xi_1-\mathcal{M}(\mathbf{w}^{\mathrm{T}}\boldsymbol{\gamma})\right]|s=s]=0$.

Moreover,

$$\begin{aligned}&\mathbb{COV}\left[z,\xi_1-\mathcal{M}(\mathbf{w}^T\boldsymbol{\gamma})|s=s\right]\\&=\underbrace{\mathbb{E}[z\left[\xi_1-\mathcal{M}(\mathbf{w}^{\mathrm{T}}\boldsymbol{\gamma})\right]|s=s]}_{0}\\&\quad-\mathbb{E}[z|s=s]\underbrace{\mathbb{E}[\xi_1-\mathcal{M}(\mathbf{w}^{\mathrm{T}}\boldsymbol{\gamma})|s=s]}_{0}=0.\end{aligned}\tag{18}$$

■

Therefore, a valid instrumental variable z is orthogonal to the truncated distribution (non-contaminated) disturbance ϵ_{1i}^{**} in (14), even though z and $\mathbf{w}$ are dependent.

The joint dependence of (ξ_1,ξ_2,v) implies the violation of zero mean expectation (under truncation) in the x_{1i} regression equation (4), such that $\mathbb{E}[v|\xi_2>-\boldsymbol{w'\gamma}]=\mathcal{M}_2(\boldsymbol{w}^T\boldsymbol{\gamma})\neq\mathbb{E}[v]=0$. That is, the conditional expectation of v given an endogenous truncation is a function of the covariate vector $\boldsymbol{w}$, while in the population it does not depend on $\boldsymbol{w}$ and has a zero mean expectation. This violation is a precondition for the endogeneity of $\{\boldsymbol{x_{-1i}},z_i\}$ with respect to v_i given participation in the regression of x_{1i}.[21] The following theorem indicates that such violation is also obtained in cases where the comovement of v and ξ_2 is entirely related to a variation in ξ_1.

Theorem 4 (Conditional independence): Let ξ_1 and ξ_2 be two jointly distributed random disturbances of the substantive and selection equations, respectively. Let v *be a random variable which depends on ξ_1 such that* v *and ξ_2 are conditionally independent given ξ_1. Denote a random variables vector $\mathbf{w}\in\mathbb{R}^l$ independent of (ξ_1,ξ_2,v) with a realization $\boldsymbol{w}$, a parameters vector $\boldsymbol{\gamma}\in\mathbb{R}^l$ and a truncated environment using the indicator variable $s=I(\xi_2>-\boldsymbol{w'\gamma})$.*

Assume the following conditions are satisfied: (i) the conditional expectation of the random disturbance given participation is $\mathbb{E}[\xi_1|\xi_2>-\boldsymbol{w'\gamma}]=\mathcal{M}_1(\boldsymbol{w}^T\boldsymbol{\gamma})$ [16]; (ii) $\mathbb{E}[v|\xi_1,\xi_2>-\boldsymbol{w'\gamma}]=\mathbb{E}[v|\xi_1]=\mathcal{H}(\xi_1)$, (endogeneity); (iii) $\mathcal{H}(.)$, a monotonic mapping $\mathbb{R}\mapsto\mathbb{R}$.

Under conditions (i)-(iii) above, $\mathbb{E}[v|\xi_2>-\boldsymbol{w'\gamma}]\neq\mathbb{E}[v]$ regardless of the conditional independence of v *and ξ_2 given ξ_1.*

Proof: Applying Tower property to $\mathbb{E}[v|s=1]$:

$$\begin{aligned}&\mathbb{E}[v|\xi_2>-\boldsymbol{w'\gamma}]\\&=\mathbb{E}[v|s=1]=\mathbb{E}_{\xi_1}\{\mathbb{E}[v|\xi_1,s]|s=1\}\\&=\mathbb{E}[\mathcal{H}(\xi_1)|s=1]=\mathcal{M}_2(\boldsymbol{w}^T\boldsymbol{\gamma})\neq\mathbb{E}[v].\end{aligned}$$

It can be shown that ξ_1 mediates between v and s (participation), in that it generates a comovement between the random variables v and s. The last equality relies on the fact that the random variable $\mathcal{H}(\xi_1)$ is a monotonic mapping of ξ_1, implying dependence on s due to the dependency between ξ_1 and s. ■

Next we show that the conventional IV estimator is inconsistent in the presence of a truncated environment in which the expectation of the instrumental variable and the random disturbance are functions of the selection equation's covariates vector $\mathbf{w}$. The proof in section II-D to follow, relies on a linear dependence assumption between these two functions of $\mathbf{w}$. The rationale for the linear dependence is due to the fact that the random disturbance's (ξ_1) conditional expectation generally satisfies monotonicity with respect to the index variable $\mathbf{w'\gamma}$. Therefore, it is enough to assume that, on average, z is affected monotonically by the index variable $\mathbf{w'\gamma}$ to generate a linear dependence between z and the conditional expectation of ξ_1 given participation.[22]

D. THE CONVENTIONAL IV ESTIMATOR'S ASYMPTOTIC BIAS

The IV estimator's asymptotic bias is:

$$\begin{aligned}&\widehat{\boldsymbol{\beta}}_{\mathbf{iv}}=(\mathbf{z}^T\mathbf{x})^{-1}\mathbf{z}^T\mathbf{y}_1=(\mathbf{z}^T\mathbf{x})^{-1}\mathbf{z}^T(\mathbf{x}\boldsymbol{\beta}+\mathcal{M}_1(\mathbf{w}^T\boldsymbol{\gamma})+\epsilon_{1i}^{**})\\&\widehat{\boldsymbol{\beta}}_{\mathbf{iv}}=(\mathbf{z}^T\mathbf{x})^{-1}\mathbf{z}^T(\mathbf{x}\boldsymbol{\beta})+(\mathbf{z}^T\mathbf{x})^{-1}\mathbf{z}^T\mathcal{M}_1(\mathbf{w}^T\boldsymbol{\gamma})+(\mathbf{z}^T\mathbf{x})^{-1}\mathbf{z}^T\epsilon_{1i}^{**}\\&\widehat{\boldsymbol{\beta}}_{\mathbf{iv}}=\boldsymbol{\beta}+(\mathbf{z}^T\mathbf{x})^{-1}\mathbf{z}^T\mathcal{M}_1(\mathbf{w}^T\boldsymbol{\gamma})+(\mathbf{z}^T\mathbf{x})^{-1}\mathbf{z}^T\epsilon_{1i}^{**}\\&\operatorname*{plim}_{N\to\infty}\left[\widehat{\boldsymbol{\beta}}_{\mathbf{iv}}\right]\\&\quad=\boldsymbol{\beta}+\underbrace{\operatorname*{plim}_{N\to\infty}\left[(N^{-1}\mathbf{z}^T\mathbf{x})^{-1}\right]\operatorname*{plim}_{N\to\infty}\left[N^{-1}\mathbf{z}^T\mathcal{M}_1(\mathbf{w}^T\boldsymbol{\gamma})\right]}_{\text{Asymptotic bias}}\end{aligned}\tag{19}$$

Given any correlation between $\mathbf{z}$ and $\mathcal{M}_1(\mathbf{w}^T\boldsymbol{\gamma})$, $\operatorname*{plim}_{N\to\infty}\left[\mathbf{z}^T\mathcal{M}_1(\mathbf{w}^T\boldsymbol{\gamma})\right]\not\to 0$. Thus, the $\widehat{\boldsymbol{\beta}}_{\mathbf{iv}}$ estimator is an inconsistent estimator for $\boldsymbol{\beta}$.

Next we discuss the two types of joint dependence which are present in our model. This is done is order to facilitate the understanding of our proposed procedure, which is intended to correct for the bias propagated by each type of joint dependence.

III. PRELIMINARIES

The objective is to eliminate the selection bias term captured by $\mathcal{M}_1(\cdot)$ in (14). As we don't want to impose a specific distribution function on the random disturbances, the aforementioned elimination should be performed in a nonparametric manner. This can be achieved using a semiparametric estimation method, which is distribution-free. However, the bias term might be a discontinuous function with different levels of smoothness that must be considered. These issues can be alleviated using multi-resolution analysis by employing the wavelet estimator [20]. Wavelet is a bandwidth-free estimator, that is based on the idea of multi-scale representation of

[21] As been discussed in [3], the fact that the conditional disturbance (given participation) in the substantive equation of x_{1i} is a function of the selection equation's covariates, leads to a potential correlation between the disturbance and the substantive equation's covariates. This correlation implies the endogeneity of the substantive equation's covariates $\{\boldsymbol{x_{-1i}},z_i\}$ with respect to its random disturbance v_i given participation.

[22] Both functions are dependent through $\mathbf{w}$ by construction, generally leading to some degree of linear dependence.

the data [21][23] and is used as a denoising technique by simple thresholding, which is based on the concept of sparsity.[24]

The applicability of the classical wavelet estimator is problematic in several important aspects. First, it limits the sample size to be represented as 2^J, with J a non-negative integer, and the observations to be equispaced, which challenges the estimation in case of irregular-spaced data.[25] Second, the classical wavelet estimator imposes the parametric assumption that the disturbances are independent identically distributed normal variables [23]. Lastly, there are the problems of coefficient expansion and border discontinuities.[26]

In order to overcome these limitations, second generation wavelets have been introduced [24] which define wavelets in terms of lifting-steps instead of matrices to reduce computational complexity.[27] An earlier attempt to deal with irregular-spaced data using second generation wavelets is presented in [21] by postulating a prior distribution function for the wavelet coefficients.[28] Alternative approaches extend Haar wavelet transform to accommodate for irregular data [27].

Both first as well as second generation wavelet estimation methods involve three steps: coefficient estimation (forward transform); (ii) denoising by using element-wise thresholding (coefficients selection) and (iii) reconstruction of the data without the noise (inverse transform). It is important to notice that the sequential nature of the estimation that relies on element-wise thresholding is applicable for limited types of wavelets, referred to as orthogonal wavelets which consist of the above described limitations. The main shortcoming of orthogonal wavelets is that the compact support and the symmetry properties which are useful in denoising are conflicting.[29] To preserve both these properties, the biorthogonal wavelet-based JPEG is used [29], [30].[30]

In what follows we briefly explain the concept of biorthogonality. Denote a set of functions $\{\varphi_k(t)\}$ which spans a vector space $\mathcal{F}$, referred to as the expansion set. By construction, any function $g(t) \in \mathcal{F}$ can be expressed by using a series expansion, such that $g(t) = \sum_k \eta_k \varphi_k(t)$, where η_k and φ_k are the expansion coefficients and expansion functions, respectively. The set $\{\varphi_k(t)\}$ is biorthogonal to the set $\{\tilde{\varphi}_k(t)\}$ if $\langle \varphi_k, \tilde{\varphi}_{k'} \rangle = \mathfrak{d}(k-k')\ \forall k$ and k', with $\langle \cdot \rangle$ being the L_2 inner product and the function $\mathfrak{d}(\cdot)$ is the Kronecker delta.[31] These two sets form a biorthogonal system, in which $\{\tilde{\varphi}_k(t)\}$ is referred to as the dual basis of $\{\varphi_k(t)\}$. Thus, we get the following unique representation:

$$\eta_k = \langle g(t), \tilde{\varphi}_k(t) \rangle \tag{20}$$

Substituting each η_k coefficient with its analytic expression in (20), to obtain:

$$g(t) = \sum_k \langle g(t), \tilde{\varphi}_k(t) \rangle \varphi_k(t) \tag{21}$$

Obviously, in the present case of biorthogonality, the coefficients in (20) are obtained by using the dual basis and the function is reconstructed in (21) by using another basis which is the expansion set. In cases where $\{\tilde{\varphi}_k(t)\} = \{\varphi_k(t)\}$ we have an orthogonal basis $\{\varphi_k(t)\}$, which is referred to as self-dual. Therefore, biorthogonality is a generalization of orthogonality that allows for a larger class of expansions.

Recall that our objective is to estimate the bias term for an unknown functional form, captured by $\mathcal{M}_1(\cdot)$ in (14), the conditional expectation of ε_i, given participation defined as $\mathbb{E}[\varepsilon_i | y_{2i} = 1] = \mathcal{M}_1(\boldsymbol{w}_i^T \boldsymbol{\gamma})$.[32] In what follows, we attend to the estimation of $\mathcal{M}_1(\cdot)$ using the wavelet estimator.

We use the concept of *a frame* in (1) to define *Riezs basis* in (2). *Riezs basis* is a building block in the definition of biorthogonal wavelets in (3) to follow.

Let $\mathbb{H}$ be a separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and a norm $\left\| \cdot \right\|_2^2$. We denote a sequence $\mathcal{F} = \{f_k, k \in \Lambda\} \subset \mathbb{H}$, in which $\Lambda \subset \mathbb{Z}$.

We use the following *frame* and *Riesz basis* definitions [31]:

Definition 1 (Frame): $\mathcal{F}$ is called a frame if there are constants $0 < A \leq B$ such that $\forall f \in \mathbb{H}$ $A\left\|f\right\|_2^2 \leq \sum_{k\in\Lambda} |\langle f, f_k \rangle|^2 \leq B\left\|f\right\|_2^2$.

Definition 2 (Riesz Basis): A sequence $\mathcal{F}$ is a Riesz basis if and only if it is a frame having the additional property that upon the removal of any element from the sequence, it ceases to be a frame.

Let $L_2(\mathbb{R})$ be the space of square integrable and real-valued functions on $\mathbb{R}$. We use the following biorthogonal wavelet definition [32]:

Definition 3 (Biorthogonal Wavelet): A pair of functions $\varphi_{j,k}, \tilde{\varphi}_{j,k} \in L_2(\mathbb{R})$ is a pair of biorthogonal wavelets if the sets $\{\varphi_{j,k} | j,k \in \mathbb{Z}\}$ and $\{\tilde{\varphi}_{j,k} | j,k \in \mathbb{Z}\}$ form the Riesz basis for $L_2(\mathbb{R})$ and if any function $g \in L_2(\mathbb{R})$ has the representation: $g = \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \langle g, \tilde{\varphi}_{j,k} \rangle \varphi_{j,k}$.

It is worth noticing that both existence as well as uniqueness of the series representation are satisfied in definition 3.

[23] Due to its multi-scale property, we can distinguish between the important information, the function's average behavior, from the noise. The coarse scales (lower resolution-levels) usually convey important information, while at fine scales there is usually more noise.

[24] Sparsity implies that the majority of wavelet coefficients are small, and can be replaced by zero [22].

[25] The observations location in space or time must be of equal distance.

[26] The standard orthogonal wavelet transform has the shortcoming in that it requires a large number of coefficients (coefficient expansion) to represent the original data [15].

[27] The lifting-steps are consecutive operations of prediction (scaled-moving average) and update (scaled-first difference) to obtain the wavelet coefficients.

[28] [21] adopt the parametric Bayesian denoising approach introduced by [25], [26] to obtain the wavelet coefficients assuming the coefficients are distributed according to a continuous mixture of a normal by a Beta density.

[29] Unlike biorthogonality, orthogonality and symmetry are conflicting properties for design of compactly supported nontrivial wavelets (see Theorem 8.1.4 in [28]).

[30] The JPEG algorithm used here is termed 'wavelet CDF 9/7'.

[31] $\mathfrak{d}(k) = \begin{cases} 1 & \text{if } |k| = 0 \\ 0 & \text{if } |k| > 0 \end{cases}$. The set $\varphi_k(t)$ is orthogonal if $\langle \varphi_k, \varphi_{k'} \rangle = 0$ $\forall k \neq k'$.

[32] For brevity, we present $\mathcal{M}_1(\cdot)$ only. Identical treatment is applied to $\mathcal{M}_2(\cdot)$.

However, our proposed nonparametric estimator might be unstable, rendering the estimation problem ill-posed, which is one of the challenges in nonparametric estimation of unknown functions.[33] This ill-posed problem can be alleviated by employing regularization on the wavelet series expansion coefficients [33], [34].

Let $u_i = y_{1i} - \boldsymbol{x}_i^T \boldsymbol{\beta}$ be the i'th element in vector $\boldsymbol{u}$ of size $n \times 1$, which satisfies:

$$u_i = \mathcal{M}_1(t_i) + \epsilon_{1i} \tag{22}$$

where $\{t_i\}_{i=1}^n$ is a sequence in which the i'th element satisfies $t_i = \boldsymbol{w}_i^T \boldsymbol{\gamma}$ and ϵ_{1i} is the white noise described in (14).[34]

We use $\boldsymbol{\Phi}_I$ and $\boldsymbol{\Phi}_F \equiv \boldsymbol{\Phi}_I^{-1}$ to denote the inverse and forward transformation matrices, respectively, each of size $n \times n$ [35] in an orthogonal wavelet.[35] We note that using orthogonal wavelets one obtains the closed-form solution to the wavelet coefficients as follows:

$$\widehat{\mathcal{M}_1} = \boldsymbol{\Phi}_I \rho_\lambda(\boldsymbol{\Phi}_F \boldsymbol{u}) \tag{23}$$

where $\widehat{\mathcal{M}_1}(\cdot)$ is the estimate of the unknown function $\mathcal{M}_1(\cdot)$, and $\rho_\lambda(\cdot)$ represents the element-wise thresholding generated by some penalty function, in which the tuning parameter is represented by λ. The procedure in (23) to obtain $\widehat{\mathcal{M}_1}(\cdot)$ by employing a thresholding operator is referred to as denoising.

We depart from the denoising procedure in (23) by employing biorthogonal wavelets, as we are interested in the applicability of the general case where the penelization is not an element-wise due to correlations among wavelet regressors.[36] In such cases, there is no such a closed-form solution, which necessitates the regularized least squares optimization method to follow.

Let $\boldsymbol{\Psi}_I$ and $\boldsymbol{\Psi}_F \equiv (\boldsymbol{\Psi}_I^T \boldsymbol{\Psi}_I)^{-1} \boldsymbol{\Psi}_I^T$ denote the inverse and forward transformation matrices, respectively, each of size $n \times n$ of the wavelet-based JPEG, which is a biorthogonal wavelet.[37,38]

Let $\rho_{\lambda,\gamma}(\cdot)$ be the minimax concave penalty (MCP) function [37], defined as [38][39]:

$$\rho_{\lambda,\gamma}(\theta) = \begin{cases} \lambda\theta - \dfrac{\theta^2}{2\gamma} & \text{if } \theta \le \gamma\lambda, \\ \dfrac{1}{2}\lambda^2\gamma & \text{if } \theta > \gamma\lambda \end{cases} \tag{24}$$

where $\theta \in (-\infty, \infty)$ is the parameter to be penalized, $\lambda > 0$ and $\gamma \in (1, \infty)$.

We define resolution-dependent regularized least squares at resolution levels 1, ..., J:

$$\widehat{\boldsymbol{\delta}} = \arg\min_{\boldsymbol{\delta}} \frac{1}{2n} \left\| \boldsymbol{u} - \boldsymbol{\Psi}_I \boldsymbol{\delta} \right\|_2^2 + \sum_{j=1}^{J} P_{\lambda_j,\gamma_j}(\boldsymbol{\delta}_j) \tag{25}$$

where $\boldsymbol{\delta} = [\boldsymbol{\delta}_1^T, \boldsymbol{\delta}_2^T, ..., \boldsymbol{\delta}_J^T]^T$ is the wavelet coefficient vector of size $n \times 1$ and $\boldsymbol{\delta}_j$ is of size $n_j \times 1$. $\|\cdot\|_2$ is the usual ℓ_2 (Euclidean) norm, defined as $\|\boldsymbol{b}\|_2 = \left(\sum_{i=1}^n |b_i|^2\right)^{1/2}$. The penalty function is $P_{\lambda_j,\gamma_j}(\boldsymbol{\delta}_j) = \sum_{k=1}^n \rho_{\lambda_j,\gamma_j}(|\delta_{j,k}|)$.

It is evident that when the $\boldsymbol{\delta}_j \to 0$, the bias propagated by the endogenous truncation approaches zero and thus, our algorithm is reduced to the conventional IV estimator.

The univariate solution of a regularized least squares problem using the penalty function in (24) is denoted by $S_\alpha(\cdot)$ and defined as[40]:

$$S_\alpha(\tilde{\delta}; \lambda, \gamma) = \begin{cases} \dfrac{1}{1 - 1/(\alpha\gamma)} \text{sign}(\tilde{\delta}) \max(\left|\tilde{\delta}\right| - \dfrac{\lambda}{\alpha}) & \text{if } \left|\tilde{\delta}\right| \le \gamma\lambda, \\ \tilde{\delta} & \text{if } \left|\tilde{\delta}\right| > \gamma\lambda \end{cases} \tag{26}$$

where $\alpha \in (0, \infty)$. It is worth noting that if $\gamma \to \infty$ the solution is soft-thresholding introduced by [40]; in case that $\alpha\gamma \to 1^+$ the solution is hard-thresholding (see proof in the Appendix VI-A).

To reduce computational complexity, the optimization problem in (25) is reformulated as:

$$\boldsymbol{\delta}^{(iter+1)} = \arg\min_{\boldsymbol{\delta}} \frac{1}{2n} ((\boldsymbol{\delta} - \boldsymbol{\delta}^{(iter)})^T \boldsymbol{\Psi}_I^T (\boldsymbol{u} - \boldsymbol{\Psi}_I \boldsymbol{\delta}^{(iter)})) + \alpha/2 \left\| \boldsymbol{\delta} - \boldsymbol{\delta}^{(iter)} \right\|_2^2 + \sum_{j=1}^{J} P_{\lambda_j,\gamma_j}(\boldsymbol{\delta}_j) \tag{27}$$

where Ψ_I^T is the transpose of matrix Ψ_I, $\alpha \boldsymbol{I}$ is an approximation of the Hessian and $\boldsymbol{I}$ is the identity matrix of size $n \times n$. The number of iterations is denoted by the integer *iter*.

[33] An estimator violating at least one of the requirements: existence, uniqueness and stability is referred to as ill-posed.

[34] A more general formulation solves the ill-posed problem by employing regularization in cases where a linear transform of the unknown function replaces the original function [33].

[35] The forward transform is referred to as the Discrete Wavelet Transform (DWT).

[36] It has been shown that the performance of wavelet estimator can be improved when the dependencies among coefficients were taken into account [36].

[37] For a definition of biorthogonal wavelets see [24].

[38] Unlike biorthogonality, orthogonality implies that the wavelet regressors are mutually uncorrelated and that the inverse transform is the transpose of the forward transform. This simplifies the computation as the wavelet coefficients are obtained analytically (closed-form) using element-wise thresholding operators (e.g., hard and soft thresholding operators). However, we opted for the biorthogonality wavelet to exploit the correlation structure of the regressors. Biorthogonal wavelets preserve the perfect reconstruction property (by employing dual-filters) as well, but is more flexible in that the inverse of $\boldsymbol{X}$ is not required to be its transpose. Consequently, the thresholding is applied to the entire coefficient vector.

[39] The penalty function in (24) represents a family of penalty functions as a generalization of the soft thresholding (if $\gamma \to \infty$) and hard thresholding (if $\gamma \to 1^+$) [38].

[40] In order to utilize the min-max concave (MCP) penalty function in (24), we depart from the regularized least squares algorithm in [39], as it is limited to its special case of the LASSO penalty function. We introduce α as an approximation to the Hessian of the least squares problem in order to obtain an element-wise thresholding. This amounts to a dimensional reduction technique for reducing computational complexity. For the special case of $\alpha = 1$, see [38].

For brevity, we divide the argument to be minimized by α and complete the squares using the expressions in $\|\cdot\|_2^2$ [39] to get:

$$\boldsymbol{\delta}^{(iter+1)} = \arg\min_{\boldsymbol{\delta}} \frac{1}{2}\left\|\boldsymbol{\delta} - \boldsymbol{\delta}^{(iter)} + 1/(\alpha n)\boldsymbol{\Psi}_I^T(\boldsymbol{u} - \boldsymbol{\Psi}_I\boldsymbol{\delta}^{(iter)})\right\|_2^2 + \frac{1}{\alpha}\sum_{j=1}^{J} P_{\lambda_j,\gamma_j}(\boldsymbol{\delta}_j) \quad (28)$$

The iterative procedure performs MCP-thresholding on a proximal gradient-descent update for $k = 1, ..., n$ (see Algorithm 5 in the Appendix):

$$\delta_{j,k}^{(iter+1)} = S_\alpha\left(\delta_{j,k}^{(iter)} + 1/(\alpha n)\boldsymbol{\psi}_{I_k}^T(\boldsymbol{u} - \boldsymbol{\Psi}_I\boldsymbol{\delta}^{(iter)}); \lambda_j, \gamma_j\right) \quad (29)$$

where $\delta_{j,k}^{(iter)}$ is the k'th coefficient in vector $\boldsymbol{\delta}_j^{(iter)}$ and $\boldsymbol{\psi}_{I_k}$ is k'th column in $\boldsymbol{\Psi}_I$ (the inverse wavelet transform). The notation $\boldsymbol{\psi}_{I_k}^T$ implies the transpose of $\boldsymbol{\psi}_{I_k}$. We use (29) to update the wavelet coefficients iteratively until the update is negligible, such that the following convergence criterion is satisfied:

$$\left\|\boldsymbol{\delta}^{(iter+1)} - \boldsymbol{\delta}^{(iter)}\right\|_2^2 / \left\|\boldsymbol{\delta}^{(iter)}\right\|_2^2 < \tau \quad (30)$$

where τ is the tolerance which is a positive real number that we arbitrarily set to 10^{-16}.

The optimization method in (29) involves matrices multiplication which is computationally infeasible for large data sets. To alleviate this computational complexity we develop a lifting scheme to be employed in order to perform simultaneously the transposed-inverse of the wavelet transform, consisting of lifting steps (see Algorithms 1-4 to follow). Conventionally, a lifting step can be either a prediction, that is a procedure generating a smoothed version of the data (the scaled coefficients), or an update that is the procedure to generate the remainder (the detail coefficients) between the data and its smoothed version. For the present case we define a new operator because the existing lifting steps do not provide an analytic representation of the transposed-inverse, as discussed in [7].

In the next section we discuss the main idea behind lifting steps, in order to obtain analytically the transposed-inverse transform. First we describe the lifting steps in a regular-spaced data given a sample size of 2^J for a non-negative integer J. Then in equations (1)-(IV-E) to follow, we alleviate these two restrictions by formulating our proposed algorithm.

A. LIFTING STEPS TO OBTAIN THE WAVELET COEFFICIENTS

Let $w = (w_1, ...w_n)$ be a discrete sequence of data consisting of n real numbers, such that the sequence is referred to as *dyatic* iff $n = 2^J$ for some integer $J \geq 0$. The sequence can be expressed uniquely in terms of detail (difference) and summation coefficients denoted by $\{d_{J-1,k}\}_{k=1}^{n/2}$ and $\{c_{J-1,k}\}_{k=1}^{n/2}$, respectively. The former capture the variation in the sequence at different scales and locations and the latter are a smooth representation of the original sequence.

The multi-scale representation of a function $g \in L_2(\mathbb{R})$ is obtained as follows:

$$g(t) = \sum_{k\in\mathbb{Z}} c_{0,k}\phi_{0,k}(t) + \sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} d_{j,k}\varphi_{j,k}(t) \quad (31)$$

The first set of terms, $\phi_{0,k}$, represents the average level of function g and the second set of terms $\varphi_{j,k}$ represents its details by accumulating information at a set of scales $j \in \mathbb{Z}$.

Let $\{0, ..., J-1\}$ denotes a set of scales (resolution levels). We define $d_{J-1,k}$ and $c_{J-1,k}$ as follows [41]:

$$\begin{aligned} d_{J-1,k} &= w_{2k} - w_{2k-1}, \quad k = 1, ..., 2^{J-1}. \\ c_{J-1,k} &= w_{2k} + w_{2k-1}, \quad k = 1, ..., 2^{J-1}. \end{aligned} \quad (32)$$

The key idea is that a lower detail coefficient $d_{J-1,k}$ implies that w_{2k} is very close to w_{2k-1} and vice versa, as such a smoother function is represented by a small sequence of detail coefficients.

In order to represent the sequence in a coarser-scale (using a lower resolution), we define the coefficients:

$$\begin{aligned} d_{J-2,k} &= c_{J-1,2k} - c_{J-1,2k-1}, \quad k = 1, ..., 2^{J-2}. \\ c_{J-2,k} &= c_{J-1,2k} + c_{J-1,2k-1}, \quad k = 1, ..., 2^{J-2}. \end{aligned} \quad (33)$$

By repeating the procedure in (33) we obtain detailed and smoothed coefficients for lower resolutions. The multiscale algorithm stops when the $c_{0,1}$ coefficient is produced.

Next we discuss how to select optimally the thresholding (tuning) parameter in (25) for each resolution-level.

B. OPTIMAL THRESHOLDING BY A REFERENCE-FREE CRITERION FUNCTION

Since we deal with truncated distributions, the source (the complete non-truncated distribution) is intrinsically unobservable and thus, we cannot assess the success of denoising by comparing it to the original non-truncated distribution. Therefore, we utilize the "two-fold cross-validation" in [42] which is a reference-free criterion function assesing the quality of the function estimated by denoising[41]:

$$\lambda_{j,\gamma_j} = \arg\min_{\lambda_{j,\gamma_j}} \left\{\frac{1}{2}\left\|\widehat{f}^o_{\lambda_{j,\gamma_j}} - \boldsymbol{u}_j^e\right\|_2^2 + \frac{1}{2}\left\|\widehat{f}^e_{\lambda_{j,\gamma_j}} - \boldsymbol{u}_j^o\right\|_2^2\right\} \quad (34)$$

where λ_{j,γ_j} is the tuning (thresholding) parameter being used in (29), in which γ_j is a specific penalty function. The odd sample and even sample are denoted by $\boldsymbol{u}_j^o$ and $\boldsymbol{u}_j^e$, respectively, and their corresponding estimates are $\widehat{f}^o_{\lambda_{j,m}}$ and $\widehat{f}^e_{\lambda_{j,m}}$. These estimates are obtained by employing the iterative procedure in (29).

As previously discussed, our proposed truncation-proof IV estimator requires controlling for the bias terms $\mathcal{M}_1(\cdot)$ and $\mathcal{M}_2(\cdot)$. For generality and applicability purposes of the proposed estimator, we adopt a semiparametric approach which

[41] The methodology implemented in [42] chooses one threshold that is applicable to all resolution levels in the wavelet transform. In the present case, however, we select a threshold for each level in order to implement multi-resolution analysis increasing our proposed estimator's accuracy.

is not subjected to distributional assumptions and consequently, does not require specifying the functional form of these unknown functions.

The wavelet-based JPEG semiparametric estimator is chosen for its many advantages. It enables a multi-resolution representation of the noisy data points, implying that the data points are characterized both globally as well as locally.[42] Such a multi-resolution decomposition facilitates distinguishing between the noise and the systemic part. The systemic part is the functional relationship between the covariates and the dependent variables in the regression equations in (14).

An additional advantage of incorporating the aforementioned newly introduced JPEG estimator is that it accommodates for various data set forms of different types of irregularities, such as non-equispaced design that will be described in section IV-B to follow. These irregularities are alleviated by introducing the locations in space of the various data points as an additional covariate that is unique for each resolution-level. An additional merit of our approach is enabling a group-wise denoising rather than the traditional element-wise JPEG denoising in the cases of image processing. Group-wise denoising plays an important role in data denoising, as it takes into account potential dependencies among the various data points. Thus, our contribution to the JPEG algorithm are controlling for irregularities, group-wise thresholding on the entire data and utilizing a reference-free criterion function to choose the optimal thresholding.

In next section we describe the JPEG algorithm which is introduced to estimate each of the bias terms $\mathcal{M}_1(\cdot)$ and $\mathcal{M}_2(\cdot)$. Although our proposed denoising procedure can be applicable to both even as well as odd sample sizes (as will be demonstrated in Algorithm 1 to follow), for ease of presentation and without loss of generality, the denoising procedure is formulated as a function of a data set consisting of $2n$ observations.

IV. THE WAVELET-BASED JPEG DENOISING

Let $\{(u_i, t_i)\}_{i=1}^{2n}$ be a pairwise sequence of $2n$ data points as described in (22), such that $t_i < t_j \; \forall i < j$. The sequence $\{u_i\}_{i=1}^{2n}$ indicates the noisy data points (or colors of pixels in image processing) and their respective locations in space are represented by $\{t_i\}_{i=1}^{2n}$. The JPEG algorithm is a procedure generating a multi-resolution denoised representation of the sequence $\{u_i\}_{i=1}^{2n}$, which is denoted by the sequence $\{\widehat{u}_i\}_{i=1}^{2n}$. The purpose of the present section is three-fold: first, to describe the JPEG algorithm to be employed in order to obtain a noise-free representation of the noisy data; second, to extend the JPEG algorithm to be compatible with irregularities in the data[43]; and thirdly, to incorporate a reference-free criterion to evaluate the denoising procedure accuracy.

[42]Global representation is a weighed average (smoothing) of the data, while local representation consists of more detailed information regarding first differences between neighboring data points.

[43]We define $\Delta_i \equiv t_i - t_{i-1} \; \forall \; 2 \leq i \leq n$, such that equispaced (regular) data is a sequence of data points satisfying $\Delta_i = \Delta_j \; \forall i$ and j. Other cases are referred to as non-equispaced (irregular) spaced data.

Applying the conventional JPEG algorithm on a vector of data points is equivalent to employing three different procedures on the noisy data: (i) the JPEG forward transform $\mathcal{T}_F : \mathbb{R}^{2n\times 2} \to \mathbb{R}^{2n\times 1}$ to obtain the wavelet-based JPEG coefficients (as will be shown in (45) to follow); (ii) coefficients selection $\mathcal{T}_S : \mathbb{R}^{2n\times 2} \to \mathbb{R}^{2n\times 1}$ by applying a thresholding procedure (as will be shown in (49)) and (iii) the JPEG inverse transform $\mathcal{T}_I : \mathbb{R}^{2n\times 2} \to \mathbb{R}^{2n\times 1}$, which recovers the noise-free data by utilizing the selected coefficients (as will be shown in (46) to follow).

In the ensuing section we introduce auxiliary matrices to be used in each of the JPEG transforms, which are essential to construct the covariate matrix in the wavelet-based JPEG regression (in (48) to follow).

A. AUXILIARY MATRICES FOR THE JPEG ALOGRITHM

The implementation of the JPEG algorithm necessitates the construction of $\mathcal{T}_F$ and $\mathcal{T}_I$ operators. For this purpose, we construct auxiliary matrices $\mathcal{A}_{2n}$, $\mathcal{S}_{2n}$, $\left\{\mathcal{H}_{2n,\ell}^{(t)}\right\}_{\ell=1}^{4}$ which are the shifting, rescaling and smoothing operator matrices, respectively, each of size $2n \times 2n$.

Let $\mathcal{S}_{2n}$ and $\mathcal{S}_{2n}^{-1}$ be the rescaling and inverse-rescaling matrices, respectively each of size $2n \times 2n$. Its elements are defined for $m = 0, ..., n$ as:

$$(\mathcal{S}_{2n})_{i,j} = \begin{cases} 1/\varphi & \text{if } i = j = 2m \\ \varphi & \text{if } i = j = 2m+1 \\ 0 & \text{if } i \neq j, \end{cases}$$
$$(\mathcal{S}_{2n}^{-1})_{i,j} = \begin{cases} \varphi & \text{if } i = j = 2m \\ 1/\varphi & \text{if } i = j = 2m+1 \\ 0 & \text{if } i \neq j \end{cases} \tag{35}$$

The rescaling operator $\tilde{\boldsymbol{\nu}} = \mathcal{S}_{2n}\boldsymbol{\nu}$ takes a vector $\boldsymbol{\nu}$ of size $2n \times 1$ and return a rescaled vector $\tilde{\boldsymbol{\nu}}$ of the same size, such that even and odd elements of the original vector are multiplied by the scalars $1/\varphi$ and φ, respectively.

Let $\mathcal{A}_{2n}$ be a shifting operator matrix of size $2n \times 2n$, its elements are defined for $m = 0, ..., n$ as:

$$(\mathcal{A}_{2n})_{i,j} = \begin{cases} 1 & \text{if } (i > n, \;\; j = 2m) \text{ or } (i \leq n, \;\; j = 2m+1) \\ 0 & \text{otherwise.} \end{cases} \tag{36}$$

The $\tilde{\boldsymbol{\nu}} = \mathcal{A}_{2n}\boldsymbol{\nu}$ operator takes a vector $\boldsymbol{\nu} = [v_1, ..., v_{2n}]^T$ of size $2n \times 1$ and return the vector $\tilde{\boldsymbol{\nu}} = [\boldsymbol{\nu}_{\text{odd}}^T, \boldsymbol{\nu}_{\text{even}}^T]^T$. The vectors $\boldsymbol{\nu}_{\text{odd}} = [v_1, ..., v_{2n-3}, v_{2n-1}]^T$ and $\boldsymbol{\nu}_{\text{even}} = [v_2, ..., v_{2n-2}, v_{2n}]^T$ consist of the odd and even elements of $\boldsymbol{\nu}$, respectively.

Unlike the conventional JPEG, we allow for data irregularities by controlling for the data set location in space. For doing so, we denote a sequence of matrices $\left\{\mathcal{H}_{2n,\ell}^{(t)}\right\}_{\ell=1}^{4}$, such that the elements of matrix $\mathcal{H}_{2n,\ell}^{(t)} \; \forall \ell \in \{1, 2, 3, 4\}$ of size $2n \times 2n$

are defined for $m = 1, ..., n$ as:

$$(\mathcal{H}^{(t)}_{\substack{2n,\ell\\ \ell\in\{1,3\}}})_{i,j} = \begin{cases} 2\pi_\ell \omega^{(t)}_{\ell,i-1} & \text{if } i{=}2m \text{ and } j{=}i{-}1 \\ 1 & \text{if } j{=}i \\ 2\pi_\ell(1-\omega^{(t)}_{\ell,i-1}) & \text{if } i{=}2m \text{ and } j{=}i{+}1 \\ 0 & \text{otherwise} \end{cases} \tag{37}$$

$$(\mathcal{H}^{(t)}_{\substack{2n,\ell\\ \ell\in\{2,4\}}})_{i,j} = \begin{cases} 2\pi_\ell \omega^{(t)}_{\ell,i-1} & \text{if } i{=}2m{+}1 \text{ and } j{=}i{-}1 \\ 1 & \text{if } j{=}i \\ 2\pi_\ell(1-\omega^{(t)}_{\ell,i-1}) & \text{if } i{=}2m{+}1 \text{ and } j{=}i{+}1 \\ 0 & \text{otherwise .} \end{cases} \tag{38}$$

where each of sequences $\left\{\omega^{(t)}_{\ell,l}\right\}$ $\forall \ell \in \{1, 2, 3, 4\}$ are the interpolation weights (defined in (39) to follow) to control for the location in space of data points (enabling irregular non-equispaced data to be used) and $\pi_1, \pi_2, \pi_3, \pi_4$ are scalar constants described in [43], which are referred to as the filter coefficients of the wavelet-based JPEG. In the special case in which $\omega^{(t)}_{\ell,l} = 0.5\ \forall l$ and $\ell \in \{1, 2, 3, 4\}$ the algorithm is reduced to the regular-spaced wavelet-based JPEG.

We define the linear interpolation weights:

$$\omega^{(t)}_{\ell,2i} = \begin{cases} \dfrac{\tilde{t}_{2i+1} - \tilde{t}_{2i}}{\tilde{t}_{2i+1} - \tilde{t}_{2i-1}} & \text{if } \ell \in \{1, 3\} \\ \dfrac{\tilde{t}_{2i+2} - \tilde{t}_{2i+1}}{\tilde{t}_{2i+2} - \tilde{t}_{2i}} & \text{if } \ell \in \{2, 4\} \end{cases} \tag{39}$$

$$\tilde{t}_l = \begin{cases} t_2 & \text{if } l = 2m, l < 2 \\ t_l & \text{if } 1 \le l \le 2n \\ t_{2n-1} & \text{if } l = 2m + 1,\ l > 2n - 1 \end{cases} \tag{40}$$

For tractability, we formulate the JPEG coefficients estimation problem as a linear regression estimation, which necessitates obtaining a closed-form expressions of the JPEG forward and inverse transforms. These closed-form expression are required to characterize the JPEG covariate matrix to be used in the wavelet-based JPEG regression. In the following section we express analytically each of the forward and inverse transforms using matrix notation as a function of the auxiliary matrices presented above.

B. THE VARIOUS JPEG TRANSFORMS IN MATRIX NOTATION

Employing our proposed JPEG algorithm on a data set involves representation of data set in multiple resolution levels, a property which referred to as a multi-resolution analysis. Let J be the highest resolution level, which requires the same number of data points as in the noisy data set. The data set representation in j'th resolution-level $\forall j < J$ is a transformation of the data set representation in the finer (higher) resolution-level $j+1$. Consequently, the JPEG noise-free representation $\forall j < J$ can be formulated recursively. However, the implication of this formulation is that the location in space of the data points in any given resolution-level is also determined recursively. This fact stems from depicting the noisy data set $\boldsymbol{u} = [u_1, ..., u_{2n}]^T$ and its location in space $\boldsymbol{t} = [t_1, ..., t_{2n}]^T$ as a pairwise sequence. For ease of notation we construct the adjusted space location operator $\forall j < J$:

$$\begin{aligned} \boldsymbol{t}_j &\equiv \left(\prod_{h=j+1}^{J} \tilde{\mathcal{A}}_h\right)\boldsymbol{t}, \\ \tilde{\mathcal{A}}_j &= \begin{bmatrix} \mathcal{A}_{\mathfrak{m}(j)\times\mathfrak{m}(j)} & \mathbf{0}_{\mathfrak{m}(j)\times(\mathfrak{m}(J)-\mathfrak{m}(j))} \\ \mathbf{0}_{(\mathfrak{m}(J)-\mathfrak{m}(j))\times\mathfrak{m}(j)} & \boldsymbol{I}_{(\mathfrak{m}(J)-\mathfrak{m}(j))\times(\mathfrak{m}(J)-\mathfrak{m}(j))} \end{bmatrix} \end{aligned} \tag{41}$$

where $\mathfrak{m}(j) \equiv \lceil 2n/2^{J-j} \rceil$, $J = \lceil \log_2(2n) \rceil$ is the number of resolution-levels and j is a specific resolution level.[44]

This recursive formulation takes the noisy data points locations in space as control variables, which are essential for alleviating irregularities in the noisy data.

Using the adjusted space location sequence $\{\boldsymbol{t}_j\}$ in (41), we define matrix $\boldsymbol{\Psi}^{(t)}_{\boldsymbol{F}}$ (to be used in (45) to follow) for $J \in \{1, ..., \lceil \log_2(2n) \rceil\}$ resolution levels as:

$$\boldsymbol{\Psi}^{(t)}_{\boldsymbol{F}} \equiv \left(\prod_{j=1}^{J-1} \tilde{\boldsymbol{\Phi}}^{(\boldsymbol{t}_j)}_{j}\right) \boldsymbol{\Phi}^{(t)}_{\mathfrak{m}(J)\times\mathfrak{m}(J)} \tag{42}$$

$$\tilde{\boldsymbol{\Phi}}^{(t)}_{j} = \begin{bmatrix} \boldsymbol{\Phi}^{(t)}_{\mathfrak{m}(j)\times\mathfrak{m}(j)} & \mathbf{0}_{\mathfrak{m}(j)\times(\mathfrak{m}(J)-\mathfrak{m}(j))} \\ \mathbf{0}_{(\mathfrak{m}(J)-\mathfrak{m}(j))\times\mathfrak{m}(j)} & \boldsymbol{I}_{(\mathfrak{m}(J)-\mathfrak{m}(j))\times(\mathfrak{m}(J)-\mathfrak{m}(j))} \end{bmatrix} \tag{43}$$

where $\boldsymbol{\Phi}^{(t)}_{m\times m} \equiv \mathcal{A}_m \mathcal{S}_m \mathcal{H}^{(t)}_{m,4} \mathcal{H}^{(t)}_{m,3} \mathcal{H}^{(t)}_{m,2} \mathcal{H}^{(t)}_{m,1}$ and $\boldsymbol{I}_{m\times m}$ is the identity matrix of size $m \times m$. It worth noticing that $\boldsymbol{\Phi}^{(t)}_{m\times m}$ is a product of invertible matrices and consequently, its inverse is characterized as:

$$\begin{aligned} \left(\boldsymbol{\Phi}^{(t)}_{m\times m}\right)^{-1} &= \left(\mathcal{H}^{(t)}_{m,1}\right)^{-1} \left(\mathcal{H}^{(t)}_{m,2}\right)^{-1} \left(\mathcal{H}^{(t)}_{m,3}\right)^{-1} \\ &\quad \times \left(\mathcal{H}^{(t)}_{m,4}\right)^{-1} \mathcal{S}^{-1}_m \mathcal{A}^{-1}_m \end{aligned} \tag{44}$$

The multi-resolution JPEG forward transform operator $\mathcal{T}_F(\boldsymbol{u}, \boldsymbol{t})$ is the linear transform:

$$\boldsymbol{\delta} = \boldsymbol{\Psi}^{(t)}_{\boldsymbol{F}} \boldsymbol{u}, \quad \mathcal{T}_F(\boldsymbol{u}, \boldsymbol{t}) = \boldsymbol{\Psi}^{(t)}_{\boldsymbol{F}} \boldsymbol{u} \tag{45}$$

where $\boldsymbol{u}, \boldsymbol{t}$ and $\boldsymbol{\delta}$ are the noisy data, the location in space and the JPEG coefficient vectors, respectively, each of size $2n \times 1$.

Similarly, the multi-resolution JPEG inverse transform operator $\mathcal{T}_I(\boldsymbol{\delta}, \boldsymbol{t})$ is the linear transform:

$$\boldsymbol{u} = \boldsymbol{\Psi}^{(t)}_{\boldsymbol{I}} \boldsymbol{\delta}, \quad \mathcal{T}_I(\boldsymbol{\delta}, \boldsymbol{t}) = \boldsymbol{\Psi}^{(t)}_{\boldsymbol{I}} \boldsymbol{\delta} \tag{46}$$

where $\boldsymbol{u}, \boldsymbol{t}$ and $\boldsymbol{\delta}$ are the noisy data, the location in space and the JPEG coefficient vectors, respectively, each of size $2n \times 1$.

Matrix $\boldsymbol{\Psi}^{(t)}_{\boldsymbol{I}}$ in (46) is defined for $J \in \{1, ..., \lceil \log_2(2n) \rceil\}$ resolution levels as:

$$\boldsymbol{\Psi}^{(t)}_{\boldsymbol{I}} \equiv \left(\boldsymbol{\Phi}^{(t)}_{\mathfrak{m}(J)\times\mathfrak{m}(J)}\right)^{-1} \left(\prod_{j=J-1}^{1} \left(\tilde{\boldsymbol{\Phi}}^{(\boldsymbol{t}_j)}_{j}\right)^{-1}\right) \tag{47}$$

where $\boldsymbol{\Psi}^{(t)}_{\boldsymbol{I}}$ is the analytic inverse transform operator.

We introduce the wavelet-based JPEG nonparametric regression given a non-equispaced irregular data set:

$$u(t) = \left(\boldsymbol{\Psi}^{(t)}_{\boldsymbol{I}} \boldsymbol{\delta}\right)_t = \sum_{k\in\mathbb{Z}} c_{0,k} \phi_{0,k}(t) + \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} d_{j,k} \varphi_{j,k}(t) \tag{48}$$

[44]The operator's notation $\lceil \cdot \rceil$ represents the ceiling of a real number.

where $\boldsymbol{\delta}$ consists of the sequences of coefficients $\{c_{0,k}\}$ and $\{d_{j,k}\}$, which capture the function's average behavior and details, respectively. The matrix $\boldsymbol{\Psi}_I^{(t)}$ consists of the sequences of covariates $\{\phi_{0,k}(t)\}$ and $\{\varphi_{j,k}(t)\}$. Unlike the present case which employs a group-wise denoising on the entire data, in the conventional JPEG the coefficients selection operator, $\mathcal{T}_S(\boldsymbol{\delta}, \boldsymbol{t})$, is constructed to be used element-wise (for each resolution-level j), e.g, $\mathcal{T}_S(\delta_{ij}, \boldsymbol{t}) = S_\alpha(\delta_{ij}; \lambda_{j,\gamma_j}, \gamma_j)$ using $S_\alpha(\cdot)$ in (26), where λ_{j,γ_j} and γ_j are defined in (34) using $\alpha = 1$, such that $\{\delta_{ij}\}_{i=1}^{\mathfrak{m}(j)}$ is a subset of vector $\boldsymbol{\delta}$ consisting of the j'th resolution-level coefficients.

In the following section, we discuss about the JPEG group-wise coefficients selection to perform denoising.

1) JPEG GROUP-WISE COEFFICIENTS SELECTION FOR IRREGULAR DATA DENOISING

Lastly, we formulate the transpose of the inverse wavelet transform, in order to employ the group-wise denoising procedure depicted in (29):

$$\tilde{\boldsymbol{u}} = \left(\boldsymbol{\Psi}_I^{(t)}\right)^T \boldsymbol{u}, \quad \left(\boldsymbol{\Psi}_I^{(t)}\right)^T \equiv \left(\prod_{j=1}^{J-1}\left[\left(\tilde{\boldsymbol{\Phi}}_j^{(t_j)}\right)^{-1}\right]\right)^T \times \left[\left(\boldsymbol{\Phi}_{\mathfrak{m}(J)\times\mathfrak{m}(J)}^{(t)}\right)^{-1}\right]^T \tag{49}$$

where $\left[\left(\boldsymbol{\Phi}_{m\times m}^{(t)}\right)^{-1}\right]^T$ is constructed as:

$$\left[\left(\boldsymbol{\Phi}_{m\times m}^{(t)}\right)^{-1}\right]^T = \left[\mathcal{A}_m^{-1}\right]^T\left[\mathcal{S}_m^{-1}\right]^T\left[\left(\mathcal{H}_{m,4}^{(t)}\right)^{-1}\right]^T \times \left[\left(\mathcal{H}_{m,3}^{(t)}\right)^{-1}\right]^T\left[\left(\mathcal{H}_{m,2}^{(t)}\right)^{-1}\right]^T\left[\left(\mathcal{H}_{m,1}^{(t)}\right)^{-1}\right]^T \tag{50}$$

The JPEG estimator, denoted by $\widehat{\boldsymbol{\delta}}$, is obtained by minimizing the objective function in (28) using the iterative procedure in (29) given the chosen thresholding level. The latter is determined by minimizing the reference-free criterion function depicted in section III-B.

The construction of the wavelet-based JPEG transforms matrix involves computational complexity, a problem which can be alleviated by employing a faster algorithm, referred to as 'a lifting step'. In the succeeding section we discuss the algorithm to obtain a denoised representation of non equispaced data design by using a procedure that does not necessitate matrix operation to reduce computational complexity.

C. THE IRREGULAR FORWARD TRANSFORM

The irregular forward transform in (45) is the procedure to obtain the wavelet coefficients, $\boldsymbol{\delta}$, as follows:

The Filter function in algorithm 1 defined as follows:

Algorithm 1 The Wavelet-Based JPEG Forward Transform

procedure FORWARDTRANSFORM(u, Grid, Level)
Require: (i) Two real-number vectors u and Grid of size $n \times 1$; (ii) Level$\in \{1, ..., \log_2(n)\}$;
Ensure: Output $\leftarrow$ a real-number coefficient vector δ of size $n \times 1$;
    The JPEG filter coefficients:
    $\pi_1 \leftarrow -1.5861343420693648$;
    $\pi_2 \leftarrow -0.0529801185718856$;
    $\pi_3 \leftarrow \ \ 0.8829110755411875$;
    $\pi_4 \leftarrow \ \ 0.4435068520511142$
    $\varphi \ \leftarrow \ \ 1.1496043988602418$
    Start:
    $n \leftarrow$ length of u
    $m \leftarrow$ ceiling of $(n/2)$
    $Q \leftarrow$ ceiling of $(m/2)$
    $d \leftarrow$ copy the odd elements of u
    $s \leftarrow$ copy the even elements of u
    top:
    NewGrid $\leftarrow$ Grid
    The parameter vector ϕ_δ is used to generate a zero wavelet coefficient for the generated observation:
    $\phi_\delta[1] = -2 \times (\pi_1 \times \pi_2 \times \pi_3)/(1 + 2 \times \pi_2 \times \pi_3)$
    $\phi_\delta[2] = 2 \times (\pi_2 \times \pi_3)/(1 + 2 \times \pi_2 \times \pi_3)$
    $\phi_\delta[3] = 2 \times (\pi_1 + \pi_3 + 3 \times \pi_1 \times \pi_2 \times \pi_3)/(1 + 2 \times \pi_2 \times \pi_3)$
    if n is an odd number **then**
        $s[m] \leftarrow d[m-1] * \phi_\delta[1] + s[m-1] * \phi_\delta[2] + d[m] * \phi_\delta[3]$
        NewGrid[n+1] $\leftarrow$ NewGrid[n] + NewGrid[n]-NewGrid[n-1]
    end if
    OddGrid $\leftarrow$ copy the odd elements of NewGrid
    EvenGrid $\leftarrow$ copy the even elements of NewGrid
    $s \leftarrow s +$ Filter(d, 0, π_1, NewGrid)
    $d \leftarrow d +$ Filter(s, 1, π_2, NewGrid)
    $s \leftarrow s +$ Filter(d, 0, π_3, NewGrid)
    $d \leftarrow d +$ Filter(s, 1, π_4, NewGrid)
    u[1:m] $\leftarrow$ d[1:m]
    u[(m+1):n] $\leftarrow$ s[1:(n-m)]
    Grid[1:m] $\leftarrow$ OddGrid
    Grid[(m+1):n] $\leftarrow$ EvenGrid
    Scaling the wavelet coefficients:
    $d \leftarrow \delta[1 : Q] * \varphi$
    $s \leftarrow \delta[(Q+1) : m]/\varphi$
    if Level > 1 **then**
        $\begin{Bmatrix} \text{u[1:m]} \\ \text{Grid[1:m]} \end{Bmatrix} \leftarrow$ ForwardTransform(u[1:m], Grid[1:m], Level-1)
    end if
    return (u, Grid)
end procedure

D. THE IRREGULAR INVERSE TRANSFORM

The irregular inverse transform is the procedure to reconstruct the vector $\boldsymbol{u}$ in (46), as follows:

Algorithm 2 The Wavelet Forward and Inverse Filter

procedure FILTER(Series, Even, π, Grid)
Require: (i) Two real-number vectors: Series of size $n \times 1$ and Grid of size $2n \times 1$; (ii) two scalars: Even $\in \{0, 1\}$ and $\pi \in \mathbb{R}$.
Ensure: Output $\leftarrow$ Filter, a real-number vector of size $n \times 1$ consisting of the predicted series;
    n $\leftarrow$ length of Series
    O $\leftarrow$ copy the odd elements of Grid
    E $\leftarrow$ copy the even elements of Grid
    if Even $= 0$ **then**
        Low $\leftarrow$ O[1:n-1]
        High $\leftarrow$ O[2:n]
        weights $\leftarrow$ (High-E[1:(n-1)])/(High-Low)
        $\omega_l \leftarrow$ [weights, 0.5]
        $\omega_h \leftarrow 1 - \omega_l$
        S $\leftarrow$ [Series[1], Series]
        Filter $\leftarrow 2\pi\ \omega_l$*S[1:n] + $2\pi\ \omega_h$*S[2:(n+1)]
    else
        Low $\leftarrow$ E[1:(n-1)]
        High $\leftarrow$ E[2:n]
        weights $\leftarrow$ (High-O[2:n])/(High-Low)
        $\omega_l \leftarrow$ [0.5, weights]
        $\omega_h \leftarrow 1 - \omega_l$
        S $\leftarrow$ [Series, Series[n]]
        Filter $\leftarrow 2\pi\ \omega_l$*S[1:n] + $2\pi\ \omega_h$*S[2:(n+1)]
    end if
    return (Filter)
end procedure

E. THE IRREGULAR TRANSPOSE OF THE INVERSE TRANSFORM

The irregular transpose of the inverse transform in (49) enables to obtain the vector $\tilde{\boldsymbol{u}}$, as follows (see transposed-inverse filter function, TransInvFilter, in algorithm 6):

In the ensuing section we describe the estimation procedure of the parameter vector of interest, $\boldsymbol{\beta}$, using the wavelet-based JPEG estimate of $\widehat{\mathcal{M}}_1(\cdot)$ obtained from (29).

F. THE INSTRUMENT VARIABLE ESTIMATOR: TRUNCATED SAMPLE

Denote the truncated data by a sequence of observations $\{y_{1i}, \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{z}_i\}_{i=1}^{n}$, such that each observation is an independent realization of the conditional joint distribution function of the random variables $\{y_1, \mathbf{x}, \mathbf{w}, \mathbf{z}\}$ given that they are selected into the sample ($y_2 = 1$). The endogenous variable is denoted by x_1 and is included in vector $\mathbf{x}$. There are two types of joint dependence between the covariate vector and the substantive equation's random disturbance. The first type is intrinsic in the model and is generated by a variation in v (the endogenous part of x_1) leading to a comovement between x_1 and ξ_1. The second type is related to the sample selection and is generated by a variation in $\mathbf{w}$ leading to a comovement between the covariate vector $\mathbf{x}$ and ξ_1. This implies that there are two sources of endogeneity to be taken into consideration:

Algorithm 3 The Wavelet-Based JPEG Inverse Transform

procedure INVERSETRANSFORM(δ, Grid, Level)
Require: (i) Two real-number vectors δ and Grid of size $n \times 1$; (ii) Level$\in \{1, ..., \log_2(n)\}$;
Ensure: Output $\leftarrow$ a real-number vector u of size $n \times 1$;
    The JPEG filter coefficients:
    $\pi_1 \leftarrow -1.5861343420693648$; $\pi_2 \leftarrow -0.0529801185718856$;
    $\pi_3 \leftarrow 0.8829110755411875$; $\pi_4 \leftarrow 0.4435068520511142$
    $\varphi \leftarrow 1.1496043988602418$
    Start:
    $n \leftarrow$ length of δ
    $m \leftarrow$ ceiling of $n/2^{(\text{Level}-1)}$
    $Q \leftarrow$ ceiling of $m/2$
    Rescaling the wavelet coefficients:
    $d \leftarrow \delta[1 : Q]/\varphi$
    $s \leftarrow \delta[(Q + 1) : m] * \varphi$
    OldGrid $\leftarrow$ Grid
    for i=1 To m **do**
        index $\leftarrow (2(i-1)+1)*(i \le Q)+2(i-Q)*(i > Q)$
        Grid[index] $\leftarrow$ OldGrid[i]
    end for
    top:
    NewGrid $\leftarrow$ Grid[1:m]
    if m is an odd number **then**
        $s[Q] \leftarrow 0$
        NewGrid[m+1] $\leftarrow$ Grid[m] - Grid[m-1] + Grid[m]
    end if
    $d \leftarrow d -$ Filter(s, 1, π_4, NewGrid)
    $s \leftarrow s -$ Filter(d, 0, π_3, NewGrid)
    $d \leftarrow d -$ Filter(s, 1, π_2, NewGrid)
    $s \leftarrow s -$ Filter(d, 0, π_1, NewGrid)
    for i=1 To m **do**
        index $\leftarrow (2(i-1)+1)*(i \le Q)+2(i-Q)*(i > Q)$
        δ[index] $\leftarrow$ d[i]$*(i \le Q)+$s[i-Q]$*(i > Q)$
    end for
    u $\leftarrow \delta$
    if Level > 1 **then**
        $\left\{\begin{matrix}\text{u}\\ \text{Grid}\end{matrix}\right\} \leftarrow$ InverseTransform(δ, Grid, Level-1)
    end if
    return (u, Grid)
end procedure

the first source is related to the endogenous covariate, while the second source is due to the truncation environment of the data.

Next we discuss the two-step estimation procedure to be employed for the correction of both endogeneity and truncation bias propagated by truncation.

G. THE ESTIMATION PROCEDURE

In this section we introduce a two-step estimation procedure to eliminate the two sources of bias discussed. To eliminate

Algorithm 4 The Wavelet-Based JPEG Transposed-Inverse Transform

procedure TRANSINVERSETRANSFORM(u, Grid, Level)
Require: (i) Two real-number vectors u and Grid of size $m \times 1$; (ii) Level$\in \{1, ..., \log_2(m)\}$
Ensure: Output $\leftarrow \Psi_I^T u$
    The JPEG filter coefficients:
    $\pi_1 \leftarrow -1.5861343420693648$; $\pi_2 \leftarrow -0.0529801185718856$;
    $\pi_3 \leftarrow 0.8829110755411875$; $\pi_4 \leftarrow 0.4435068520511142$
    $\varphi \leftarrow 1.1496043988602418$
    Start:
    Output $\leftarrow u$
    $m \leftarrow$ length of u
    $Q \leftarrow$ ceiling of $(m/2)$
    $d \leftarrow$ copy the odd elements of u
    $s \leftarrow$ copy the even elements of u
    OddGrid $\leftarrow$ copy the odd elements of Grid
    EvenGrid $\leftarrow$ copy the even elements of Grid
    top:
    NewGrid $\leftarrow$ Grid
    if m is an odd number **then**
        $s[Q] \leftarrow 0$
        NewGrid[m+1] $\leftarrow$ NewGrid[m] + NewGrid[m]-NewGrid[m-1]
    end if
    $d \leftarrow d -$ TransInvFilter(s, 0, π_1, NewGrid)
    $s \leftarrow s -$ TransInvFilter(d, 1, π_2, NewGrid)
    $d \leftarrow d -$ TransInvFilter(s, 0, π_3, NewGrid)
    $s \leftarrow s -$ TransInvFilter(d, 1, π_4, NewGrid)
    Rescaling the wavelet coefficients:
    $d \leftarrow d[1 : Q]/\varphi$
    $s \leftarrow s[(Q + 1) : m] * \varphi$
    Output[1:Q] $\leftarrow$ d[1:Q]
    Output[(Q+1):m] $\leftarrow$ s[1:(m-Q)]
    Grid[1:Q] $\leftarrow$ OddGrid
    Grid[(Q+1):m] $\leftarrow$ EvenGrid
    if Level > 1 **then**
        Output[1:Q] $\leftarrow$ TransInverseTransform(Output[1:Q], Grid[1:Q], Level-1)
    end if
    return (Output, Grid)
end procedure

the endogeneity bias term we adapt a similar approach to the two step procedure in [44] for a partially linear single index model estimation, in which the first stage is a regression of the endogenous covariate on all the exogenous covariates and the instrumental variable. In the second stage, the endogenous covariate is substituted with the fitted values obtained from the first stage. However, the estimation approach in [44] cannot be implemented in a truncated environment, because it treats the first stage regression as a linear population regression (as if the entire covariates distribution function is observed). We alleviate this by modeling both the first as well as the second stage equations as endogenously truncated equations. In order to eliminate the endogenous truncation bias, we control for this source of bias by including the truncation bias term as an additional covariate in the substantive equations, as depicted in (14). Thus, the partial linearity is applied to both the first as well as the second stage equations.

In the first stage, we regress the endogenous covariate on the instrumental and exogenous variables, by minimizing the partially linear index model:

$$(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\theta}_{\mathbf{1}f}}) = \underset{(\boldsymbol{\delta},\boldsymbol{\theta}_{1f})\in\Theta\times\Delta_K}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \times \left(x_{1i} - \left[\boldsymbol{x}_{-1_i}^T, \boldsymbol{z}_i^T\right]\boldsymbol{\delta} - \widehat{\mathcal{M}}_2(\boldsymbol{w_i}; \boldsymbol{\theta_{1f}})\right)^2 \quad (51)$$

In the second stage, the endogenous variable is replaced by its predicted value obtained from the first stage in (51), and we minimize the following function:

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}_{\mathbf{2}f}}) = \underset{(\boldsymbol{\beta},\boldsymbol{\theta}_{2f})\in\Theta\times\Delta_K}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}\left(y_{1i} - \left[\widehat{x_{1i}}, \boldsymbol{x}_{-1_i}^T\right]\boldsymbol{\beta} - \widehat{\mathcal{M}}_1(\boldsymbol{w_i}; \boldsymbol{\theta_{2f}})\right)^2 \quad (52)$$

As can be seen in (52) the two sources of endogeneity bias we deal with are: (i) the bias propagated by the endogenous covariate is alleviated by utilizing the covariate set $\left[\widehat{x_{1i}}, \boldsymbol{x}_{-1_i}^T\right]$ consisting entirely of exogenous covariates and (ii) the bias propagated by the endogenous truncation is alleviated by controlling for the selection bias term $\widehat{\mathcal{M}}_1(\cdot)$.

Next we present Monte Carlo simulation to examine our semiparametric IV estimator's performance in a truncated environment.

V. SIMULATION

In this section, we generate multiple random data sets to be used for the examination of our model's performance, using different sample sizes.

First, we discuss the procedure for the data generation process (DGP).

A. DATA GENERATION PROCESS

Denote the sample size by $N \in \{500, 2000, 3000, 5000, 8000, 10000\}$. In order to not restrict the data generation process to the family of symmetric unimodal distribution functions, a mixture of distribution functions is utilized to generate each of the selection model's disturbances that are jointly dependent (as will be discussed in section V-A.1 to follow). In order to verify that our proposed model performs well under different data generating processes (DGP), we construct a data set consisting of 2,000,000 distribution functions,[45] practically generating 100 millions realizations which are not i.i.d. By construction, each observation is randomly drawn from a unique mixture of distribution functions.

[45]The estimates obtained given the various data distribution functions will be supplied upon request.

1) THE DISTURBANCES' JOINT DISTRIBUTION FUNCTION

Each triple of disturbances $\{\xi_{1i}, \xi_{2i}, \mathrm{v}_i\}$ is randomly and independently drawn from $F_{\xi_1,\xi_2,\mathrm{v}}$, which is the substantive and participation equations' disturbances joint distribution function. The aforementioned joint density function consists of two components: a Copula function,[46] which characterizes the disturbances' dependence structure, and three marginal distribution functions F_{ξ_1}, F_{ξ_2} and F_{v}. In order to verify our model's performance in the presence of random disturbances' distribution functions that are not restricted to the family of symmetric and unimodal distribution functions, each one of the sample selection model's disturbances ξ_1 and ξ_2 is marginally-distributed according to a mixture of three different distribution functions: (i) a normal distribution function with expectation and standard deviation parameters (μ, σ_a) denoted by $\mathcal{N}(\mu, \sigma_a^2)$; (ii) a normal distribution function with expectation and standard deviation parameters $(-\mu, \sigma_b)$ denoted by $\mathcal{N}(-\mu, \sigma_b^2)$; (iii) a gamma distribution function with scale and shape parameters $(\mu\varphi, \varphi)$ denoted by $\Gamma_{\text{Gamma}}(\mu\varphi, \varphi)$.[47] This mixture distribution function is defined as:

$$\begin{cases} \mathrm{v} \sim \mathcal{N}(0, \sigma_{\mathrm{v}}^2) \\ \xi_{\mathrm{j}} \sim 0.4\mathcal{N}(\mu, \sigma_a^2) + 0.5\mathcal{N}(-\mu, \sigma_b^2) \\ \qquad + 0.1\Gamma_{\text{Gamma}}(\mu\varphi, \varphi), \quad j = 1, 2. \end{cases} \tag{53}$$

where $\mathbb{E}\left[\xi_{\mathrm{j}}\right] = 0$ and $\mathbb{E}\left[\mathrm{v}\right] = 0$.

The parameters set $(\mu, \sigma_a, \sigma_b, \varphi, \sigma_{\mathrm{v}}) = (4, 2.5, 1.5, 2, 1)$ is arbitrarily chosen. Due to its simplicity, the Clayton Copula (as will be discussed in section V-A.2 to follow) with a degree of dependence parameter is set to equal 1, assuring a mild correlation between the disturbances, is used for controlling the dependence structure. Choosing a mild correlation, is important in order to be conservative by examining the potential bias in the parameter estimates under conditions which are not extreme.

Next we employ a function characterizing the dependence properties of the Copula [46], referred to as *a generator function* to construct the joint dependence of the random disturbances in (53).

2) ARCHIMEDEAN COPULA FUNCTION

An Archimedean Copula is a Copula characterized by a non-increasing, continuous generator function $\psi\colon [0, \infty] \to [0, 1]$, which satisfies $\psi(0) = 1$, $\psi(\infty) = 0$ and is strictly decreasing on $[0, \inf\{t : \psi(t) = 0\}]$. In particular, we are interested in the d dimensional Archemdean Copula family (3 in the present case[48]) which has the simple algebraic form [46][49]:

$$\mathcal{C}(u_1, .., u_d) = \psi(\psi^{-1}(u_1), ..., \psi^{-1}(u_d)), \quad (u_1, ..., u_d) \in [0, 1]^d \tag{54}$$

where ψ is a specific function known as the generator of $\mathcal{C}$. To generate the disturbances, the Clayton Copula's generator $\psi(t) = (1 + t)^{-1/\theta}$ is chosen.

The covariates vector of i'th observation is a realization of the random variables $[\mathrm{z}, \mathrm{x}_2, \mathrm{w}_1, \mathrm{w}_2]$ which are jointly distributed with their corresponding marginal distribution functions $\left\{\mathcal{F}^i_{\mathrm{z}}, \mathcal{F}^i_{\mathrm{x}_2}, \mathcal{F}^i_{\mathrm{w}_1}, \mathcal{F}^i_{\mathrm{w}_2}\right\}$. The dependence structure is modeled by utilizing a Gaussian Copula which is a convenient way to generate high dimensional data. By construction, each datum is generated by utilizing a different sequence of marginal distribution functions (constructed as a finite mixture drawn from a menu of 2, 000, 000 continuous distribution functions). These random variables expectation vector $\boldsymbol{\mu} = [0, 0, 0, 0]^T$ and a covariance matrix $\boldsymbol{\Sigma_{4\times 4}}$. The arbitrarily chosen covariance matrix is:

$$\begin{aligned} \boldsymbol{\Sigma_{4\times 4}} &= \begin{bmatrix} \sigma^2_{\mathrm{z}} & \sigma_{\mathrm{z},\mathrm{x}_2} & \sigma_{\mathrm{z},\mathrm{w}_1} & \sigma_{\mathrm{z},\mathrm{w}_2} \\ \sigma_{\mathrm{z},\mathrm{x}_2} & \sigma^2_{\mathrm{x}_2} & \sigma_{\mathrm{x}_2,\mathrm{w}_1} & \sigma_{\mathrm{x}_2,\mathrm{w}_2} \\ \sigma_{\mathrm{z},\mathrm{w}_1} & \sigma_{\mathrm{x}_2,\mathrm{w}_1} & \sigma^2_{\mathrm{w}_1} & \sigma_{\mathrm{w}_1,\mathrm{w}_2} \\ \sigma_{\mathrm{z},\mathrm{w}_2} & \sigma_{\mathrm{x}_2,\mathrm{w}_2} & \sigma_{\mathrm{w}_1,\mathrm{w}_2} & \sigma^2_{\mathrm{w}_2} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0.4 & 0.8 & -0.6 \\ 0.4 & 1.264 & 0.36 & -0.48 \\ 0.8 & 0.36 & 2 & -0.4 \\ -0.6 & -0.48 & -0.4 & 2 \end{bmatrix} \end{aligned}$$

We generate the data y_{1i}, y_{2i}, x_{1i} according to the following data generation process (DGP) [49]:

$$\text{DGP1}: \begin{cases} y^*_{1i} = \alpha_1 + \beta_1\, x_{1i} + \beta_2\, x_{2i} + \xi_{1\mathrm{i}} \\ y^*_{2i} = \alpha_2 + \gamma_1\, w_{1i} + \gamma_2\, w_{2i} + \xi_{2\mathrm{i}} \\ x^*_{1i} = \delta_1\, z_i + \delta_2\, x_{2i} + \mathrm{v}_{\mathrm{i}} \end{cases} \tag{55}$$

where each i element in the sequence $\{x_{2i}, z_i, w_{1i}, w_{2i}\}_{i=1}^N$ is an independent realization of the random variables $(\mathrm{x}_2, \mathrm{z}, \mathrm{w}_1, \mathrm{w}_2)$. We choose the parameter setting $[\alpha_1, \alpha_2, \beta_1, \beta_2, \delta_1, \delta_2, \gamma_1, \gamma_2] = [2, 0.5, 1, 1.25, 0.5, 1, 2, -1]$.

The truncated data set is characterized by the following equations:

$$\begin{bmatrix} y_{1i} \\ x_{1i} \end{bmatrix} = \begin{cases} \begin{bmatrix} \alpha_1 + \beta_1\, x_{1i} + \beta_2\, x_{2i} + \xi_{1\mathrm{i}} \\ \delta_1\, z_i + \delta_2\, x_{2i} + \mathrm{v}_{\mathrm{i}} \end{bmatrix} & \text{if } y^*_{2i} \geq 0 \\ \text{Unobserved} & \text{if } y^*_{2i} < 0, \end{cases} \tag{56}$$

where x_{1i} is an endogenous variable included in vector $\boldsymbol{x}_i \in \mathbb{R}^p$, in which all the elements (except for x_i) are exogenous variables and $\boldsymbol{\beta} \in \mathbb{R}^p$ is a covariates vector. The substantive equation's random disturbance is denoted by ξ_{1i}.

[46] Any continuous joint distribution function can be characterized by a set of marginal distribution functions and a joint distribution function determining the dependence structure which is referred to as a Copula function (Sklar's Theorem [45]).

[47] The scale and shape parameters imply that the expectation and standard deviation parameters are $(\mu, \sqrt{\mu/\varphi})$, respectively.

[48] $d = 3$ representing the three-dimensional vector of random disturbances $(\mathrm{v}_{\mathrm{i}}, \xi_{1i}, \xi_{2i})$.

[49] Knowing the distribution corresponding to a generator ψ, [47] presented a sampling algorithm for exchangeable Archimedean copulas which does not require the knowledge of the copula density. This algorithm is therefore applicable to large dimensions [48].

B. SIMULATIONS RESULT

We have randomly generated for each sample size $N \in \{500, 2000, 3000, 5000, 8000, 10000\}$, a total of 10, 000 data sets using the data generation process elaborated on in V-A. For a given number of observations N, different models are estimated: (i) an *OLS* estimator utilizing a sample consisting of random realizations from the complete distribution function, without correcting for the endogeneity of x_{1i} covariate; (ii) a conventional IV estimator, correcting for the endogeneity of x_{1i} covariate using the aforementioned entire distribution function; (iii) both *OLS* as well as a conventional IV estimators are applied to a truncated portion of the data distribution function consisting of participants only (without correcting for the self-selection bias); (iv) truncated sample model's estimates using the developed wavelet-based JPEG IV estimator, correcting for both truncation as well as endogeneity biases.[50]

Table 1 presents summary statistics of estimates for models (i) and (ii), while Table 2 presents summary statistics of estimates for models (iii) and (iv). In Table 3, different convergence measures of these estimates are presented.

TABLE 1. Monte Carlo Simulation - Non-truncated (complete) data set with (without) IV correction.

True Parameter[a]	Estimate	Model Setup: Sample size 500	2000	3000	5000	8000	10000
Full sample *OLS* estimates (**without IV correction**)							
$\beta_1 = 1$	Mean	2.6848	2.6798	2.6807	2.6810	2.6809	2.6805
	Median	2.6857	2.6815	2.6811	2.6812	2.6819	2.6812
	Std	0.1602	0.1138	0.0858	0.0605	0.0534	0.0507
$\beta_2 = 1.25$	Mean	-0.7009	-0.6953	-0.6988	-0.6983	-0.6963	-0.6976
	Median	-0.7085	-0.6983	-0.6967	-0.6994	-0.6968	-0.6976
	Std	0.2420	0.1743	0.1275	0.0838	0.0710	0.0652
Full sample *OLS* estimates (**with IV correction**)							
$\beta_1 = 1$	Mean	0.9826	0.9995	1.0025	1.0020	0.9986	0.9991
	Median	1.0050	1.0057	1.0083	0.9996	0.9997	0.9998
	Std	0.4397	0.3072	0.2174	0.1393	0.1116	0.1006
$\beta_2 = 1.25$	Mean	1.2687	1.2493	1.2457	1.2469	1.2510	1.2517
	Median	1.2365	1.2382	1.2438	1.2466	1.2497	1.2506
	Std	0.5406	0.3771	0.2655	0.1692	0.1366	0.1211

Note: [a] The parameters that are used in the data generation process.
We estimate by ordinary least squares (OLS) method the parameters for the full sample and truncated sample without correction for the selectivity bias. Then, we calculate for these estimates the mean, median and standard deviation (Std.) over all data sets. The standard deviations are obtained using the estimates from the Monte Carlo simulations.

Entries in Table 1 indicate that regardless of sample size, the means of the *OLS* estimates are biased, such that e.g., for a sample size of 3000 observations $\beta_1 = 2.6807$ and $\beta_2 = -0.6988$, while the mean of the full sample IV's estimates are $\beta_1 = 1.0025$ and $\beta_2 = 1.2457$ for the same sample size. For a sample size of 500 observations, the standard deviation obtained for β_1 (the endogenous covariate's coefficient) using the IV estimator is 2.73 times larger than in the *OLS* estimator and decreases from 0.4397 to 0.1006, when the sample size increases from 500 to 10, 000 observations.

In Table 2 to follow, the estimates obtained by using the truncated data are presented. It is evident that the mean of the truncated sample *OLS* estimates are $\beta_1 = 2.26$ and $\beta_2 = -0.3084$ for a sample size of 500 observations, whereas the true parameter values are $\beta_1 = 1$ and $\beta_2 = 1.25$, respectively. This is a huge bias which is hardly improved as

[50]For sake of brevity, we delegate results of the first stage to Appendix VI-B.

TABLE 2. Monte Carlo Simulation - Truncated data set with (without) truncation bias correction.

True Parameter[a]	Estimate	Model Setup: Sample size 500	2000	3000	5000	8000	10000
Truncated sample *OLS* estimates (without truncation bias correction)							
$\beta_1 = 1$	Mean	2.2600	2.2493	2.2504	2.2485	2.2488	2.2492
	Median	2.2711	2.2514	2.2534	2.2510	2.2484	2.2493
	Std	0.2676	0.1878	0.1362	0.0865	0.0723	0.0675
$\beta_2 = 1.25$	Mean	-0.3084	-0.2926	-0.2974	-0.2955	-0.2939	-0.2940
	Median	-0.3170	-0.2884	-0.3048	-0.2974	-0.2944	-0.2942
	Std	0.3693	0.2647	0.1909	0.1162	0.0940	0.0851
Truncated sample *OLS* estimates - IV correction (without truncation bias correction)							
$\beta_1 = 1$	Mean	0.1084	0.1805	0.1942	0.1910	0.1850	0.1847
	Median	0.1410	0.2095	0.2074	0.2003	0.1901	0.1876
	Std	0.7595	0.5288	0.3588	0.2235	0.1777	0.1634
$\beta_2 = 1.25$	Mean	2.0544	2.0148	1.9973	2.0018	2.0083	2.0088
	Median	2.0022	1.9798	1.9833	1.9968	2.0058	2.0082
	Std	0.8910	0.6219	0.4212	0.2601	0.2055	0.1878
Truncated sample **JPEG** *IV* estimates							
$\beta_1 = 1$	Mean	0.9455	0.9875	1.0058	1.0010	1.0004	1.0000
	Median	0.9717	1.0085	1.0079	1.0047	1.0013	1.0009
	Std	0.7325	0.3641	0.2972	0.2178	0.1732	0.1577
$\beta_2 = 1.25$	Mean	1.3174	1.2830	1.2617	1.2465	1.2506	1.2499
	Median	1.2945	1.2623	1.2579	1.2456	1.2518	1.2469
	Std	0.8145	0.4081	0.3324	0.2433	0.1933	0.1754

Note: [a] The parameters that are used in the data generation process.
We estimate the parameters for the truncated sample without correction for the selectivity bias by employing the OLS and the (conventional) IV methods. Additionally, we estimate the parameters using the same truncated sample by employing our JPEG IV estimator, correcting for both truncation as well as endogeneity bias. For brevity, we introduce here only the second stage results and delegate the first stage results to Table 4 Appendix B. In each of the estimation procedures (OLS, IV and JPEG IV), we calculate for these estimates the mean, median and standard deviation (Std.) over all data sets.

the sample size increases. Further, applying a conventional IV produces estimates which still represent a huge bias, particularly $\beta_1 = 0.1084$ and $\beta_2 = 2.0544$ for the same sample size (500 observations).

Entries in Table 2 indicate that regardless of sample size, the means of the truncated sample IV's estimates are biased (ranges from one-tenth to one-fifth of the estimate that would have emerged by employing the conventional IV method in the absence of truncation).[51] Note that estimates' accuracy hardly improved as sample size increases. This is due to the presence of two sources of bias. The mean estimate of β_1 (the endogenous covariate's parameter) which is obtained from implementing our proposed methodology, basically mimics the results obtained using a random sample from the entire data distribution function for sample sizes, above 2, 000 observations. The standard deviations of this estimate for sample sizes of 500 and 10, 000 observations are 0.7325 and 0.1577, respectively. For a sample size of 5, 000 observations (or above), the mean estimate of β_2 (the exogenous covariate's parameter) approximates the estimate obtained by employing the conventional IV, using a random sample from the entire data distribution function. However, the estimate of β_2 obtained by employing the conventional IV in a truncated sample is biased even for 10, 000 observations.

We conduct sensitivity test to measure the influence of an increase in number of observations on the accuracy of the truncated sample's parameter estimates.

The first accuracy measure we use is the standardized root mean square error, $RMSE_j$, measuring the bias in the truncated regression estimate relative to the true parameter value that would have been obtained in an non truncated

[51]For sake of brevity, we have omitted the estimates of the nuisance parameters which can be furnished upon request as well as the parameter estimates of the first stage, which are delegated to Table 4 Appendix VI-B.

distribution, defined as:

$$RMSE_j\left(\Omega\right) = \left(\frac{1}{\Omega}\sum_{i=1}^{\Omega}\left(\frac{\hat{\beta}_{i,j}^{s} - \beta_j^s}{\beta_j^s}\right)^2\right)^{1/2}, \qquad (57)$$

where $\hat{\beta}_{i,j}^{s}$ and β_j^s stand for the substantive (s) equation's j'th coefficient estimated in the i'th sample and the coefficient in the theoretical model that would have been obtained in the entire population, respectively. Ω is the number of data sets generated for the Monte Carlo simulations, which is 5000 data sets (each one consists of N observations).

Another measure is based on a formula similar to the one described in (57), and is intended to find the relative accuracy of the truncated sample's estimates, in comparison to full sample estimates, defined as:

$$R_j\left(\Omega\right) = \left(\frac{1}{\Omega}\sum_{i=1}^{\Omega}\left(\frac{\hat{\beta}_{i,j}^{ts} - \hat{\beta}_{i,j}^{s}}{\hat{\beta}_{i,j}^{s}}\right)^2\right)^{1/2}, \qquad (58)$$

where $\hat{\beta}_{i,j}^{ts}$ and $\hat{\beta}_{i,j}^{s}$ stand for the substantive (s) equation's j'th coefficient estimated using the truncated (t) sample and the full sample, respectively. This measure evaluates the relative model's performance in the truncated sample, with respect to the conventional IV using the full sample.

The last estimates' accuracy measure is the δ coefficient used for the calculation of the estimators' standard deviations convergence rate n^{δ} with respect to the sample size. It depicts the speed of standard deviation's shrinkage resulting from increasing the sample size. This coefficient is calculated based on the following ratio:

$$\delta = \frac{\ln\left(\sigma_1/\sigma_2\right)}{\ln\left(n_2/n_1\right)}, \qquad (59)$$

where σ_1 and σ_2 are the estimate's standard deviations that are calculated for data sets with n_1 and n_2 number observations, respectively (calculated for a given estimate).

Table 3 entries indicate that the root mean squares error (RMSE) measure of the estimates obtained by employing the conventional IV estimator, using a random sample from the entire data distribution function, gets smaller as the sample size increases, as can be expected. However, applying the same procedure to the truncated data set leads to RMSE measures, which are in the range of 2 to 8-fold larger, given a sample size of 2, 000 to 10, 0000 observations, respectively. This is indeed a huge bias generated by the conventional IV, which is not immune to truncation bias. Additionally, the RMSE measures show negligible improvements as a function of number of observations for the conventional IV, whereas there is a huge improvement of the RMSE, as a function of the number of observations for the JPEG IV estimator provided by our model. The proximity between the JPEG IV and the full sample IV estimates increases with the sample size, as reflected by the R_j proximity measure. Using the same measure, we find that there is a much smaller improvement in the proximity between the truncated sample conventional IV and the full sample IV, relative to the improvement in the proximity between the JPEG IV and the full sample IV estimates.

TABLE 3. Monte Carlo Simulation - Convergence measures.

Parameter	Model's estimates					
	Number of observations					
	500	2000	3000	5000	8000	10000
RMSE measure						
Full sample OLS estimates - IV correction						
β_1	0.4395	0.3069	0.2172	0.1393	0.1115	0.1005
β_2	0.4321	0.3013	0.2123	0.1353	0.1092	0.0969
Truncated sample OLS estimates - IV correction						
β_1	1.1702	0.9749	0.8818	0.8392	0.8341	0.8315
β_2	0.9590	0.7882	0.6861	0.6363	0.6284	0.6253
Truncated sample **JPEG** *IV* estimates						
β_1	0.7328	0.3639	0.2968	0.2177	0.1731	0.1575
β_2	0.6524	0.3271	0.2658	0.1946	0.1546	0.1402
R_j (n) measure - relative to full sample IV						
Truncated sample OLS estimates - IV correction						
β_1	5.5322	1.0441	0.9006	0.8434	0.8354	0.8328
β_2	1.5804	0.7607	0.6797	0.6382	0.6266	0.6217
Truncated sample **JPEG** *IV* estimates						
β_1	4.5652	0.5544	0.3176	0.2015	0.1588	0.1422
β_2	1.4772	0.4244	0.2675	0.1721	0.1349	0.1210
δ consistency measure ($n^{\delta} \equiv$ the convergence rate)						
Truncated sample OLS estimates - IV correction						
β_1	-	0.7627	0.3538	0.2949	0.2254	0.1793
Truncated sample **JPEG** *IV* estimates						
β_1	-	0.5040	0.5017	0.6072	0.4875	0.4201

Note: The parameters that are used in the data generation process.
We examine three different measures for the parameters presented in Table (3). First, the standardized root mean square error $RMSE$ between each model's estimates and the true parameters is calculated based on equation (57). Second, the R_j measure is calculated based on equation (58). Third, the convergence rate which is measured n^{δ} is calculated for both the conventional IV as well as the JPEG IV estimates. This convergence rate measure implies that multiplying the sample size by 2 shrinks the estimators' standard deviations by 2^{δ}.

It is evident that JPEG IV is a $\sqrt{n}$ consistent estimator, as depicted by the δ consistency measure, which is about 0.5, implying that multiplying the sample size by 2 shrinks the estimators' standard deviations by $2^{\delta} = \sqrt{2}$. It is also evident that the truncated data conventional IV is poorly functioning in terms of consistency, as is shown by entries in Table 3.

VI. CONCLUSION

We provide an analytical proof showing that in an endogenously truncated data the conventional IV estimator does not perform the task it was intended to, but rather introduces an additional unintended bias into the parameter estimates of the substantive equation. The instrumental variable is endogenous by itself in the context of endogenously truncated data due to a comovement between the instrumental variable and the substantive equation's random disturbance, generated by mediating covariates. We offer a truncation-proof JPEG IV, shown to be a proper estimator under endogenous truncation. Employing Monte Carlo simulations attests to the JPEG IV estimator's high accuracy and its $\sqrt{n}$ consistency. These results have been verified by utilizing 2,000,000 different distribution functions (not restricted to the unimodal symmetric family), generating 100 million realizations to construct the covariates' data sets which are not imposed to be i.i.d. The various distribution functions attest to a very high accuracy of the model as depicted by the parameter estimates that closely mimic the true parameters.

TABLE 4. Monte Carlo Simulation - Truncated data set with truncation bias correction.

True Parameter[a]	Estimate	Model Setup: Sample size					
		500	2000	3000	5000	8000	10000
Truncated sample **JPEG** *IV* estimates (First stage)							
$\delta_1 = 0.5$	Mean	0.4944	0.4975	0.4973	0.4996	0.5000	0.4999
	Median	0.4951	0.4963	0.4971	0.4999	0.4998	0.5000
	Std	0.1236	0.0675	0.0514	0.0461	0.0352	0.0298
$\delta_2 = 1$	Mean	0.9976	0.9983	0.9994	1.0001	1.0005	0.9999
	Median	0.9977	0.9989	0.9989	1.0001	1.0006	1.0003
	Std	0.0902	0.0444	0.0360	0.0279	0.0221	0.0197

Note: [a] The parameters that are used in the data generation process.
We estimate the parameters using the truncated sample by employing our JPEG IV estimator, correcting for both truncation as well as endogeneity bias. We calculate for these estimates the mean, median and standard deviation (Std.) over all data sets. The standard deviations are obtained using the estimates from the Monte Carlo simulations.

APPENDIX

A. AN ELEMENT-WISE THRESHOLDING

In this appendix we show that for any $\gamma \in (1, \infty)$, $S_\alpha(\cdot)$ in (26) is the solution to the min-max concave penalty function in (24) $\forall \alpha \in (1/\gamma, \infty)$.

Proof: Let $\tilde{\boldsymbol{\delta}} \equiv \boldsymbol{\delta}^{(iter)} - 1/(\alpha n)\boldsymbol{\Psi}_I^T(\boldsymbol{u} - \boldsymbol{\Psi}_I \boldsymbol{\delta}^{(iter)})$

$$\boldsymbol{\delta} = \arg\min_{\boldsymbol{\delta}} \frac{1}{2}\left\|\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}\right\|_2^2 + \frac{1}{\alpha} P_{\lambda_j,\gamma_j}(\boldsymbol{\delta}) \tag{60}$$

As $P_{\lambda_j,\gamma_j}(\boldsymbol{\delta})$ is a separable function of $\boldsymbol{\delta}$, the minimization problem can be implemented in an element-wise fashion. Let δ_k be the solution to the following univariate regularized least squares problem:

$$\delta_k = \begin{cases} \arg\min_{\delta} \frac{1}{2}\left\|\delta - \tilde{\delta}_k\right\|_2^2 + \lambda|\delta| - \frac{\delta^2}{2\gamma} & \text{if } |\delta_k| \le \lambda\gamma \\ \arg\min_{\delta} \frac{1}{2}\left\|\delta - \tilde{\delta}_k\right\|_2^2 + \frac{1}{2}\lambda^2\gamma & \text{if } |\delta_k| > \lambda\gamma \end{cases} \tag{61}$$

We obtain the first order condition of (61) with respect to δ:

$$\delta_k = \begin{cases} \frac{1}{1 - 1/(\alpha\gamma)}\left[\tilde{\delta}_k - \text{sign}(\delta_k)\frac{\lambda}{\alpha}\right] & \text{if } |\delta_k| \le \lambda\gamma \\ \tilde{\delta}_k & \text{if } |\delta_k| > \lambda\gamma \end{cases} \tag{62}$$

Note that if $\left|\tilde{\delta}_k\right| > \lambda\gamma$ and $|\delta_k| \le \lambda\gamma$ it implies that either $\text{sign}(\delta_k) = -1$ and $-\lambda\gamma \le \tilde{\delta}_k \le \lambda\gamma - 2\lambda/\alpha$ or $\text{sign}(\delta_k) = 1$ and $-\lambda\gamma + 2\lambda/\alpha \le \tilde{\delta}_k \le \lambda\gamma$. Both cases contradict the fact that $\left|\tilde{\delta}_k\right| > \lambda\gamma$. Similarly, if $\left|\tilde{\delta}_k\right| \le \lambda\gamma$ and $|\delta_k| > \lambda\gamma$, it contradicts the fact that $\delta_k = \tilde{\delta}_k$ which follows directly from (62). Consequently, $\left|\tilde{\delta}_k\right| > \lambda\gamma \Leftrightarrow |\delta_k| > \lambda\gamma$. We get:

$$\delta_k = \begin{cases} \frac{\tilde{\delta}_k - \frac{\lambda}{\alpha}}{1 - 1/(\alpha\gamma)} & \text{if } 0 < \frac{\tilde{\delta}_k - \frac{\lambda}{\alpha}}{1 - 1/(\alpha\gamma)} \le \lambda\gamma \\ \frac{\tilde{\delta}_k + \frac{\lambda}{\alpha}}{1 - 1/(\alpha\gamma)} & \text{if } -\lambda\gamma \le \frac{\tilde{\delta}_k + \frac{\lambda}{\alpha}}{1 - 1/(\alpha\gamma)} < 0 \end{cases} \tag{63}$$

and provided that $\forall \alpha \in (1/\gamma, \infty)$ (63) is simplified to:

$$\delta_k = \begin{cases} \frac{\tilde{\delta}_k - \frac{\lambda}{\alpha}}{1 - 1/(\alpha\gamma)} & \text{if } \frac{\lambda}{\alpha} < \tilde{\delta}_k \le \lambda\gamma \\ \frac{\tilde{\delta}_k + \frac{\lambda}{\alpha}}{1 - 1/(\alpha\gamma)} & \text{if } -\lambda\gamma \le \tilde{\delta}_k < -\frac{\lambda}{\alpha} \end{cases} \tag{64}$$

Algorithm 5 The Wavelet-Based JPEG Penalized Regression

procedure PENALIZEDLINEARREGRESSION(u, Grid, Level, λ, γ)
Require: (i) Two real-number vectors u and Grid of size $n \times 1$; (ii) Level$\in \{1, ..., \log_2(n)\}$;
(iii) $\lambda \in (0, \infty)$ and $\gamma \in (1, \infty)$.
Ensure: Output $\leftarrow$ a real-number coefficient vector $\widehat{\delta}$ of size $n \times 1$;
  $n \leftarrow$ length of u
  top:
  $\delta_{\text{old}} \leftarrow$ ForwardTransform(u, Grid, Level-1)
  $\widehat{u}_{\text{old}} \leftarrow$ u
  $\text{residual}_{\text{old}} \leftarrow$ u - $\widehat{u}_{\text{old}}$
  $\text{Obj}_{\text{old}} \leftarrow \frac{1}{2n}\sum_{i=1}^{n} \text{residual}_{\text{old}}^2[\text{i}] + \sum_{i=1}^{n}$ MCP.penalty($|\delta_{\text{old}}[i]|$,λ, α, γ))
  $\text{v}_{\text{old}} \leftarrow$ TransInverseTransform($\text{residual}_{\text{old}}$, Grid, Level-1)
  flag $\leftarrow$ TRUE
  gap $\leftarrow$ infinite
  tolerance $\leftarrow 10^{-9}$
  maxiter $\leftarrow 1000$
  $\eta \leftarrow 1.2$
  while flag=TRUE and gap $>$ tolerance **do**
    $\alpha \leftarrow 1$
    Iter $\leftarrow 1$
    repeat
      Update $\leftarrow \delta_{\text{old}} + \frac{1}{\alpha n}\text{v}_{\text{old}}$
      $\delta_{\text{new}} \leftarrow S_\alpha$(Update, λ, α, γ)
      $\widehat{u}_{\text{new}} \leftarrow$ InverseTransform(δ_{new}, Grid, Level-1)
      $\text{residual}_{\text{new}} \leftarrow$ u - $\widehat{u}_{\text{new}}$
      $\text{Obj}_{\text{new}} \leftarrow \frac{1}{2n}\sum_{i=1}^{n} \text{residual}_{\text{new}}^2[\text{i}] + \sum_{i=1}^{n}$ MCP.penalty($|\delta_{\text{new}}[i]|$,λ, α, γ))
      $\alpha \leftarrow \alpha\eta$
      flag $\leftarrow \text{Obj}_{\text{new}} < \text{Obj}_{\text{old}}$
    until flag=TRUE or *iter* $>$ *maxiter*
    if flag = TRUE **then**
      $\delta_{\text{old}} \leftarrow \delta_{\text{new}}$
      $\text{Obj}_{\text{old}} \leftarrow \text{Obj}_{\text{new}}$
      $\text{residual}_{\text{old}} \leftarrow \text{residual}_{\text{new}}$
      $\text{v}_{\text{old}} \leftarrow$ TransInverseTransform($\text{residual}_{\text{new}}$, Grid, Level-1)
    end if
    gap $\leftarrow \left\|\delta_{\text{new}} - \delta_{\text{old}}\right\|_2^2 / \left\|\delta_{\text{old}}\right\|_2^2$
    Iter $\leftarrow$ Iter + 1
  end while
  return (δ_{old})
end procedure

After some algebraic manipulation we obtain:

$$\delta_k = \begin{cases} \frac{1}{1 - 1/(\alpha\gamma)}\text{sign}(\tilde{\delta}_k)\left(\left|\tilde{\delta}_k\right| - \lambda/\alpha\right) & \text{if } \lambda/\alpha < \left|\tilde{\delta}_k\right| \le \lambda\gamma \\ \tilde{\delta}_k & \text{if } \left|\tilde{\delta}_k\right| > \lambda\gamma \end{cases} \tag{65}$$

■

Algorithm 6 The Transposed-Inverse filter

procedure TRANSINVFILTER(Series, Even, π, Grid)
Require: (i) Two real-number vectors: Series of size $n \times 1$ and Grid of size $2n \times 1$; (ii) two scalars: Even $\in \{0, 1\}$ and $\pi \in \mathbb{R}$.
Ensure: Output $\leftarrow$ Filter, a real-number vector of size $n \times 1$ consisting of the predicted series;
    n $\leftarrow$ length of Series
    O $\leftarrow$ copy the odd elements of Grid
    E $\leftarrow$ copy the even elements of Grid
    if Even $= 1$ **then**
        Low $\leftarrow$ O[1:n-1]
        High $\leftarrow$ O[2:n]
        weights $\leftarrow$ (High-E[1:(n-1)])/(High-Low)
        $\varpi_l \leftarrow$ [0, 1-weights]
        $\varpi_h \leftarrow \leftarrow$ [weights, 1]
        S $\leftarrow$ [Series[1], Series]
        Filter $\leftarrow 2\pi\ \varpi_l \leftarrow$*S[1:n] $+ 2\pi\ \varpi_h \leftarrow$*S[2:(n+1)]
    else
        Low $\leftarrow$ E[1:(n-1)]
        High $\leftarrow$ E[2:n]
        weights $\leftarrow$ (High-O[2:n])/(High-Low)
        $\varpi_l \leftarrow$ [1, 1-weights]
        $\varpi_h \leftarrow$ [weights, 0]
        S $\leftarrow$ [Series, Series[n]]
        Filter $\leftarrow 2\pi\ \varpi_l$*S[1:n] $+ 2\pi\ \varpi_h$*S[2:(n+1)]
    end if
    return (Filter)
end procedure

Next we show how to obtain an analytic representation of the transpose of the wavelet inverse transform. For doing so, we based our arguments on the equivalence between the matrices-product and the lifting scheme representations of the wavelet transform. Then using this equivalence, we generate a filter to render the costly matrices building useless. Our objective is to reduce computational complexity.

Next we present the first stage estimates.

B. FIRST STAGE ESTIMATES

See Table 4 and Algorithm 5.

C. ALGORITHMS

See Algorithm 6.

ACKNOWLEDGMENT

The authors would like to thank Boaz Nadler, Yaniv Tenzer and seminar participants of the faculty of Mathematics and Computer Science, The Weizmann Institute of Science and Statistic departments of Tel-Aviv and Haifa universities, for very constructive comments.

REFERENCES

[1] J. D. Angrist and G. M. Kuersteiner, "Semiparametric causality tests using the policy propensity score," Nat. Bur. Econ. Res., Cambridge, MA, USA, Tech. Rep. w10975, 2004.
[2] N. Billfeld and M. Kim, "Semiparametric correction for endogenous truncation bias with vox populi-based participation decision," *IEEE Access*, vol. 7, pp. 12114–12132, 2019.
[3] J. J. Heckman, "Sample selection bias as a specification error," *Econometrica*, vol. 47, no. 1, pp. 153–161, 1979.
[4] W. K. Newey, "Two-step series estimation of sample selection models," *Econ. J.*, vol. 12, no. s1, pp. S217–S229, 2009.
[5] J. L. Powell, "Semiparametric estimation of censored selection models," in *Proc. 13th Int. Symp. Econ. Theory Econ., Essays Honor Takeshi Amemiya Nonlinear Stat. Modeling*, vol. 165. Cambridge, U.K.: Cambridge Univ. Press, 2001, p. 96.
[6] H. Xie, L. E. Pierce, and F. T. Ulaby, "SAR speckle reduction using wavelet denoising and Markov random field modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 10, pp. 2196–2212, Oct. 2002.
[7] S. Voronin, D. Mikesell, and G. Nolet, "Compression approaches for the regularized solutions of linear systems from large-scale inverse problems," *GEM-Int. J. Geomath.*, vol. 6, no. 2, pp. 251–294, 2015.
[8] P. Hall and B. A. Turlach, "Interpolation methods for nonlinear wavelet regression with irregularly spaced design," *Ann. Statist.*, vol. 25, no. 5, pp. 1912–1925, 1997.
[9] S. Sardy, D. B. Percival, A. G. Bruce, H.-Y. Gao, and W. Stuetzle, "Wavelet shrinkage for unequally spaced data," *Statist. Comput.*, vol. 9, no. 1, pp. 65–75, 1999.
[10] M. Carnec, P. Le Callet, and D. Barba, "Full reference and reduced reference metrics for image quality assessment," in *Proc. IEEE 7th Int. Symp. Signal Process. Appl.*, vol. 1, Jul. 2003, pp. 477–480.
[11] J. Farah, M.-R. Hojeij, J. Chrabieh, and F. Dufaux, "Full-reference and reduced-reference quality metrics based on SIFT," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 161–165.
[12] A. Cohen, I. Daubechies, and J.-C. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Commun. Pure Appl. Math.*, vol. 45, no. 5, pp. 485–560, 1992.
[13] H. Ichimura, "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models," *J. Econ.*, vol. 58, no. 1, pp. 71–120, 1993.
[14] A. Lewbel and O. Linton, "Nonparametric matching and efficient estimators of homothetically separable functions," *Econometrica*, vol. 75, no. 4, pp. 1209–1227, 2007.
[15] B. E. Usevitch, "A tutorial on modern lossy wavelet image compression: Foundations of JPEG 2000," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 22–35, Sep. 2001.
[16] P. M. Robinson, "Root-N-consistent semiparametric regression," *Econ., J. Econ. Soc.*, vol. 56, no. 4, pp. 931–954, 1988.
[17] H. Ichimura and L. F. Lee, "Semiparametric least squares estimation of multiple index models: Single equation estimation," in *Proc. 5th Int. Symp. Econ. Theory Econ. Nonparametric Semiparametric Methods Econ. Statist.*, Cambridge, U.K., 1991, pp. 3–49.
[18] A. Lewbel and S. M. Schennach, "A simple ordered data estimator for inverse density weighted expectations," *J. Econ.*, vol. 136, no. 1, pp. 189–211, 2007.
[19] D. Williams, *Probability with Martingales*. Cambridge, U.K.: Cambridge Univ. Press, 1991.
[20] A. Haar, "Zur theorie der orthogonalen funktionensysteme," *Math. Ann.*, vol. 69, no. 3, pp. 331–371, Sep. 1910.
[21] V. Delouille, M. Jansen, and R. von Sachs, "Second-generation wavelet denoising methods for irregularly spaced data in two dimensions," *Signal Process.*, vol. 86, no. 7, pp. 1435–1450, 2006.
[22] E. Vanraes, M. Jansen, and A. Bultheel, "Stabilised wavelet transforms for non-equispaced data smoothing," *Signal Process.*, vol. 82, no. 12, pp. 1979–1990, 2002.
[23] B. W. Silverman, "Wavelets in statistics: Beyond the standard assumptions," *Philos. Trans. Roy. Soc. London A, Math., Phys. Eng. Sci.*, vol. 357, no. 1760, pp. 2459–2473, 1999.
[24] I. Daubechies and W. Sweldens, "Factoring wavelet transforms into lifting steps," *J. Fourier Anal. Appl.*, vol. 4, no. 3, pp. 247–269, 1998.
[25] I. M. Johnstone and B. W. Silverman, "Empirical Bayes selection of wavelet thresholds," *Ann. Statist.*, vol. 33, no. 4, pp. 1700–1752, 2005.
[26] I. M. Johnstone and B. W. Silverman, "Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences," *Ann. Statist.*, vol. 32, no. 4, pp. 1594–1649, 2004.
[27] V. Delouille, J. Simoens, and R. von Sachs, "Smooth design-adapted wavelets for nonparametric stochastic regression," *J. Amer. Stat. Assoc.*, vol. 99, no. 467, pp. 643–658, 2004.

[28] I. Daubechies, *Ten Lectures on Wavelets* (CBMS-NSF Regional Conference Series in Applied Mathematics), vol. 61. Philadelphia, PA, USA: SIAM, 1992.
[29] I.-L. Chern, "Interpolating wavelets and difference wavelets," in *Proc. Joint Austral.-Taiwanese Workshop Anal. Appl.*, 1999, pp. 133–147.
[30] D. Wei, J. Tian, R. O. Wells, Jr., and C. S. Burrus, "A new class of biorthogonal wavelet systems for image transform coding," *IEEE Trans. Image Process.*, vol. 7, no. 7, pp. 1000–1013, Jul. 1998.
[31] R. A. Zalik, "Riesz bases and multiresolution analyses," *Appl. Comput. Harmon. Anal.*, vol. 7, no. 3, pp. 315–331, 1999.
[32] V. Dicken and P. Maaß, "Wavelet–Galerkin methods for ill-posed problems," *J. Inverse Ill-Posed Problems*, vol. 4, no. 3, pp. 203–222, 1996.
[33] F. U. Abramovich and B. Silverman, "Wavelet decomposition approaches to statistical inverse problems," *Biometrika*, vol. 85, no. 1, pp. 115–129, 1998.
[34] J. L. Horowitz, "Adaptive nonparametric instrumental variables estimation: Empirical choice of the regularization parameter," *J. Econ.*, vol. 180, no. 2, pp. 158–173, 2014.
[35] X. He *et al.*, Eds., *Computer, Informatics, Cybernetics and Applications: Proceedings of the CICA 2011*, vol. 107. Springer, 2011.
[36] D. Tomassi, D. Milone, and J. D. B. Nelson, "Wavelet shrinkage using adaptive structured sparsity constraints," *Signal Process.*, vol. 106, pp. 73–87, Jan. 2015.
[37] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, vol. 38, no. 2, pp. 894–942, 2010.
[38] P. Breheny and J. Huang, "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection," *Ann. Appl. Statist.*, vol. 5, no. 1, p. 232, 2011.
[39] Z. Yang, Z. Wang, H. Liu, Y. C. Eldar, and T. Zhang, "Sparse nonlinear regression: Parameter estimation and asymptotic inference." 2015, *arXiv:1511.04514*. [Online]. Available: https://arxiv.org/abs/1511.04514
[40] D. L. Donoho and J. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.
[41] G. Nason, *Wavelet Methods in Statistics With R*. Springer, 2010.
[42] G. P. Nason, "Wavelet shrinkage using cross-validation," *J. Roy. Stat. Soc. B, Methodol.*, vol. 58, no. 2, pp. 463–479, 1996.
[43] P. Schelkens, A. Skodras, and T. Ebrahimi, *The JPEG 2000 Suite*, vol. 15. Hoboken, NJ, USA: Wiley, 2009.
[44] Y. Zhou, Y. Yang, J. Han, and P. Zhao, "Estimation for partially linear single-index instrumental variables models," *Commun. Statist.-Simul. Comput.*, vol. 45, no. 10, pp. 3629–3642, 2016.
[45] A. Sklar, *Fonctions de Répartition án Dimensions et Leurs Marges*. Paris, France: Univ. Paris, 1959.
[46] A. J. McNeil and J. Nešlehová, "Multivariate Archimedean copulas, d-monotone functions and ℓ_1-norm symmetric distributions," *Ann. Statist.*, vol. 37, no. 5B, pp. 3059–3097, 2009.
[47] A. W. Marshall and I. Olkin, "Families of multivariate distributions," *J. Amer. Stat. Assoc.*, vol. 83, no. 403, pp. 834–841, 1988.
[48] M. Hofert, "Sampling archimedean copulas," *Comput. Statist. Data Anal.*, vol. 52, no. 12, pp. 5163–5174, 2008.
[49] J. C. Escanciano, "A simple and robust estimator for linear regression models with strictly exogenous instruments," *Econ. J.*, vol. 21, no. 1, pp. 36–54, 2017.

NIR BILLFELD received the B.A. degree in economics and statistics from the University of Haifa, Israel, in 2006, the M.A. degree in economics from Tel-Aviv University, in 2010, and the Ph.D. degree from the University of Haifa, in 2019. His work has appeared in the IEEE, where he is currently a Researcher.

MOSHE KIM received the Ph.D. degree from the University of Toronto. He is currently a Professor of economics with the University of Haifa, Israel. He is the Founder and former Director of the Barcelona Banking Summer School, Universitat Pompeu Fabra, Barcelona, the former Director of the endowed chair of banking with the Humboldt University of Berlin, a Senior Distinguished Fellow at the Swedish School of Economics, Helsinki (Hanken), the Institute Research Professor with the German Institute for Economic Research (DIW), a Consultant at the Central Bank of Norway, and recently visited NYU Shanghai. His research interests are econometrics, banking, financial markets, and industrial organization. His books include the *Microeconometrics of Banking: Methods, Applications, and Results* (Oxford University Press, 2009). His work has appeared in the IEEE, the *Journal of Finance*, the *Journal of Monetary Economics*, the *Journal of Financial Intermediation*, the *Journal of Business and Economics Statistics*, the *Journal of Money Credit and Banking*, the *International Economic Review*, the *Journal of Accounting and Economics*, the *Journal of Public Economics*, the *Journal of Urban Economics*, the *Journal of Law and Economics*, the *Journal of Banking and Finance*, the *Journal of Industrial Economics*, and the *International Journal of Industrial Organization*. He was recently declared high-end foreign expert by the Chinese Foreign Ministry, and is a recent recipient of the Outstanding Tutor Award from the Chinese Ministry of Education.