

Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models

Susanne M. Schennach*
Department of Economics
University of Chicago
1126 East 59th Street
Chicago IL 60637
smschenn@uchicago.edu

First draft: October 2003, This version: December 2005

Abstract

This paper establishes that instruments enable the identification of nonparametric regression models in the presence of measurement error by providing a closed form solution for the regression function in terms of Fourier transforms of conditional expectations of observable variables. For parametrically specified regression functions, we propose a root n consistent and asymptotically normal estimator taking the familiar form of a generalized method of moment estimator with a plugged-in nonparametric kernel density estimate. Both the identification and the estimation methodologies rely on Fourier analysis and on the theory of generalized functions. The finite-sample properties of the estimator are investigated through Monte Carlo simulations.

Keywords: errors-in-variables model, Fourier transform, generalized function, semiparametric model.

*This work was made possible in part through financial support from the National Science Foundation via grant SES-0452089. The author would like to thank Jeremy Fox, Ricardo Mayer, Derek Neal and Xiaohong Chen, as well as participants at seminars given at the Universities of Rochester, Chicago, Maryland, Michigan, UCSD and UC-Riverside, the 2004 summer meetings of the Econometric Society and the CIRANO/CIREQ “Operator Methods in Microeconometrics, Time Series and Finance” conference for their helpful comments. Three anonymous referees and a co-editor provided helpful suggestions for a greatly improved presentation.

1 Introduction

Estimators based on instrumental variables (IV) have long been used to estimate linear regressions models of the form $Y = \theta X + \varepsilon$ where Y is the dependent variable, θ is the parameter of interest and where the error term ε is potentially correlated with the explanatory variable X . This correlation between ε and X could arise either from endogeneity or measurement error in the regressors. Indeed, if the observed regressor X and the unobserved true regressor X^* are related through $X = X^* + \Delta X$, where ΔX is a zero mean measurement error that is uncorrelated with X^* , the true model $Y = \theta X^* + \Delta Y$ is related to the observed Model $Y = \theta X + \varepsilon$ by

$$Y = \theta X^* + \Delta Y = \theta X - \theta \Delta X + \Delta Y = \theta X + \varepsilon \quad (1)$$

where the disturbance term $\varepsilon = -\theta \Delta X + \Delta Y$ is correlated with X , which prompts the need for IV estimation. For a nonlinear specification, $g(x^*, \theta) \equiv E[Y|X^* = x^*]$, IV estimation admits a straightforward extension when the correlation between ε and X is due to endogeneity, but not when it is due to measurement error. As noted by Amemiya (1985), the simple additive separation between the observed regressor and the measurement error illustrated in Equation (1) is no longer possible. The same issue also invalidates the use of recent nonparametric instrumental variable methods (e.g., Darolles, Florens, and Renault (2002), Newey and Powell (2003)) in the presence of measurement error. This problem has prompted a long search for a solution. A wide variety of measurement error bias-reduction approaches in nonlinear models with classical measurement error have been proposed (among many others, see Hsiao (1989), Chesher (1991), Lewbel (1996), Chesher (1998), Lewbel (1998), Hsiao and Wang (2000), Wang (2002), the review by Carroll, Ruppert, and Stefanski (1995) and numerous methods based on validation data). While the problem of identifying and estimating nonlinear errors-in-variables models when repeated measurements are available has been studied extensively (Hausman, Newey, Ichimura, and Powell (1991), Hausman, Newey, and Powell (1995), Li (2002), Schennach (2004a)), the

present paper uses more widely available instrumental variables, since instruments have arguably been the most common way to overcome measurement error problems for linear models in empirical work.

In the special case of polynomial specifications, Hausman, Newey, Ichimura, and Powell (1991) have provided a proof of identification using instruments and a corresponding asymptotically normal and root n consistent estimator that requires no distributional assumptions regarding the model's variables. Subsequently, Newey (2001) has shown that with distributional assumptions, root n consistent and asymptotically normal estimation is possible for general functional forms and that, without distributional assumptions, consistent estimation is possible. However, Newey assumes identification of the model and does not attempt to establish it in terms of primitive assumptions.

Wang and Hsiao (1995) provide a root n consistent semiparametric estimator under the assumption that the regression function $g(x^*, \theta)$ is absolutely integrable (that is, $\int_{-\infty}^{\infty} |g(x^*, \theta)| dx^*$ is finite). Unfortunately, most specifications used in empirical econometrics are not absolutely integrable, including the widely used logistic function and any model containing a polynomial, even reduced to a simple constant term. Wang and Hsiao (1995) can also show identification for models having at most $N_x + 1$ parameters, where N_x is the dimension of the mismeasured regressor. Since a linear specification already requires $N_x + 1$ parameters, the number of parameters identifiable via Wang and Hsiao's approach will be insufficient for many nonlinear applications.

While substantial progress has also been made to handle nonlinear or nonparametric models with nonseparable endogenous errors, existing methods are unable to handle endogeneity of the specific form associated with the presence of mismeasured regressors (as discussed in Schennach (2005)), either because the model does not take the form of a triangular system (as required in Chesher (2003) and Imbens and Newey

(2003)) or because the disturbances in the reduced form equations are necessarily correlated with the instruments (thus violating the assumptions made in Chernozhukov and Hansen (2005) and Chernozhukov, Imbens, and Newey (2006)).

Despite these contributions, a general proof that instruments enable the identification of nonlinear or nonparametric specifications with measurement error in the regressor has so far remained elusive, and existing estimators of nonlinear models exhibit important limitations. The present paper fills these gaps in the challenging case where the true regressor is continuously distributed.

First, we show nonparametric identification of the regression function in the absence of distributional assumptions by deriving a closed form solution for the regression function of interest in terms of the Fourier transforms of various conditional expectations involving observable variables. For estimation purposes, we consider the case where the regression function of interest and the relationship between the regressor and the instrumental variables are parametrically specified, while still avoiding any distributional assumptions. We devise a root n consistent and asymptotically normal estimator that takes the form of a generalized method of moment estimator with a plugged-in nonparametric kernel density estimate. Our approach thus provides, for the first time, a general nonlinear extension of the instrumental variable treatment of the linear errors-in-variables model found in most econometric textbooks. It enables a measurement error-robust treatment of common nonlinear models such as tobit, logit, probit, polynomials, piecewise-linear models or splines. Multivariate, nonparametric and quantile extensions are also discussed. The finite-sample properties of the proposed estimator are investigated through Monte Carlo simulations. An example of an application to the estimation of the black-white wage gap is also presented in the Supplementary Material available at the *Econometrica* web site.

2 Identification

We consider the model:

$$\begin{aligned}
 Y &= g(X^*) + \Delta Y & E[\Delta Y|W, \Delta X^*] &= 0 \\
 X &= X^* + \Delta X & E[\Delta X|W, \Delta X^*, \Delta Y] &= 0 \\
 X^* &= m(W) + \Delta X^* & \Delta X^* \text{ independent from } W \text{ and }^1 E[\Delta X^*] &= 0
 \end{aligned}
 \tag{2}$$

where, $g(\cdot)$ is the function to be determined, while $m(\cdot)$ is an unknown function of a random vector W of instruments. The variables X, Y, W are observable, while the variables $X^*, \Delta X, \Delta Y, \Delta X^*$ are not. Lowercase letters denote particular values of the corresponding uppercase random variable. For simplicity of exposition, we consider $Y, X, X^*, \Delta Y, \Delta X$ and ΔX^* to be scalar random variables, although a multivariate extension will be discussed in Section 5. The assumptions made are the same as in Newey (2001), except that we consider a nonparametric $g(x^*)$ instead of a parametric $g(x^*, \theta)$. They allow for the presence of conditional heteroskedasticity in the disturbances ΔY and ΔX , but not in ΔX^* . Nonparametric identification will require that at least one of the components of W be continuously distributed (which implies that X^* also has a continuous distribution). All other variables (including the measurement error ΔX) can be either discrete or continuous.

We will now derive a closed-form expression for $g(x^*)$ in terms of the observed variables, as summarized in Theorem 1 later in this section. Since

$$X = X^* + \Delta X = m(W) + \Delta X^* + \Delta X \tag{3}$$

where $E[\Delta X^* + \Delta X|W] = 0$, the function $m(w)$ can be determined from a standard nonparametric least-squares projection of X on W (both of which are observable) and is therefore identified. Hence, for the purpose of establishing identification, we define the observed scalar random variable

$$Z = m(W). \tag{4}$$

¹The assumption that that $E[\Delta X^*] = 0$ results in no loss of generality since this can always be achieved by allowing for a constant shift in the function $m(w)$.

Model (2) can then be rewritten as

$$\begin{aligned} Y &= g(X^*) + \Delta Y & E[\Delta Y|Z, U] &= 0 \\ X &= X^* + \Delta X & E[\Delta X|Z, U, \Delta Y] &= 0 \\ X^* &= Z - U. & U \text{ independent from } Z \text{ and } E[U] &= 0 \end{aligned} \tag{5}$$

where, for convenience, we have set $U = -\Delta X^*$. Note that we require full independence between U and Z instead of the more common mean independence. While a formal statistical test of the validity of this assumption is not possible, since X^* is not observed, the observable data nevertheless enables a test of the arguably stricter restriction that the residuals $(\Delta X - U)$ of the regression of X on W are independent from Z . This test can be considered more stringent because, even if it failed, it could be the result of a dependence between ΔX and Z , which would not violate the assumptions of our estimation procedure.²

Newey (2001) suggests that the function $g(x^*)$ may be identified from the knowledge of the conditional expectations $E[Y|Z = z]$ and $E[XY|Z = z]$ through the following two equalities implied by the assumptions of Model (5):

$$E[Y|Z = z] = \int g(z - u) dF(u) \tag{6}$$

$$E[XY|Z = z] = \int (z - u) g(z - u) dF(u) \tag{7}$$

where $F(u)$ denotes the cdf of U and where the integrals extend over the whole real line.³ The heuristic argument supporting this suggestion is the fact that this model is characterized by two unknown functions $g(x^*)$ and $F(u)$ and we have two functional equations available. However, a formal and general proof of identification of this model has so far been missing and existing treatments of special parametric cases provide ambiguous indications regarding whether identification is possible in a general class of models: A polynomial $g(x^*)$ is identified (Hausman, Newey, Ichimura,

²Of course, a dependence between U and ΔX could fortuitously yield a $(\Delta X - U)$ that is independent of Z even if U is dependent on Z , but this appears highly unlikely.

³In their proof of identification, Wang and Hsiao (1995) integrate these equations over all z , thereby reducing this very rich set of nonparametric functional restrictions to only two scalar equations in the case of a scalar X^* . As a result, they can only show identification of up to 2 parameters.

and Powell (1991)) from these equations, while the case of an exponential $g(x^*)$ has been shown not to be identified (Schemm (2004b)).

Our approach relies on the fact that Fourier transforms convert integral equations such as (6) and (7), which are difficult to solve directly, into much simpler algebraic equations, thanks to the Convolution and the Moment Theorems, as summarized in the following Lemma and as proven in the Appendix.⁴

Assumption 1 $|g(x^*)|$, $|E[Y|Z = z]|$ and $|E[XY|Z = z]|$ are defined and bounded by polynomials for $x^*, z \in \mathbb{R}$.

Lemma 1 Under Assumption 1, Equations (6) and (7) are equivalent to

$$\varepsilon_y(\zeta) = \gamma(\zeta) \phi(\zeta) \tag{8}$$

$$\mathbf{i}\varepsilon_{xy}(\zeta) = \dot{\gamma}(\zeta) \phi(\zeta) \tag{9}$$

where $\mathbf{i} = \sqrt{-1}$, where dots denote derivatives (e.g. $\dot{\gamma}(\zeta) \equiv d\gamma(\zeta)/d\zeta$) and where

$$\begin{aligned} \varepsilon_y(\zeta) &= \int E[Y|Z = z] e^{\mathbf{i}\zeta z} dz & \gamma(\zeta) &= \int g(x^*) e^{\mathbf{i}\zeta x^*} dx^* \\ \varepsilon_{xy}(\zeta) &= \int E[XY|Z = z] e^{\mathbf{i}\zeta z} dz & \phi(\zeta) &= \int e^{\mathbf{i}\zeta u} dF(u). \end{aligned} \tag{10}$$

Given the simplicity of this result, it would be tempting to simply manipulate Equations (8) and (9) to eliminate the infinite-dimensional nuisance parameter $\phi(\zeta)$ and get a single differential equation in $\gamma(\zeta)$ (the quantity of interest) in terms of the observable functions $\varepsilon_y(\zeta)$ and $\varepsilon_{xy}(\zeta)$. However, these standard algebraic manipulations can break down, because the Fourier transforms of functions that are not necessarily absolutely integrable, such as $g(x^*)$, $E[Y|Z = z]$, and $E[XY|Z = z]$, are typically not functions in the usual sense, but more abstract and singular objects known as *generalized functions* or as *tempered distributions* (see, for instance, Lighthill (1962), Gel'fand and Shilov (1964), Schwartz (1966), or the summary in the

⁴Although Wang and Hsiao (1995) employ Fourier transforms, they did not make use of the fact that Equation (7) also admits a simple representation in terms of Fourier transforms, which is crucial to establish a general identification result.

Supplementary Material). The most widely known example of generalized function is Dirac’s delta “function”, denoted $\delta(\cdot)$, which can be viewed as the limit of a sequence of normal densities of shrinking width. The delta function is the Fourier transform of a constant function. Delta function “derivatives” of any finite order, denoted $\delta^{(k)}(\cdot)$, can also be defined through k differentiations of a sequence defining a delta function. The Fourier transform of a polynomial, for instance, yields a linear combination of delta function derivatives. Assumption 1 limits the rate at which various functions diverge as their argument goes to infinity, in order to ensure that the Fourier transforms we consider are well-defined tempered distributions and can therefore always be decomposed as a sum of an ordinary function and a linear combination of delta function derivatives of some finite order. In the sequel we will distinguish these “ordinary” and “singular” components by subscripts “ o ” and “ s ”, respectively. For instance, we can write $\gamma(\zeta) \equiv \gamma_o(\zeta) + \gamma_s(\zeta)$.

In manipulating Equations (8) and (9), it will be important to keep in mind that two generalized functions cannot be multiplied or divided by one another. However, multiplication of a generalized function with an ordinary function, such as $\phi(\zeta)$, is allowed. $\phi(\zeta)$ is an ordinary function, because it is the Fourier transform of a probability measure (also called a characteristic function) and can be shown to be a bounded and continuous function (Loève (1977)), using the fact that probability measures are absolutely integrable.

Wang and Hsiao (1995) avoid dealing with generalized functions by assuming that $g(x^*, \theta)$ is absolutely integrable (which implies that $E[Y|Z = z]$ is as well). Most specifications used in applied work are not absolutely integrable, and merely assuming that X^* is compactly supported, while arbitrarily extrapolating $g(x^*, \theta)$ outside of that compact support to make it absolutely integrable, does not provide a viable way to meet the absolute integrability requirement. This would demand that the estimated $E[Y|Z = z]$ and $E[XY|Z = z]$ be similarly extrapolated outside the

support of the data in a *mutually compatible* way (so that Equations (6) and (7) remain satisfied in the extrapolated tails). This is only possible if one already knows the distribution of the disturbance in the instrument equation, which is not the case.

A similar situation arises when the distribution of X^* is supported on \mathbb{R} , but one is only interested in $E[Y|X^* = x^*]$ for x^* in some compact interval. Even in that case, the knowledge of $E[Y|Z = z]$ and $E[XY|Z = z]$ over the whole real line is still required. The value of the unobserved conditional expectations $E[Y|X^* = x^*]$ at a given point x^* depends on the whole shape of the observed conditional expectations $E[Y|Z = z]$ and $E[XY|Z = z]$ at all points $z \in \mathbb{R}$, because their Fourier transforms (used to establish identification) are integrals with respect to z over the whole real line. Truncating the tails of $E[Y|Z = z]$ and $E[XY|Z = z]$ beyond a certain value of z in the hope of solving this problem results in a large error that can only be properly shown to converge to zero, as the truncation point is moved towards infinity, by using the theory of generalized functions. In other words, the tails of these conditional expectations are, in fact, responsible for the generalized function components in the Fourier transforms considered and the theory of generalized functions is therefore essential to handle these tails without introducing nonnegligible errors.

It is worth noting that Assumption 1 is very weak, as the only commonly used function that does not satisfy it is the exponential. However, as shown in Schennach (2004b), with an exponential specification, $g(x^*)$ is actually not identified from Equations (6) and (7), so allowing for exponentials would bring no additional benefits. We then need a few more conditions to state our identification result.

Assumption 2 (i) $E[|U|] < \infty$ and (ii) $\phi(\zeta) \neq 0$ for all $\zeta \in \mathbb{R}$.

The assumption of existence of the first absolute moment of U is very weak. Requiring the characteristic function $\phi(\zeta)$ of the disturbance U to be nonvanishing everywhere is a standard assumption in the deconvolution literature (Carroll, Rupert, and Stefanski (1995), Fan (1991), Fan and Truong (1993), Li and Vuong (1998),

Li (2002), Horowitz and Markatou (1996), Schennach (2004a)). The only common distributions that are excluded by the requirement that $\phi(\zeta) \neq 0$ are the uniform and the triangular distributions. The normal, t , χ^2 , gamma, and double exponential distributions all satisfy this assumption. When $\phi(\zeta) = 0$ over some set, $\gamma(\zeta)$ can take any value over the interior of that set without changing the observables $\varepsilon_y(\zeta)$ and $\varepsilon_{xy}(\zeta)$ and it is therefore impossible to fully recover $\gamma(\zeta)$.⁵

Assumption 3 *There exists a positive finite or infinite constant $\bar{\zeta}$ such that (i) $\gamma(\zeta) \neq 0$ almost everywhere in $[-\bar{\zeta}, \bar{\zeta}]$ and (ii) $\gamma(\zeta) = 0$ for all $|\zeta| > \bar{\zeta}$.*⁶

While Assumption 3 requires that $\gamma(\zeta)$ vanish beyond some frequency $\bar{\zeta}$, it allows $\bar{\zeta}$ to be infinite, so that the case $\gamma(\zeta) \neq 0$ almost everywhere in \mathbb{R} is included as a particular case. It is important to note that the constant $\bar{\zeta}$ does not need to be known. Assumption 3 is also fairly weak, as it essentially excludes specifications whose Fourier transforms vanish on a finite interval (in which case $\gamma(\zeta)$ would not be fully identified). Such functions exist (e.g. consider a purely sinusoidal function), but are not commonly encountered in practical application. The asymmetry in the assumptions regarding $\phi(\zeta)$ and $\gamma(\zeta)$ comes from the fact that our main focus is on identifying $\gamma(\zeta)$ and not $\phi(\zeta)$. If we wanted to identify $\phi(\zeta)$ everywhere we would need to impose that $\gamma(\zeta) \neq 0$ almost everywhere in \mathbb{R} . As discussed below, Assumptions 2 and 3 can be relaxed when parametric constraints on $g(x^*)$ are imposed, since it may then be sufficient to identify $\gamma(\zeta)$ for some, but not necessarily all, ζ . The following nonparametric identification result gives an explicit expression for the regression function $g(x^*)$ in terms of Fourier transforms of observable quantities, which automatically implies that there cannot be two different $g(x^*)$ that are observationally equivalent.

⁵While it may be possible to relax this assumption to $\phi(\zeta) \neq 0$ almost everywhere, at the expense of strengthening our assumptions on $\gamma(\zeta)$ and $\phi(\zeta)$, we do not do so here for conciseness.

⁶There are no constraints on the behavior of $\gamma(\zeta)$ at $\zeta = \pm\bar{\zeta}$. Also note that if $\gamma(\zeta)$ contains delta function derivatives at some point ξ , $\gamma(\zeta)$ is not equal to the zero function at $\zeta = \xi$ and therefore $\gamma(\xi) \neq 0$.

Theorem 1 Under Assumptions 1-3, $g(x^*)$ in Model (5) is nonparametrically identified. Moreover, if $\bar{\zeta} > 0$ in Assumption 3,

$$g(x^*) = (2\pi)^{-1} \int \gamma(\zeta) e^{-i\zeta x^*} d\zeta \quad (11)$$

where⁷

$$\gamma(\zeta) = \begin{cases} 0 & \text{if } \varepsilon_y(\zeta) = 0 \\ \varepsilon_y(\zeta) / \phi(\zeta) & \text{otherwise} \end{cases}, \quad (12)$$

where $\phi(\zeta)$ is characteristic function of $U \equiv -\Delta X^*$, given⁸ for $|\zeta| \leq \bar{\zeta}$, by

$$\phi(\zeta) = \exp\left(\int_0^\zeta \frac{\mathbf{i}\varepsilon_{(z-x)y,o}(\xi)}{\varepsilon_{y,o}(\xi)} d\xi\right) \quad (13)$$

and where $\varepsilon_{y,o}(\xi)$ and $\varepsilon_{(z-x)y,o}(\xi)$ denote the ordinary function components of $\varepsilon_y(\xi) \equiv \int E[Y|Z=z] e^{i\xi z} dz$ and $\varepsilon_{(z-x)y}(\xi) \equiv \int E[(Z-X)Y|Z=z] e^{i\xi z} dz$, respectively.

The proof of this result, given in the appendix, is based on the following intuition. Starting from Equations (8) and (9) and calculating $d\varepsilon_y(\zeta)/d\zeta - \mathbf{i}\varepsilon_{xy}(\zeta)$, we obtain

$$\mathbf{i}\varepsilon_{(z-x)y}(\zeta) = \gamma(\zeta) \dot{\phi}(\zeta). \quad (14)$$

Next, in Equations (8) and (14), we replace each generalized function by its decomposition as a sum of an ordinary function and a purely singular component. Equating the ordinary function components amongst themselves and using the fact that $\phi(\zeta)$ and $\dot{\phi}(\zeta)$ are ordinary functions, we obtain analogues of Equations (8) and (14) involving only ordinary functions. Dividing the ordinary part of (14) by the ordinary

⁷When the ratio $\mathbf{i}\varepsilon_{(z-x)y,o}(\xi)/\varepsilon_{y,o}(\xi)$ takes the forms $0/0$ or ∞/∞ , we take the convention that $\mathbf{i}\varepsilon_{(z-x)y,o}(\xi)/\varepsilon_{y,o}(\xi) \equiv \lim_{\xi^* \rightarrow \xi} \mathbf{i}\varepsilon_{(z-x)y,o}(\xi^*)/\varepsilon_{y,o}(\xi^*)$, a limit that is shown to always exist in the proof of the theorem. Also, by convention, the statement $\varepsilon_y(\zeta) = 0$ is false when $\varepsilon_y(\zeta)$ contains a delta function derivative at ζ .

⁸Although Equation (13) is reminiscent of an identity due to Kotlarski (see Rao (1992), p. 21), it differs substantially in that it involves the Fourier transforms of conditional expectations rather than probability densities. Also, in the repeated measurement case covered by Kotlarski's result, the distribution of the true regressor can be expressed solely in terms of the joint distribution of two error-contaminated measurements but not of the dependent variable Y . In contrast, in the more generally applicable instrumental variable case covered by Equation (13), the dependent variable Y , the mismeasured regressor X and the instrument Z are intricately interrelated and all play an essential role in obtaining this result.

part of (8), we obtain the following differential equation in $\phi(\zeta)$

$$\frac{\dot{\phi}(\zeta)}{\phi(\zeta)} = \frac{\mathbf{i}\varepsilon_{(z-x)y,o}(\zeta)}{\varepsilon_{y,o}(\zeta)} \quad (15)$$

which can be solved, with the initial condition $\phi(0) = 1$, to yield Equation (13). Once $\phi(\zeta)$ is known, $\gamma(\zeta)$ can be obtained from Equation (8), wherever it is nonzero.

Interestingly, while $\gamma(\zeta)$ is identified for all ζ , $\phi(\zeta)$ is only identified by Equation (13) for $|\zeta| \leq \bar{\zeta}$. In the relatively common case where $\bar{\zeta} = \infty$, $\phi(\zeta)$ is fully identified, thus implying that the density of U is identified and so is the density of X^* , since the characteristic function of X^* is given by $E[e^{i\zeta Z}] \phi(\zeta)$.

In the case where $g(x^*)$ is parametrically specified as $g(x^*, \theta)$, where the parameter vector θ belongs to some set Θ , identification can be shown under even weaker assumptions. Letting $\gamma(\zeta, \theta)$ denote the Fourier transform of $g(x^*, \theta)$ with respect to x^* and letting θ^* denote the true value of θ , we make the following assumption.

Assumption 4 $\{\theta \in \Theta : \gamma(\zeta, \theta^*) = \gamma(\zeta, \theta) \text{ for all } \zeta \in]-\zeta^*, \zeta^*[\} = \{\theta^*\}$ where ζ^* denotes the smallest $\zeta > 0$ such that $\phi(\zeta) = 0$ and such that $\gamma(\zeta, \theta^*) \neq 0$ almost everywhere in $]-\zeta^*, \zeta^*[$. (If $\phi(\zeta) \neq 0$ for all $\zeta \in \mathbb{R}$, then $\zeta^* = \infty$.)

Assumption 4 requires that the knowledge of the function $\gamma(\zeta, \theta)$ over some interval $\zeta \in]-\zeta^*, \zeta^*[$ is sufficient to determine θ^* , which is plausible for a number of reasons. First, parametric identification only requires the determination of a finite number of degrees of freedom, so the knowledge of the value of the function $\gamma(\zeta, \theta)$ everywhere is clearly not needed. Second, examples of functions where some of the elements of θ only affect the value of $\gamma(\zeta, \theta)$ outside of a neighborhood of the origin would be difficult to construct and involve specifications with an oscillating behavior that are rarely encountered in applications. Finally, the smallest ζ such that $\phi(\zeta) = 0$ is never zero, because $\phi(\zeta)$ is a continuous function that satisfies $\phi(0) = 1$, thus implying that $\phi(\zeta)$ is necessarily nonzero in a neighborhood of the origin. Our

parametric identification result (proven in the Appendix) has the important feature that it permits $\phi(\zeta)$ to vanish, and can be stated as follows.

Corollary 1 *Under Assumptions 1, 2(i) and 4, the parameter vector θ^* is identified.*

3 Semiparametric Estimation

Although $g(x^*)$ and $m(w)$ in Model (5) are actually nonparametrically identified, we primarily focus on estimation in the case where $g(x^*)$ and $m(w)$ are parametrically specified, in order to provide a root n consistent and asymptotically normal estimator. Accordingly, we denote the regression function by $g(x^*, \theta)$ and its Fourier transform by $\gamma(\zeta, \theta)$ where $\theta \in \mathbb{R}^{N_\theta}$ is the parameter to be determined. Similarly, the unknown function $m(w)$ entering the instrumental equation is written as $m(w, \alpha)$ where $\alpha \in \mathbb{R}^{N_\alpha}$ is to be determined. Note that the distributions of all the disturbances remain nonparametric, making this a semiparametric estimation problem.

The following three-step estimation procedure relies on the fact that the equations obtained in Lemma 1 imply conventional moment conditions, which suggests an estimator taking the familiar form of a Generalized Method of Moment (GMM) estimator with a plugged-in nonparametric density estimate.

Step 1. The parameter α in Model (2) is estimated using standard (nonlinear) least-squares on the specification $X = m(W, \alpha) + (\Delta X^* + \Delta X)$ where $E[(\Delta X^* + \Delta X) | W] = 0$ by the assumptions of Model (2).

Step 2. The variable Z is then constructed from the instruments W via $Z = m(W, \alpha)$ and its density $p(z)$ is estimated using a standard kernel density estimator with bandwidth h .

Step 3. The parameter vector θ is estimated using the empirical analogue of the

moment condition $E [Q (X, Y, Z, \theta, p)] = 0$, where

$$Q(x, y, z, \theta, p) = \left[\begin{array}{c} yr_y(z, \theta) + xy r_{xy}(z, \theta) \\ \frac{yr_{1y}(z, \theta)}{p(z)} - 1 \end{array} \right], \quad (16)$$

and where the functions $r_{1y}(z, \theta)$, $r_y(z, \theta)$, and $r_{xy}(z, \theta)$ are vectors of known functions of $\gamma(\zeta, \theta)$ to be subsequently defined. An explicit algorithm to construct these moment conditions is given in Section 3.2, while an intuitive derivation is provided in Section 3.1.

Clearly, regularity conditions will be needed to ensure that the expectations in Step 3 exist and can be root n consistently estimated, despite the presence of a division by the density $p(z)$. Also, as the density $p(z)$ needs to be estimated, standard trimming techniques will be needed to handle divisions by a potentially vanishing random quantity.⁹

Our focus on the estimation of parametric specifications parallels the emphasis on parametric models found in the empirical literature and is governed by the realization, in the theoretical literature, that nonparametric estimation in the presence of endogeneity or measurement error typically suffers from slow convergence issues (e.g. Fan (1991), Darolles, Florens, and Renault (2002), Newey and Powell (2003), Schennach (2004c)) to which semiparametric restrictions provide a pragmatic solution (e.g. Schennach (2004a), Chen, Hong, and Tamer (2005), Hu and Ridder (2004), Newey (1994)). Even when a semiparametric estimation approach is selected, the availability of a nonparametric identification result guarantees that the identifiability of the model is not crucially dependent on the particular parametric specification of the model (see Chesher (2005), p. 1542, for a discussion of these issues).

⁹While efficiency considerations would suggest stacking the moment conditions of steps 1 and 3 to yield a single-step estimator, the multistep nature of the estimator proposed above simplifies its implementation by avoiding repeated nonparametric estimation of the density of Z while the parameter α is being optimized.

3.1 Heuristic derivation of the moment conditions

3.1.1 Ordinary function case

To provide some intuition regarding the form of the moment conditions, we start by providing suitable functions $r_y(z, \theta)$, $r_{xy}(z, \theta)$ and $r_{1y}(z, \theta)$ in the simple case where the Fourier transforms of $g(x^*, \theta)$ and $x^*g(x^*, \theta)$ (and hence, all Fourier transforms in Lemma 1) are ordinary functions. We will relax this assumption in the next section.

Equations (8) and (9) can be manipulated to eliminate $\phi(\zeta)$ and yield a single equation in one unknown function $\gamma(\zeta, \theta)$

$$\varepsilon_{xy}(\zeta) \gamma(\zeta, \theta) = -\mathbf{i} \varepsilon_y(\zeta) \dot{\gamma}(\zeta, \theta). \quad (17)$$

The functional forms of $\gamma(\zeta, \theta)$ and $\dot{\gamma}(\zeta, \theta)$ are known from the assumed functional form of $g(x^*, \theta)$, while $\varepsilon_{xy}(\zeta)$ and $\varepsilon_y(\zeta)$ are Fourier transforms of conditional expectations involving observable variables. Equation (17) effectively provides us with an infinite number of restrictions, as it must hold for all $\zeta \in \mathbb{R}$. Since $g(x^*, \theta)$ is parametric, a finite number of restriction suffices and we can replace Equation (17) by a finite system of equations defined by

$$\int \varepsilon_y(\zeta) \mathbf{i} \dot{\gamma}(\zeta, \theta) \omega(\zeta) d\zeta + \int \varepsilon_{xy}(\zeta) \gamma(\zeta, \theta) \omega(\zeta) d\zeta = 0 \quad (18)$$

for some vector of weighting functions $\omega(\zeta)$ chosen so that basic rank conditions hold in order to avoid colinearity among the equations. The choice of the weighting functions $\omega(\zeta)$ is conceptually analogous to the choice of which nonlinear functions of a given set of instruments are to be used in conventional instrumental variable estimation, when the disturbances are assumed to satisfy conditional mean restrictions that, in principle, imply an infinite family of moment restrictions. Next, if we define

$$\rho_y(\zeta, \theta) = \mathbf{i} \dot{\gamma}(-\zeta, \theta) \omega(-\zeta), \quad (19)$$

$$\rho_{xy}(\zeta, \theta) = \gamma(-\zeta, \theta) \omega(-\zeta) \quad (20)$$

Equation (18) can be written as $\int \varepsilon_y(\zeta) \rho_y(-\zeta, \theta) d\zeta + \int \varepsilon_{xy}(\zeta) \rho_{xy}(-\zeta, \theta) d\zeta = 0$ and, by Parseval's identity,¹⁰ this equality can be expressed as

$$\int E[Y|Z=z] r_y(z, \theta) dz + \int E[XY|Z=z] r_{xy}(z, \theta) dz = 0 \quad (21)$$

where $r_y(z, \theta)$ and $r_{xy}(z, \theta)$ denote the inverse Fourier transforms of $\rho_y(\zeta, \theta)$ and $\rho_{xy}(\zeta, \theta)$, respectively:

$$r_y(z, \theta) = (2\pi)^{-1} \int \rho_y(\zeta, \theta) e^{-i\zeta z} d\zeta \quad (22)$$

$$r_{xy}(z, \theta) = (2\pi)^{-1} \int \rho_{xy}(\zeta, \theta) e^{-i\zeta z} d\zeta. \quad (23)$$

Now, by multiplying and dividing the integrands by $p(z)$, the density of Z , Equation (21) becomes, after using iterated expectations,

$$E \left[\frac{Y r_y(Z, \theta) + XY r_{xy}(Z, \theta)}{p(Z)} \right] = 0, \quad (24)$$

which yields the upper block of moment conditions in Equation (16).¹¹

Since $p(z) \rightarrow 0$ as $|z| \rightarrow \infty$, it is essential that the numerator in Equation (24) decays sufficiently rapidly as $|z| \rightarrow \infty$ to ensure the existence of the expectation. This can be achieved by noting that a function's rate of decay as its argument goes to infinity is governed by the smoothness of its Fourier transform. Hence, choosing a $\omega(\zeta)$ such that $\rho_y(\zeta, \theta)$ and $\rho_{xy}(\zeta, \theta)$ are very smooth will ensure that the numerator of Equation (24) is rapidly decaying.

Unfortunately, Equation (24) cannot be used to identify the scale of $\gamma(\zeta, \theta)$: Since $r_y(z, \theta)$ and $r_{xy}(z, \theta)$ are both linearly related to $\gamma(\zeta, \theta)$, multiplying $\gamma(\zeta, \theta)$ by a constant maintains the equality in Equation (24). The lower block of moment conditions in Equation (16) is therefore introduced to pin down the scale. It can be derived

¹⁰ $2\pi \int a(z) b(z) dz = \int \alpha(\zeta) \beta(-\zeta) d\zeta$ where $\alpha(\zeta)$ and $\beta(\zeta)$ are the Fourier transforms of $a(z)$ and $b(z)$, respectively.

¹¹Combinations of Fourier techniques with GMM estimation have been considered by other authors (e.g. Taupin (1998), Carrasco and Florens (2002) and Singleton (2001)).

along similar lines as above, starting from Equation (8), evaluated at $\zeta = 0$, while noting that $\phi(0) = 1$, since a proper distribution must integrate to 1.

In practice, the functions $r_y(z, \theta)$, $r_{xy}(z, \theta)$ and $r_{1y}(z, \theta)$ will usually be evaluated via numerical techniques (as described in the Supplementary Material). Fortunately, these are nonrandom functions that depend on the model specification and not on the data. Hence, the magnitude of errors in the numerical approximations made can be easily checked by gradually varying the parameters controlling precision (such as the size of the grid used to store the value of the functions and to perform numerical (inverse) Fast Fourier transforms). One can push the numerical accuracy of the calculations to any desired level, regardless of sample size.

3.1.2 Generalized functions case

While the previous subsection motivates the form of Equation (16) under the assumption that all Fourier transforms are ordinary functions, we will now intuitively describe why the same basic form of moment conditions applies even when $\gamma(\zeta, \theta)$, $\varepsilon_y(\zeta)$ and $\varepsilon_{xy}(\zeta)$ are generalized functions. An explicit expression for the moment conditions is given in the next section. The idea is to first decompose $\gamma(\zeta, \theta)$ as a sum of an ordinary function and a purely singular component.

Assumption 5 $\gamma(\zeta, \theta)$ admits the decomposition¹²

$$\gamma(\zeta, \theta) = \gamma_o(\zeta, \theta) + 2\pi \sum_{k=0}^{\bar{k}} \gamma_k(\theta) (-\mathbf{i})^k \delta^{(k)}(\zeta) \quad (25)$$

where $\gamma_o(\zeta, \theta)$ is an ordinary function, $\bar{k} \in \mathbb{N}$, and the $\gamma_k(\theta)$ for $k = 0, \dots, \bar{k}$ are θ -dependent scalar parameters. Without loss of generality, $\gamma_{\bar{k}}(\theta) \neq 0$.

Since the functional form of $g(x^*, \theta)$ is known, this decomposition can be performed exactly via an analytic calculation¹³ of the Fourier transform of $g(x^*, \theta)$ (ex-

¹²The factor $(-\mathbf{i})^k$ is included so that the coefficients $\gamma_k(\theta)$ are real-valued.

¹³There exist numerous symbolic computational tools which can calculate Fourier transforms that include generalized functions, such as Maple or Mathematica. Alternatively, Table I in Lighthill (1962) provides numerous Fourier transforms.

amples are given in the Supplementary Material). Equation (25) assumes that all singularities (the delta function derivatives $\delta^{(k)}(\zeta)$) are centered at $\zeta = 0$. While it is straightforward to extend our treatment to allow for singularities at other locations, thus allowing for sines and cosines in the specification, we do not explore this eventuality here. Singularities in $\gamma(\zeta, \theta)$ located away from the origin are only possible if the tails of $g(x^*, \theta)$ have an oscillating behavior as $|x^*| \rightarrow \infty$. Model specifications having this property are not commonly encountered in practical applications. The benefit of a simplified notation therefore outweighs the slight loss in generality.

Since $\dot{\gamma}(\zeta, \theta)$, $\varepsilon_y(\zeta)$ and $\varepsilon_{xy}(\zeta)$ also admit a decomposition similar¹⁴ to Assumption 5, we can substitute all of these decompositions into Lemma 1. Equating ordinary functions amongst themselves and singular components amongst themselves yields separate equations for the ordinary and singular components.

The equations involving ordinary functions only can be handled very much like in the heuristic treatment of the previous section. The only technical difficulty lies in the fact that, when $\varepsilon_y(\zeta)$ and $\varepsilon_{xy}(\zeta)$ are *estimated*, their ordinary and singular components can no longer be easily separated via an analytic calculation. This is handled by selecting weighting functions (i.e. $\omega(\zeta)$ in the previous section) that go to zero sufficiently fast in the neighborhood of the origin so that the singular components of $\varepsilon_y(\zeta)$ and $\varepsilon_{xy}(\zeta)$ do not contribute to the value of the inner products in Equation (18). For instance, a function that behaves as ζ^2 near the origin has a zero inner product with $\delta^{(0)}(\zeta)$ and $\delta^{(1)}(\zeta)$, since the function ζ^2 and its first derivative vanish at the origin, and since an inner product of a given function with a delta function derivative extracts the corresponding derivative of that function at the origin.

The equations involving the singular components turn out to be analogous to the equations obtained by Hausman, Newey, Ichimura, and Powell (1991) in the polynomial case, which is not surprising, since the delta function derivatives in Equation (25)

¹⁴The highest order of the delta function derivatives present in $\varepsilon_y(\zeta)$ and $\varepsilon_{xy}(\zeta)$ are the same as in $\gamma(\zeta)$ and $\dot{\gamma}(\zeta)$, respectively.

represent the Fourier transforms of a polynomial. In fact, as explained in the Supplementary Material, for a suitable choice of the weighting functions, the asymptotic variance of our estimator coincides with Hausman, Newey, Ichimura, and Powell (1991). However, for general nonlinear specifications, the situation is more involved than in the purely polynomial case, because we must devise a way to extract the singular components of the estimated $\varepsilon_y(\zeta)$ and $\varepsilon_{xy}(\zeta)$. This can be achieved by selecting weighting functions which go to zero at the origin at a rate such that each singular component can be individually extracted. For instance, the inner product of a function that behaves as ζ near the origin will extract the prefactor of the $\delta^{(1)}(\zeta)$ component only, since $\int \delta^{(k)}(\zeta) \zeta d\zeta = [d^k \zeta / d\zeta^k]_{\zeta=0} = 1$ ($k = 1$). These weighting functions must also be orthogonal to the ordinary parts of $\varepsilon_y(\zeta)$ and $\varepsilon_{xy}(\zeta)$, which can be achieved under suitable regularity conditions (described in Section 3.2 below). The idea is to note that if $\phi(\zeta)$ is sufficiently smooth, it is possible to predict its value, say at the origin, from its value in a nearby region (e.g., via a Taylor series). This extrapolation operation can be expressed in terms of an inner product with some suitably chosen $\mu_a(\zeta)$, e.g. $\int \mu_a(\zeta) \phi(\zeta) d\zeta = \phi(0)$. Now consider another, different, way of making this prediction, e.g. $\int \mu_b(\zeta) \phi(\zeta) d\zeta = \phi(0)$ and note that, for the weighting function $\mu(\zeta) = \mu_a(\zeta) - \mu_b(\zeta)$, we have $\int \mu(\zeta) \phi(\zeta) d\zeta = 0$. Hence, the function $\mu(\zeta) / \gamma_o(\zeta, \theta)$ will be orthogonal to $\varepsilon_{y,o}(\zeta)$, since $\phi(\zeta) = \varepsilon_{y,o}(\zeta) / \gamma_o(\zeta, \theta)$ (from the ordinary components of Equation (8)) implies that $\int \varepsilon_{y,o}(\zeta) (\mu(\zeta) / \gamma_o(\zeta, \theta)) d\zeta = \int \phi(\zeta) \mu(\zeta) d\zeta = 0$.

3.2 Formal construction of the moment conditions

This section provides explicit expressions for the functions $r_y(z, \theta)$, $r_{xy}(z, \theta)$ and $r_{1y}(z, \theta)$ entering the moment conditions (Equation (16)) in terms of user-specified functions that must satisfy specific constraints. Although more general functions could be employed, the validity of the suggested expressions will be established in Theorem 2, proven in the Appendix, and examples are given in the Supplementary Material. This section is not essential to understand the main ideas underlying the estimation

method and can be skipped upon first reading. While we provide a way to obtain a just-identified set of moment conditions, it is straightforward to obtain overidentifying restrictions, if desired. We first need a few definitions that will serve as building blocks for the moment conditions.

Definition 1 *Let \mathcal{G} denote the set of all functions $\lambda : \mathbb{R} \mapsto \mathbb{C}$ that can be written as linear combinations of products of Gaussians and polynomials.*

These functions are smooth and rapidly decaying, and so are their inverse Fourier transforms. This will help ensure that the functions entering the moment conditions are rapidly decaying functions of z , so that the required expectations exist. The Supplementary Material provides alternative choices of the set \mathcal{G} that relax some of the assumptions needed below.¹⁵

Definition 2 *Let \mathcal{S}_c for some $c \in \mathbb{R}$ denote the set of all functions that can be written as $\sum_{k=0}^{\infty} \frac{1}{k!} d^k (\zeta^k \lambda(\zeta)) / d\zeta^k$ for some $\lambda \in \mathcal{G}$ satisfying the constraint $\int \lambda(\zeta) d\zeta = c$.*

The constraint $\int \lambda(\zeta) d\zeta = c$ is trivial to impose, since it is just a linear constraint on the coefficients of the polynomial entering the function $\lambda \in \mathcal{G}$. In practice, functions in \mathcal{S}_c can be approximated with an arbitrary accuracy by $\sum_{k=0}^{k^*} \frac{1}{k!} d^k (\zeta^k \lambda(\zeta)) / d\zeta^k$ for some $\lambda \in \mathcal{G}$ and for k^* sufficiently large.

Definition 3 *Let \mathcal{C} be the set of functions $\lambda(\zeta)$ satisfying $\lambda(-\zeta) = \lambda^\dagger(\zeta)$ for all $\zeta \in \mathbb{R}$, where \dagger denotes complex conjugation.*

Functions in this class have the property that their inverse Fourier transforms are real-valued. Note that these functions are entirely determined by the value they take for $\zeta \geq 0$. The following definition will be helpful in connection with our use of generalized functions.

¹⁵For instance, the moment generating function of U can exist only over an interval instead of over \mathbb{R} , as assumed in Section 3.2.3 below. The definition of \mathcal{G} can also be expanded to include negative powers in addition to polynomials, which can be useful to cancel out potential divergences in $\gamma_o(\zeta, \theta)$.

Definition 4 We say that a function $\lambda(\zeta)$ has k vanishing derivatives at the origin if $d^j \lambda(\zeta) / d\zeta^j = 0$ for $j = 0, \dots, k$ (with the convention that $d^0 \lambda(\zeta) / d\zeta^0 \equiv \lambda(\zeta)$).

The vector-valued functions $r_y(z, \theta)$, $r_{xy}(z, \theta)$ and $r_{1y}(z, \theta)$ entering the moment conditions are each composed of two pieces, one piece (denoted by a subscript o) aimed at determining $\gamma_o(\zeta, \theta)$ in Equation (25) and the other (denoted by a subscript s) aimed at determining the coefficients $\gamma_k(\theta)$ of the singular components in Equation (25) in Assumption 5:

$$r_y(z, \theta) = (r'_{y,o}(z, \theta), r'_{y,s}(z, \theta))' \quad (26)$$

$$r_{xy}(z, \theta) = (r'_{xy,o}(z, \theta), r'_{xy,s}(z, \theta))' \quad (27)$$

$$r_{1y}(z, \theta) = (r_{1y,o}(z, \theta), r_{1y,s}(z, \theta))'. \quad (28)$$

We will construct each one separately. Note that in special cases, detailed below, some of these subvectors may be “empty”.

3.2.1 Ordinary function components

Shape parameters Let N_o be the number of degrees of freedom of θ that affect the shape of $\gamma_o(\zeta, \theta)$, but not its overall scale.¹⁶ This can be determined by inspection from Equation (25) for the particular regression function of interest. If $N_o = 0$, then $r_{y,o}(z, \theta)$ and $r_{xy,o}(z, \theta)$ are “empty”. Otherwise, select noncolinear $\omega_j \in \mathcal{G} \cap \mathcal{C}$, for $j = 1, \dots, N_o$ such that $\dot{\gamma}_o(-\zeta, \theta) \omega_j(-\zeta)$ and $\gamma_o(-\zeta, \theta) \omega_j(-\zeta)$ have, respectively, \bar{k} and $\bar{k} + 1$ vanishing derivatives at the origin¹⁷ for all $\theta \in \Theta$. Set $\omega(\zeta) \equiv (\omega_1(\zeta), \dots, \omega_{N_o}(\zeta))$ and

$$r_{y,o}(z, \theta) = (2\pi)^{-1} \int \mathbf{i} \dot{\gamma}_o(-\zeta, \theta) \omega(-\zeta) e^{-\mathbf{i}\zeta z} d\zeta \quad (29)$$

$$r_{xy,o}(z, \theta) = (2\pi)^{-1} \int \gamma_o(-\zeta, \theta) \omega(-\zeta) e^{-\mathbf{i}\zeta z} d\zeta. \quad (30)$$

¹⁶More formally, N_o is the dimension of the space spanned by $\gamma_o(\zeta, \theta) / \gamma_o(\zeta_0, \theta)$ as θ varies over Θ for some $\zeta_0 \in \mathbb{R}$ such that $\gamma_o(\zeta_0, \theta) \neq 0$ almost everywhere in Θ .

¹⁷This is to ensure orthogonality to the singular components of $\varepsilon_y(\zeta)$ and $\varepsilon_{xy}(\zeta)$, respectively.

Scale parameter If it is impossible to vary θ in such a way that the net effect is simply to multiply $\gamma_o(\zeta, \theta)$ by a constant,¹⁸ then the scale of the regression function is already fixed by the model and the $r_{1y,o}(\zeta, \theta)$ vector is “empty”. Otherwise, let $\varpi \in \mathcal{S}_1 \cap \mathcal{C}$ be such that $\varpi(\zeta)/\gamma_o(\zeta, \theta)$ has \bar{k} vanishing derivatives at the origin for all $\theta \in \Theta$. The vanishing derivatives constraint is easy to impose by setting up a linear system of equations and by exploiting the natural linear parametrization of the set \mathcal{S}_1 . Set

$$r_{1y,o}(z, \theta) = (2\pi)^{-1} \int \frac{\varpi(-\zeta)}{\gamma_o(-\zeta, \theta)} e^{-i\zeta z} d\zeta. \quad (31)$$

Note that if $\gamma_o(\zeta, \theta) = 0$ at some point, it may be beneficial to select $\varpi(\zeta)$ so that it vanishes at the same point as $\gamma_o(\zeta, \theta)$, in order to avoid dealing with infinities.¹⁹

3.2.2 Singular components

We first introduce a few matrices needed for the definition of the moment conditions determining the singular components of $\gamma(\zeta, \theta)$. Define the $(\bar{k} + 1) \times (\bar{k} + 1)$ matrices $\Gamma_y(\theta)$ and $\Gamma_{xy}(\theta)$ as

$$\Gamma_{y,j+1 \ k+1}(\theta) = \binom{k+j}{j} \gamma_{k+j}(\theta) \mathbf{1}(k+j \leq \bar{k}) \text{ for } j, k = 0, \dots, \bar{k} \quad (32)$$

$$\Gamma_{xy,j+1 \ k+1}(\theta) = \binom{k+j+1}{j+1} \gamma_{k+j}(\theta) \mathbf{1}(k+j \leq \bar{k}) \text{ for } j, k = 0, \dots, \bar{k}. \quad (33)$$

where $\gamma_k(\theta)$ is as in Equation (25). If $\gamma_o(\zeta, \theta) = 0$ everywhere, select $\nu_{y,k} \in \mathcal{G} \cap \mathcal{C}$ for $k = 1, \dots, \bar{k} + 1$ and $\nu_{xy,k} \in \mathcal{G} \cap \mathcal{C}$ for $k = 1, \dots, \bar{k} + 2$ and skip to Equation (36), defining $\nu_{y,k}(\zeta, \theta) \equiv \nu_{y,k}(\zeta)$ and $\nu_{xy,k}(\zeta, \theta) \equiv \nu_{xy,k}(\zeta)$. Otherwise, let $\mu_{y,k} \in \mathcal{S}_0 \cap \mathcal{C}$ for $k = 1, \dots, \bar{k} + 1$ and $\mu_{xy,k} \in \mathcal{S}_0 \cap \mathcal{C}$ for $k = 1, \dots, \bar{k} + 2$ and define the vectors $\nu_y(\zeta, \theta)$ and $\nu_{xy}(\zeta, \theta)$ as

$$\nu_{y,k}(\zeta, \theta) = \frac{\mu_{y,k}(\zeta)}{\gamma_o(\zeta, \theta)} \text{ for } k = 1, \dots, \bar{k} + 1 \quad (34)$$

$$\nu_{xy,k}(\zeta, \theta) = \frac{\mu_{xy,k}(\zeta)}{i\dot{\gamma}_o(\zeta, \theta)} \text{ for } k = 1, \dots, \bar{k} + 2. \quad (35)$$

¹⁸More formally, if there exists no $\theta, \tilde{\theta} \in \Theta$ such that $\gamma_o(\zeta, \theta) = c\gamma_o(\zeta, \tilde{\theta})$ for some $c \in \mathbb{R}$, then the scale is fixed by the model.

¹⁹Points where $\varpi(\zeta)/\gamma_o(\zeta, \theta)$ takes the form $0/0$ or ∞/∞ can be “filled in” by continuity.

If $\gamma_o(\zeta, \theta)$ or $\dot{\gamma}_o(\zeta, \theta)$ vanish at some points, it may again be preferable to select functions $\mu_{y,k}(\zeta)$ and $\mu_{xy,k}(\zeta)$ that vanish at those same points in order to avoid divergences. Let $\mu_{y,k}$ and $\mu_{xy,k}$ be such that the $M_y(\theta)$ and $M_{xy}(\theta)$ matrices defined as

$$M_{y,j,k}(\theta) = (\mathbf{i})^{k-1} \frac{d^{k-1} \nu_{y,j}(0, \theta)}{d\zeta^{k-1}} \text{ for } j, k = 1, \dots, \bar{k} + 1 \quad (36)$$

$$M_{xy,j,k}(\theta) = (\mathbf{i})^{k-1} \frac{d^{k-1} \nu_{xy,j}(0, \theta)}{d\zeta^{k-1}} \text{ for } j, k = 1, \dots, \bar{k} + 2, \quad (37)$$

exist and are nonsingular. This will typically be the case if the $\mu_{y,k}$ and $\mu_{xy,k}$ are not colinear. Define the vectors $V_y(z, \theta)$ and $V_{xy}(z, \theta)$ as

$$V_y(z, \theta) = (2\pi)^{-1} \int \nu_y(-\zeta, \theta) e^{-i\zeta z} d\zeta \quad (38)$$

$$V_{xy}(z, \theta) = (2\pi)^{-1} \int \nu_{xy}(-\zeta, \theta) e^{-i\zeta z} d\zeta. \quad (39)$$

Shape parameters Let N_s be the number of degrees of freedom of θ upon which the coefficients $\gamma_k(\theta)$ of the singular part in Equation (25) depend, excluding the scale.²⁰ If $N_s = 0$, then $r_{y,s}(z, \theta)$ and $r_{xy,s}(z, \theta)$ are “empty”. Let $S_{i,j}$ denote a selection matrix extracting elements i through j (inclusively) of the vector it multiplies. Set²¹

$$r_{y,s}(z, \theta) = S_{2, N_s+1} (\Gamma_y(\theta))^{-1} (M_y(\theta))^{-1} V_y(z, \theta) \quad (40)$$

$$r_{xy,s}(z, \theta) = -S_{2, N_s+1} (\Gamma_{xy}(\theta))^{-1} S_{2, \bar{k}+2} (M_{xy}(\theta))^{-1} V_{xy}(z, \theta). \quad (41)$$

Scale parameter If it is impossible to change θ so that $(\gamma_0(\theta), \dots, \gamma_k(\theta))'$ is simply multiplied by a constant,²² then $r_{1y,s}(z, \theta)$ is “empty”. Otherwise, let

$$r_{1y,s}(z, \theta) = S_{1,1} (\Gamma_y(\theta))^{-1} (M_y(\theta))^{-1} V_y(z, \theta) \quad (42)$$

²⁰More formally, N_s is the dimension of the space spanned by $(\gamma_0(\theta), \dots, \gamma_k(\theta))' / \gamma_k(\theta)$ as θ varies over $\Theta \setminus \{\theta : \gamma_k(\theta) = 0\}$.

²¹Note that the vector $(M_{xy}(\theta))^{-1} V_{xy}(z, \theta)$ has one more element than $(M_y(\theta))^{-1} V_y(z, \theta)$ but that extra element can be shown to provide no useful information for identification purposes and is therefore deleted using the selection matrix $S_{2, \bar{k}+2}$.

²²More formally, if there exists no $\theta, \tilde{\theta} \in \Theta$ such that $\gamma_j(\theta) = c\gamma_j(\tilde{\theta})$ for $j = 0, \dots, \bar{k}$ and for some $c \in \mathbb{R}$, then $r_{s,1y}(z, \theta)$ is empty.

3.2.3 Summary

We now formally state that the moment conditions constructed in the previous section enable identification of the true value θ^* of the parameter vector. We first need a standard rank condition, often used in parametric identification results (e.g., see Theorem 5.1.1 in Hsiao (1983)).

Assumption 6 $E [Q (X, Y, Z, \theta, p)]$ exists and $E [\partial Q (X, Y, Z, \theta, p) / \partial \theta']$ is nonsingular for θ in an open neighborhood of θ^* , with $Q (x, y, z, \theta, p)$ given by Equation (16).

Given the identification result of Section 2, all that is typically needed for this assumption to hold is that the user-specified functions $\omega_k (\zeta)$, $\mu_{y,k} (\zeta)$ and $\mu_{xy,k} (\zeta)$ introduced in the previous section are not fortuitously colinear (e.g. $\omega_1 (\zeta) = \omega_2 (\zeta)$). Next, we need a constraint on the distribution of the disturbance in the instrument equation.

Assumption 7 The moment generating function of U , $E [e^{tU}]$, exists for all $t \in \mathbb{R}$.

The Supplementary Material describes how Assumption 7 can be further relaxed to the existence of the moment generating function over a finite interval. This type of assumption has been used in other treatments of nonlinear errors-in-variables models (e.g. Hausman, Newey, and Powell (1995)) and arguably stronger assumptions, such as compact support (Newey (2001)) or parametric tails (Newey and Powell (2003)) are often made. We are now ready to state our result, proven in the Appendix.

Theorem 2 Under Assumptions 1, 2(i), 4 and 5-7, if $Q (x, y, z, \theta, p)$ is as defined in Equation (16) and Section 3.2, then there exists a compact set $\Theta \subset \mathbb{R}^{N_\theta}$ containing θ^* in its interior such that $\theta = \theta^*$ is the only solution to $E [Q (X, Y, Z, \theta, p)] = 0$ in Θ . Assumption 7 is unnecessary when $r_{y,o} (z, \theta)$, $r_{xy,o} (z, \theta)$ and $r_{1y,o} (z, \theta)$ are “empty” or when $r_{y,s} (z, \theta)$, $r_{xy,s} (z, \theta)$, $r_{1y,s} (z, \theta)$ and $r_{1y,o} (z, \theta)$ are “empty”.²³

²³This is the case for the polynomial and the probit/logit models, respectively.

The additional contribution of Theorem 2, beyond what was shown in Corollary 1, is that, not only a parametric model can be identified, but a finite number of moment conditions suffice to achieve this goal.

4 Asymptotic properties

Our estimator takes the familiar form of a Generalized Method of Moment (GMM) estimator with a plugged-in nonparametric density estimate, thus admitting a relatively straightforward asymptotic analysis (paralleling, for instance, Newey (1994)). This is to be contrasted with earlier suggestions (Newey (2001)) of using a simulated method of moments sieve estimator of θ where the distribution of the disturbance U is approximated by a flexible functional form with a number of parameters that is allowed to grow with sample size. Although it would be interesting to investigate the asymptotic properties of Newey's estimator along the lines of Ai and Chen (2003), our estimator avoids the nonparametric estimation of the distribution of the disturbance U jointly with the parameter θ . As a result, our approach avoids the difficult (and so far unresolved) question of finding an asymptotically linear representation for Newey's sieve estimator.²⁴

A few fairly standard regularity conditions are needed to establish the asymptotic properties of the estimator $\hat{\theta}$, formally defined as follows.

Definition 5 $\hat{\theta}$ is the solution to $n^{-1} \sum_{j=1}^n Q(X_j, Y_j, m(W_j, \hat{\alpha}), \theta, \hat{p}(\cdot|\hat{\alpha})) 1(\hat{p}((m(W_j, \hat{\alpha})|\hat{\alpha}) \geq \tau) = 0$ where (Y_j, X_j, W_j) for $j = 1, \dots, n$ is a sample, $Q(x, y, z, \theta, p)$ is defined in Equation (16) and Section 3.2, $\hat{p}(\cdot|\alpha)$ is a nonparametric kernel density estimator (with kernel $K(\cdot)$ and bandwidth h) of $p(\cdot|\alpha)$, the density of $m(W, \alpha)$ for a given value of α , τ is a sample size-dependent trimming parameter and $\hat{\alpha}$ is the first step estimate of the model $X = m(W, \alpha) + \eta$, where $\eta \equiv \Delta X + \Delta X^*$ satisfies

²⁴In the sieve literature, finding this linear representation is usually achieved by solving an operator equation, which is highly nontrivial in the present case.

$E[\eta|W] = 0$. Let θ^* and α^* denote the true values of the corresponding parameters.

Assumption 8 (Y_j, X_j, W_j) is an iid sequence of random variables distributed as (Y, X, W) .

Assumption 9 (i) Let $\mathcal{B} \subset \mathbb{R}^{N_\alpha}$ be a compact set such that $\alpha^* = \arg \min_{\alpha \in \mathcal{B}} E[(X - m(W, \alpha))^2]$ is unique and lies in the interior of \mathcal{B} , (ii) $E[\sup_{\alpha \in \mathcal{B}} m^2(W, \alpha)] < \infty$ and $E[X^2] < \infty$, (iii) $m(w, \alpha)$ is continuous in α for $\alpha \in \mathcal{B}$, (iv) $m(w, \alpha)$ is continuously differentiable in α for $\alpha \in \mathcal{A}$, a neighborhood of α^* , (v) $E\left[\sup_{\alpha \in \mathcal{A}} \left\| \frac{\partial m(W, \alpha)}{\partial \alpha} \right\|^2\right] < \infty$, (vi) $E\left[\frac{\partial m(W, \alpha^*)}{\partial \alpha} \frac{\partial m(W, \alpha^*)}{\partial \alpha'}\right]$ is nonsingular, and (vii) $E\left[(X - m(W, \alpha^*))^2 \left\| \frac{\partial m(W, \alpha^*)}{\partial \alpha} \right\|^2\right] < \infty$.

Assumption 10 $Q(x, y, z, \theta, p)$ is continuously differentiable in θ for $\theta \in \Theta$.

Assumption 11 $E[\sup_{\alpha \in \mathcal{A}} \sup_{\theta \in \Theta} \|Q(X, Y, m(W, \alpha), \theta, p(\cdot|\alpha))\|] < \infty$.

Assumption 12 $E[\sup_{\alpha \in \mathcal{A}} \sup_{\theta \in \mathcal{N}} \|\partial Q(X, Y, m(W, \alpha), \theta, p(\cdot|\alpha)) / \partial \theta'\|] < \infty$ for some neighborhood $\mathcal{N} \subset \Theta$ of θ^* .

Assumption 13 $E[\psi_\theta(X, Y, W) \psi'_\theta(X, Y, W)]$ exists, where, letting $Z = m(W, \alpha^*)$,

$$\psi_\theta(x, y, w) = Q(x, y, m(w, \alpha^*), \theta^*, p(\cdot|\alpha^*)) - E[Q(X, Y, Z, \theta^*, p(\cdot|\alpha^*)) | Z = m(w, \alpha^*)] \quad (43)$$

Assumption 14 The kernel function $K(z)$ satisfies (i) $\int K(z) dz = 1$, (ii) $K(z) = K(-z)$ (iii) $\int K(z) z^j dz = 0$ for $j = 1, \dots, N_K - 1$ (iv) $\int |K(z)| |z|^{N_K} dz < \infty$ for some $N_K \in \mathbb{N}$ (v) $K(0) < \infty$ and (vi) $dK(z)/dz$ exists.

Assumption 15 $\sup_{\alpha \in \mathcal{A}} \sup_{z \in \mathbb{R}} |\partial^{N_K} p(z|\alpha) / \partial z^{N_K}| < \infty$.

Assumption 16 (i) $n^{1/2} h^2 \tau^2 \rightarrow \infty$ (ii) $n^{1/2} h^{N_K} \tau^{-1} \rightarrow 0$ (iii) $\tau \rightarrow 0$ (iv) $h \rightarrow 0$ as $n \rightarrow \infty$.

Assumption 17 $E [\sup_{\alpha \in \mathcal{A}} \sup_{\theta \in \Theta} Q(X, Y, m(W, \alpha), \theta, p(\cdot | \alpha)) 1(p(m(W, \alpha) | \alpha) \leq \tau)] = o(n^{-1/2})$.

Assumption 18 $Q(\theta, \alpha) \equiv E [Q(X, Y, m(W, \alpha), \theta, p(\cdot | \alpha))]$ and $\frac{\partial}{\partial \theta} Q(\theta, \alpha)$ are continuous in α for all $\alpha \in \mathcal{A}$, uniformly in θ for $\theta \in \Theta$.

Assumption 19 $E [\sup_{\alpha \in \mathcal{A}} \|\partial Q(X, Y, m(W, \alpha), \theta^*, p(\cdot | \alpha)) / \partial \alpha'\|] < \infty$.

Assumption 9 collects all the standard regularity conditions traditionally used to show asymptotic normality and root n consistency of the first-step estimator $\hat{\alpha}$ in iid settings. For added generality, it could be replaced by assuming that the first step estimate $\hat{\alpha}$ admits a stochastic expansion of the form $n^{1/2}(\hat{\alpha} - \alpha) = n^{-1/2} \sum_{i=1}^n \psi_{\alpha}(X_i, W_i) + o_p(1)$, where the influence function $\psi_{\alpha}(X, W)$ has a finite variance-covariance matrix. In light of the results found in Newey (1994), it should also be possible to allow for a nonparametric first step while still maintaining root n consistency and asymptotic normality of $\hat{\theta}$, although this is beyond the scope of the present work.

Assumptions 10, 11 and 12 impose conventional continuity and dominance conditions that imply uniform convergence in probability of the quantities that define the estimator and its limiting distribution. Assumption 13 ensures that the asymptotic variance of the estimator exists for α^* fixed, which is essential to obtain root n consistency. Assumption 14 defines a standard bias-reducing kernel of order N_K . Assumption 15 is used to show uniform convergence in probability of the kernel density estimate. Assumption 16 imposes constraints on the rates at which the bandwidth h and the trimming parameter τ can go to zero as $n \rightarrow \infty$. Assumption 17 ensures that the bias introduced by trimming is asymptotically negligible. Following standard practice (e.g. Härdle and Stoker (1989), Assumption 8 and Laverge and Vuong (1996) Theorem 1), this assumption is stated in a relatively high-level form. Finally, Assumptions 18 and 19 ensure that root n consistency of $\hat{\theta}$ is possible despite

the statistical noise in the first step estimator $\hat{\alpha}$. The following result is shown in the Supplementary Material, using standard techniques borrowed from Newey and McFadden (1994), Andrews (1995), Pagan and Ullah (1999) and Powell, Stock, and Stoker (1989).

Theorem 3 *Let the Assumptions of Theorem 2 hold. Under Assumptions 8 through 17, $n^{1/2}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathbb{N}(0, G^{-1}\Omega G^{-1})$, where $G = E[\partial Q(X, Y, Z, \theta^*, p(\cdot|\alpha^*)) / \partial \theta']$ and $\Omega = E[\Psi(X, Y, W) \Psi'(X, Y, W)]$, where*

$$\Psi(x, y, w) = \psi_{\theta}(x, y, w) + \frac{\partial Q(\theta^*, \alpha^*)}{\partial \alpha'} \psi_{\alpha}(x, w) \quad (44)$$

and

$$\psi_{\alpha}(x, w) = - \left(E \left[\frac{\partial m(W, \alpha^*)}{\partial \alpha} \frac{\partial m(W, \alpha^*)}{\partial \alpha'} \right] \right)^{-1} \frac{\partial m(w, \alpha^*)}{\partial \alpha} (x - m(w, \alpha^*)). \quad (45)$$

Note that the term added to $\psi_{\theta}(x, y, w)$ in Equation (43) is the correction term for the first-step estimation of α . As in conventional GMM, the expression for the asymptotic variance accounts for the potential presence of conditional heteroskedasticity in the disturbances ΔY and ΔX . Under standard regularity conditions (similar to the ones found in Newey and McFadden (1994) and Andrews (1995)), consistent estimates of the asymptotic variance can be obtained from the expressions of G and Ω given in Theorem 3, replacing all expectations by sample averages and all nonparametric quantities by their nonparametric kernel estimates. This explicit expression for the asymptotic variance can also be used to help select weighting functions (in Section 3.1.2) that yield a more efficient estimator.

5 Extensions

The results presented so far can be extended in a variety of useful directions. First, a vector of correctly measured regressors R can be trivially included in the identification proof by conditioning all expectations and densities on these additional regressors.

The corresponding estimator would rely on moment conditions that would now depend on R and that would be constructed by introducing a sufficient number of R -dependent weighting functions. The plugged-in density $p(z)$ in the moment conditions would become the joint density of Z and R . Sufficiently rapid convergence rates for this nonparametric first step can be achieved using higher-order kernels.

Second, multivariate mismeasured regressors X^* can be handled by using multivariate Fourier transforms, after noting that Z and ζ become multivariate as well. The identification result holds almost unchanged, after noting that the interval $[-\bar{\zeta}, \bar{\zeta}]$ is now a multidimensional connected region containing the origin and that Equation (13) becomes a path integral involving a dot product $\varepsilon_{(z-x)y,o}(\xi) \cdot d\xi$. Estimation would parallel the univariate case, except that singularities can now take the form of products of delta function derivatives along the different axes (e.g. $\delta^{(1)}(\zeta_1) \delta^{(3)}(\zeta_2)$).

By defining the dependent variable Y in Model (2) to be an indicator function such as $1(G \leq g(X^*))$, where G is the dependent variable of interest, it can be shown that the whole conditional distribution of G given X^* is, in fact, identified under suitably modified assumptions. This extension, along with a corresponding nonparametric estimator, is fully developed in a separate paper (Schennach (2005)).

Another interesting extension is to allow for a type of nonclassical errors-in-variables model recently investigated by Hyslop and Imbens (2001) for linear specifications and by Wang (2004) under parametric distributional assumptions. We can consider a variable Z contaminated with a so-called Berkson-type measurement error and an “instrument” V satisfying

$$\begin{aligned} X^* &= Z + \Delta X^* & \Delta X^* \text{ independent from } Z \text{ and } E[\Delta X^*] &= 0 \\ V &= aX^* + b + \Delta X & E[\Delta X|Z, \Delta X^*, \Delta Y] &= 0. \end{aligned}$$

This setup can be directly mapped into the form of Model (5) by first regressing V on Z to obtain the coefficients a and b and by then setting $X = (V - b)/a$. Of course, the asymptotic analysis would now have to account for the noise in the estimated X .

It is also possible to extend the estimation procedure proposed here by allowing

for the dimension of the parameter vectors θ and α (and the number of moment conditions) to go to infinity as sample size grows, thus providing a practical nonparametric estimation procedure.²⁵ Such an approach would still require some polynomial bound on the tail of $g(x^*)$, so that the maximum order of the delta function derivatives involved is known.²⁶ The asymptotics of such a nonparametric series estimator could be obtained along the following lines.²⁷ First, paralleling the results in Schenach (2004c) for repeated measurements, it should be possible to show that, under suitable regularity conditions, the bias is equal to the bias of the corresponding estimator in the absence of measurement error, up to some remainder terms that are asymptotically negligible in probability. This would enable the use of known results regarding the approximation rates of series approximations (e.g. Newey (1997), Chen (2005)). It would then be necessary to show that the finite-sample variance of $\hat{\theta}$ is well approximated by the asymptotic variance given in Theorem 3, from which an approximation to the variance of $g(x^*, \hat{\theta})$ could be obtained via the delta method. Finally, an optimal convergence rate could be obtained by choosing the number of parameters so as to balance the bias squared and the variance in the usual way. A formal statement of the requisite regularity conditions would be an interesting avenue of future research.

²⁵Using power series to represent $g(x^*)$ would probably not be wise, since their Fourier transforms are purely singular, making it difficult, if not impossible, for them to approximate the ordinary function component of the specification.

²⁶Restrictions on the tail behavior in nonparametric endogenous models are not uncommon. For instance, Newey and Powell (2003) even assume that the regression function has parametric tails, while the polynomial bound assumed here only corresponds to assuming an upper bound on a nonparametric tail.

²⁷Another possibility would be to substitute suitably trimmed nonparametric estimators into our identification result (Theorem 1). However, implementing an estimator involving random generalized functions entails technical difficulties that are avoided with the proposed moment condition approach.

6 Monte Carlo Simulations

We now investigate the finite-sample properties of the proposed estimator via Monte Carlo simulations. The data is generated according to Model (5) with $Z \sim N(0, 1)$, $U \sim N(0, 1/4)$, $\Delta X \sim N(0, 1/4)$ and where we consider three different specifications for $g(x^*, \theta)$, namely, a polynomial, a rational fraction and a probit model. All of these models satisfy the assumptions of Theorem 3. Note that the ratio of the standard deviation of the measurement error ΔX to the standard deviation of the true regressor X^* is $(1/2) / \sqrt{(1 + 1/4)} \approx 0.45$, so that the measurement error is fairly large. The distribution of Z is deliberately chosen to be a normal, in order to explore the behavior of the proposed estimator in a situation where the issue of the division by a thin-tailed density in the moment conditions is especially severe.

The Supplementary Material describes the details of the simulations, including the kernel used in the estimation of the density of Z , the construction of the moment conditions and the method used to select the bandwidth and trimming parameters. The Supplementary Material also provides an example of an empirical application.

The finite sample properties of the proposed estimator are studied by drawing 5000 samples of 1000 independent observations. As a point of comparison, we also calculate the standard instrumental variable estimator using $\partial g(z, \theta) / \partial \theta$ as a vector of instruments and X as the regressor in addition to a standard (nonlinear) least squares estimator using X as the regressor, although both of these estimators are clearly biased in the presence of measurement error.

Our main conclusion from these simulations is that, while the reduction in bias achieved with our estimator comes at the expense of increased standard errors for some coefficients, the overall Root Mean Square Error (RMSE), defined as the square root of the sum of the mean square errors of all coefficients, is still much lower for the proposed estimator than for the other two estimators (see Tables 1-3).

	Bias				Std. Dev.				RMSE				
	θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_4	all
present	-0.05	-0.07	-0.02	0.05	0.17	0.19	0.24	0.05	0.17	0.20	0.24	0.07	0.36
IV	0.00	0.42	0.00	-0.01	0.13	0.30	0.11	0.08	0.13	0.52	0.11	0.08	0.55
OLS	0.00	-0.43	0.00	0.21	0.07	0.13	0.06	0.04	0.07	0.45	0.06	0.22	0.51

Table 1: Simulations results for a polynomial specification: $g(x^*, \theta) = \theta_1 + \theta_2 x^* + \theta_3 (x^*)^2 + \theta_4 (x^*)^3$, where $\theta_1 = 1$, $\theta_2 = 1$, $\theta_3 = 0$, $\theta_4 = -0.5$ and with $\Delta y \sim N(0, 1/4)$.

	Bias			Std. Dev.			RMSE			
	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	all
present	0.107	0.117	-0.150	0.146	0.139	0.328	0.181	0.182	0.361	0.443
IV	-0.244	0.001	0.704	0.084	0.028	0.191	0.258	0.028	0.729	0.774
OLS	0.338	-0.166	-0.643	0.046	0.022	0.085	0.341	0.167	0.649	0.752

Table 2: Simulation results for the rational fraction specification: $g(x^*, \theta) = \theta_1 + \theta_2 x^* + \theta_3 (1 + (x^*)^2)^{-2}$, where $\theta_1 = 1$, $\theta_2 = 1$, $\theta_3 = 2$ and $\Delta y \sim N(0, 1/4)$.

	Bias		Std. Dev.		RMSE		
	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	all
present	0.05	-0.06	0.39	0.53	0.39	0.53	0.69
NIV	-0.38	0.75	0.57	1.11	0.68	1.34	1.50
NLS	0.39	-0.98	0.06	0.08	0.39	0.98	1.05

Table 3: Simulation results for the probit model, with specification $g(x^*, \theta) = \frac{1}{2}(1 + \text{erf}(\theta_1 + \theta_2 x^*))$, where $\theta_1 = -1$ and $\theta_2 = 2$ and Δy is equal to $1 - g(x^*, \theta)$ with probability $g(x^*, \theta)$ and equal to $-g(x^*, \theta)$ otherwise. Note that the results for the standard NIV estimator excludes the 50% of the replications that do not yield a finite estimate of θ_2 . The actual performance of NIV is therefore far worse than indicated in the table. Also note that Δy is heteroskedastic, as it is not independent of x^* in this example.

7 Conclusion

This paper addresses two unresolved issues. First, it is shown that instruments indeed permit nonparametric identification of general nonlinear regression models in the presence of measurement error. Second, when the regression function is parametrically specified, a root n consistent and asymptotically normal estimator is provided. The starting point of the proposed approach is a system of two functional equations that relate conditional expectations of observed variables to the regression function of interest, as first proposed by Hausman, Newey, Ichimura, and Powell (1991) for polynomial specifications. Both the proof of nonparametric identification and the construction of the estimator rely on a representation of these functional equations in terms of Fourier transforms. The proposed estimation procedure takes the form of a generalized method of moment estimator with a plugged-in nonparametric kernel density estimate. As a result, standard techniques borrowed from the semiparametrics literature can be used to establish its asymptotic properties.

A Proofs

Proof of Lemma 1. Assumption 1 implies that $\gamma(\zeta)$, $\varepsilon_y(\zeta)$, $\varepsilon_{xy}(\zeta)$ and $\phi(\zeta)$ are well-defined generalized functions. Therefore, the interchange of the order of integration and the interchange of derivative and integration operations performed below are allowed. (Formally, this is justified by noting that generalized functions are defined via an inner product with test functions. After a sufficient number of integrations by parts, this inner product can be written as the integral of an absolutely integrable function, thus permitting the use of Fubini's Theorem.)

$$\begin{aligned}\varepsilon_y(\zeta) &= \int \int g(z-u) dF(u) e^{i\zeta z} dz = \int \int g(z-u) e^{i\zeta z} dz dF(u) \\ &= \int \int g(x^*) e^{i\zeta(x^*+u)} d(x^*) dF(u) = \int g(x^*) e^{i\zeta x^*} d(x^*) \int e^{i\zeta u} dF(u) = \gamma(\zeta) \phi(\zeta)\end{aligned}$$

$$\begin{aligned}
\varepsilon_{xy}(\zeta) &= \int \int (z - u) g(z - u) dF(u) e^{i\zeta z} dz = \int x^* g(x^*) e^{i\zeta x^*} d(x^*) \int e^{i\zeta u} dF(u) \\
&= \left(-\mathbf{i} \frac{\partial}{\partial \zeta} \int g(x^*) e^{i\zeta x^*} d(x^*) \right) \int e^{i\zeta u} dF(u) \equiv -\mathbf{i} \dot{\gamma}(\zeta) \phi(\zeta).
\end{aligned}$$

■

Proof of Theorem 1. First note that Equations (8) and (9) are equivalent to

$$\varepsilon_y(\zeta) = \gamma(\zeta) \phi(\zeta) \quad (46)$$

$$\mathbf{i}\varepsilon_{(z-x)y}(\zeta) = \gamma(\zeta) \dot{\phi}(\zeta) \quad (47)$$

Indeed, differentiating each side of Equation (8) with respect to ζ yields

$$\frac{\partial}{\partial \zeta} \varepsilon_y(\zeta) = \frac{\partial}{\partial \zeta} \int E[Y|Z=z] e^{i\zeta z} dz = \mathbf{i} \int E[Z Y|Z=z] e^{i\zeta z} dz \equiv \mathbf{i}\varepsilon_{zy}(\zeta)$$

and, since $\dot{\phi}(\zeta) \equiv d\phi(\zeta)/d\zeta$,

$$\frac{\partial}{\partial \zeta} (\gamma(\zeta) \phi(\zeta)) = \dot{\gamma}(\zeta) \phi(\zeta) + \gamma(\zeta) \dot{\phi}(\zeta)$$

and we therefore obtain $\mathbf{i}\varepsilon_{zy}(\zeta) = \dot{\gamma}(\zeta) \phi(\zeta) + \gamma(\zeta) \dot{\phi}(\zeta)$. Now, calculating $\mathbf{i}\varepsilon_{zy}(\zeta) - \mathbf{i}\varepsilon_{xy}(\zeta)$, we obtain $\mathbf{i}\varepsilon_{(z-x)y}(\zeta) = \gamma(\zeta) \dot{\phi}(\zeta)$, which is Equation (47). Note that, although the differentiation operation causes a loss of information (as derivatives are unaffected by constant shifts), the whole system of two equations does not suffer from this loss because we keep the original equation $\varepsilon_y(\zeta) = \gamma(\zeta) \phi(\zeta)$ as part of the system. Also, since generalized functions are closed under differentiation, all quantities involved are well-defined generalized functions. We will now use Equations (46) and (47) to show identification.

It is only possible to have $\bar{\zeta} = 0$ when $g(x^*)$ is a polynomial, a case which has already been shown to be identified (Hausman, Newey, Ichimura, and Powell (1991)). Hence, we focus on the case where $\bar{\zeta} > 0$.

For $|\zeta| > \bar{\zeta}$, the fact that $\gamma(\zeta) = 0$ can be directly inferred from Equation (46) and the fact that $\varepsilon_y(\zeta) = 0$, since $|\phi(\zeta)| > 0$, as stated in the first part of Equation (12).

We next focus on $|\zeta| \leq \bar{\zeta}$. As mentioned in Section 2, any generalized function (such as $\gamma(\zeta)$) can be decomposed as the sum of an ordinary function, denoted by an “o” subscript (e.g. $\gamma_o(\zeta)$), and a purely singular component, denoted by an “s” subscript (e.g. $\gamma_s(\zeta)$), which consists of a linear combination of delta function derivatives. Decomposing $\varepsilon_y(\zeta)$ and $\varepsilon_{(z-x)y}(\zeta)$ in a similar fashion and substituting these decompositions into Equations (46) and (47) yields

$$\varepsilon_{y,o}(\zeta) + \varepsilon_{y,s}(\zeta) = (\gamma_o(\zeta) + \gamma_s(\zeta))\phi(\zeta) \quad (48)$$

$$\mathbf{i}\varepsilon_{(z-x)y,o}(\zeta) + \mathbf{i}\varepsilon_{(z-x)y,s}(\zeta) = (\gamma_o(\zeta) + \gamma_s(\zeta))\dot{\phi}(\zeta). \quad (49)$$

Since the product of an ordinary function with an ordinary function is an ordinary function, while the product of a purely singular component with an ordinary function is purely singular (as shown in Lemma 2), Equations (48) and (49) imply that

$$\varepsilon_{y,o}(\zeta) = \gamma_o(\zeta)\phi(\zeta) \quad (50)$$

$$\mathbf{i}\varepsilon_{(z-x)y,o}(\zeta) = \gamma_o(\zeta)\dot{\phi}(\zeta). \quad (51)$$

Since all quantities are now ordinary functions (including $\dot{\phi}(\zeta)$, since $E[|U|] < \infty$), Equations (50) and (51) can now be manipulated according to the usual rules of multiplication and division. Under the assumption that $\phi(\zeta) \neq 0$, and for any ζ such that $\gamma_o(\zeta) \neq 0$, we can divide each side of Equation (51) by the corresponding side of Equation (50) to obtain

$$\frac{\dot{\phi}(\zeta)}{\phi(\zeta)} = \frac{\mathbf{i}\varepsilon_{(z-x)y,o}(\zeta)}{\varepsilon_{y,o}(\zeta)}. \quad (52)$$

This equation holds almost everywhere in $[-\bar{\zeta}, \bar{\zeta}]$, since the assumption that $\gamma(\zeta) \neq 0$ almost everywhere in $[-\bar{\zeta}, \bar{\zeta}]$ also implies that $\gamma_o(\zeta) \neq 0$ almost everywhere in $[-\bar{\zeta}, \bar{\zeta}]$. By Lemma 3 in the Appendix and the assumption that $E[|U|] < \infty$, both $\dot{\phi}(\zeta)$ and $\phi(\zeta)$ are continuous. Since $\phi(\zeta) \neq 0$ for all $\zeta \in \mathbb{R}$ by assumption, the ratio $\dot{\phi}(\zeta)/\phi(\zeta)$ is continuous everywhere. Since Equation (52) holds almost everywhere in $[-\bar{\zeta}, \bar{\zeta}]$ and since $\dot{\phi}(\zeta)/\phi(\zeta)$ is continuous, the ratio $\mathbf{i}\varepsilon_{(z-x)y,o}(\zeta)/\varepsilon_{y,o}(\zeta)$ contains no

essential singularity and its value can be defined everywhere in $[-\bar{\zeta}, \bar{\zeta}]$ by taking limits (that is, we take the convention that $\mathbf{i}\varepsilon_{(z-x)y,o}(\xi)/\varepsilon_{y,o}(\xi)$ is a shorthand notation for $\lim_{\xi^* \rightarrow \xi} \mathbf{i}\varepsilon_{(z-x)y,o}(\xi^*)/\varepsilon_{y,o}(\xi^*)$). With this convention, Equation (52) holds for all $\zeta \in [-\bar{\zeta}, \bar{\zeta}]$.

Integrating each side of Equation (52) with respect to ζ , yields

$$\ln \phi(\zeta) - \ln \phi(0) = \int_0^\zeta \frac{\mathbf{i}\varepsilon_{(z-x)y,o}(\xi)}{\varepsilon_{y,o}(\xi)} d\xi$$

for $|\zeta| < \bar{\zeta}$. Making use of the boundary condition $\phi(0) = \int e^{\mathbf{i}0u} dF(u) = \int dF(u) = 1$, and taking exponentials on each side, we obtain Equation (13) stated in Theorem 1, which provides the value of $\phi(\zeta)$ for $|\zeta| \leq \bar{\zeta}$ in terms of observable quantities.

Next, multiplying each side of Equation (8) by $(\phi(\zeta))^{-1}$ establishes that, for $|\zeta| \leq \bar{\zeta}$,

$$\gamma(\zeta) = \frac{\varepsilon_y(\zeta)}{\phi(\zeta)} \quad (53)$$

where $\phi(\zeta)$ is known from Equation (13). This operation is justified because (i) $\phi(\zeta) \neq 0$ by assumption, (ii) multiplication of a generalized function by the ordinary function $((\phi(\zeta))^{-1})$ is allowed, provided that the ordinary function admits a sufficient number of continuous derivatives, which is the case here, since the result of this operation, $\gamma(\zeta)$, is a well-defined generalized function, by Assumption 1. Substituting Equation (13) into Equation (53) yields

$$\gamma(\zeta) = \varepsilon_y(\zeta) \exp\left(-\int_0^\zeta \frac{\mathbf{i}\varepsilon_{(z-x)y,o}(\xi)}{\varepsilon_{y,o}(\xi)} d\xi\right), \quad (54)$$

for $|\zeta| \leq \bar{\zeta}$, which is the second part of Equation (12). Finally, $g(x^*)$ is given by the inverse Fourier transform of $\gamma(\zeta)$. ■

Proof of Corollary 1. Since Equations (48)–(54) in the proof of Theorem 1 hold over any interval containing the origin and over which $\gamma(\zeta, \theta^*) \neq 0$ almost everywhere and $\phi(\zeta) \neq 0$, we can conclude that $\gamma(\zeta, \theta^*)$ is identified for $\zeta \in]-\zeta^*, \zeta^*[$, for ζ^* defined in Assumption 4. By Assumption 4, there exist no $\theta \in \Theta$ such that $\theta \neq \theta^*$ and

$\gamma(\zeta, \theta^*) = \gamma(\zeta, \theta)$ for all $\zeta \in]-\zeta^*, \zeta^*[$, hence the knowledge of the function $\gamma(\zeta, \theta^*)$ for $\zeta \in]-\zeta^*, \zeta^*[$ uniquely determines θ^* . ■

Lemma 2 *If $\phi(\zeta)$ is k times continuously differentiable at $\zeta = 0$, then $\delta^{(k)}(\zeta) \phi(\zeta) = (-1)^k \sum_{j=0}^k \binom{k}{j} (d^{k-j} \phi(0) / d\zeta^{k-j}) \delta^{(j)}(\zeta)$.*

Proof. This is shown by calculating $\int \delta^{(k)}(\zeta) \phi(\zeta) \psi(\zeta) d\zeta$ for any test function ψ , by using k repeated integrations by parts (see Supplementary Material for details). ■

While the following Lemma resembles a well-known result regarding characteristic functions (see Loève (1977), following Property 13.1), it generalizes it to apply to Fourier transforms of any absolutely integrable function.

Lemma 3 *If $s(z)$ is absolutely integrable, then its Fourier transform $\sigma(\zeta)$ is continuous. In particular, if $\int |z|^k |s(z)| dz < \infty$ for some $k \in \mathbb{N}$, then $d^k \sigma(\zeta) / d\zeta^k$ is continuous.*

Proof. First, $\sigma_j(\zeta) = \int e^{i\zeta z} s(z) 1(|z| \leq j) dz$ is continuous in ζ for every j :

$$\begin{aligned} |\sigma_j(\zeta) - \sigma_j(\xi)| &= \left| \int (e^{i\zeta z} - e^{i\xi z}) s(z) 1(|z| \leq j) dz \right| \\ &= \left| \int e^{i(\zeta+\xi)z/2} 2 \sin((\zeta - \xi)z/2) s(z) 1(|z| \leq j) dz \right| \\ &\leq \int |2 \sin((\zeta - \xi)z/2)| |s(z)| 1(|z| \leq j) dz \\ &\leq \int |(\zeta - \xi)z| |s(z)| 1(|z| \leq j) dz \\ &\leq |\zeta - \xi| j \int |s(z)| 1(|z| \leq j) dz \leq |\zeta - \xi| j \int |s(z)| dz. \end{aligned}$$

Next, observe that $\sigma_j(\zeta)$ converges to $\sigma(\zeta)$ uniformly in ζ because $|\sigma(\zeta) - \sigma_j(\zeta)| = \left| \int e^{i\zeta z} s(z) (1 - 1(|z| \leq j)) dz \right| \leq \int |e^{i\zeta z}| |s(z)| (1 - 1(|z| \leq j)) dz = \int_{|z| \geq j} |s(z)| dz \rightarrow 0$ as $j \rightarrow \infty$. Since $\sigma_j(\zeta)$ is a sequence of continuous functions converging uniformly to $\sigma(\zeta)$, the limiting function $\sigma(\zeta)$ must be continuous. The second assertion then follows from the fact that the Fourier transform of $(iz)^k s(z)$ is $d^k \sigma(\zeta) / d\zeta^k$. ■

Lemma 4 Under Assumption 1 and 5, $\gamma(\zeta, \theta)$, $\dot{\gamma}(\zeta, \theta)$, $\varepsilon_y(\zeta)$ and $\varepsilon_{xy}(\zeta)$ admit the decompositions²⁸

$$\gamma(\zeta, \theta) = \gamma_o(\zeta, \theta) + 2\pi \sum_{k=0}^{\bar{k}} \gamma_k(\theta) (-\mathbf{i})^k \delta^{(k)}(\zeta) \quad (55)$$

$$\dot{\gamma}(\zeta, \theta) = \dot{\gamma}_o(\zeta, \theta) + 2\pi \sum_{k=-1}^{\bar{k}} \gamma_k(\theta) (-\mathbf{i})^k \delta^{(k+1)}(\zeta) \quad (56)$$

$$\varepsilon_y(\zeta) = \varepsilon_{y,o}(\zeta) + 2\pi \sum_{k=0}^{\bar{k}} \varepsilon_{y,k} (-\mathbf{i})^k \delta^{(k)}(\zeta) \quad (57)$$

$$\mathbf{i}\varepsilon_{xy}(\zeta) = \mathbf{i}\varepsilon_{xy,o}(\zeta) + 2\pi \mathbf{i} \sum_{k=-1}^{\bar{k}} \varepsilon_{xy,k} (-\mathbf{i})^{k+1} \delta^{(k+1)}(\zeta), \quad (58)$$

where "o" subscripts denote ordinary functions. Moreover, Equations (8) and (9), are equivalent to

$$\varepsilon_{y,o}(\zeta) = \gamma_o(\zeta, \theta) \phi(\zeta) \quad (59)$$

$$\mathbf{i}\varepsilon_{xy,o}(\zeta) = \dot{\gamma}_o(\zeta, \theta) \phi(\zeta) \quad (60)$$

$$\Sigma_y = \Gamma_y(\theta) \Phi \quad (61)$$

$$\Sigma_{xy} = \Gamma_{xy}(\theta) \Phi \quad (62)$$

where the $(\bar{k} + 1) \times 1$ vectors Φ , Σ_y , Σ_{xy} are given by²⁹

$$\Phi = \left(\phi(0), -\mathbf{i} \frac{d\phi(0)}{d\zeta}, \dots, (-\mathbf{i})^{\bar{k}} \frac{d^{\bar{k}}\phi(0)}{d\zeta^{\bar{k}}} \right)' \quad (63)$$

$$\Sigma_y = (\varepsilon_{y,0}, \dots, \varepsilon_{y,\bar{k}})' \quad (64)$$

$$\Sigma_{xy} = (\varepsilon_{xy,0}, \dots, \varepsilon_{xy,\bar{k}})' \quad (65)$$

and where the elements of the $(\bar{k} + 1) \times (\bar{k} + 1)$ matrices $\Gamma_y(\theta)$ and $\Gamma_{xy}(\theta)$ are given by Equations (32) and (33). Note that Equations (61) and (62) are the matrix analogues of Equations (3.6) and (3.9) in Hausman, Newey, Ichimura, and Powell (1991).

²⁸Note that if $\gamma_o(\zeta, \theta)$ has a step discontinuity at $\zeta = 0$, $\partial\gamma_o(\zeta, \theta)/\partial\zeta$ will contain a delta function. Hence, we define $\dot{\gamma}_o(\zeta, \theta)$ to be the ordinary part of $\partial\gamma_o(\zeta, \theta)/\partial\zeta$ and $\gamma_{-1}(\zeta)$ contains the magnitude of the step in $\gamma_o(\zeta, \theta)$. The term $\gamma_{-1}(\theta)\delta(\zeta)$ will actually never be needed in the estimation procedure we propose.

²⁹Note that the vector Σ_{xy} does not contain the element $\varepsilon_{xy,-1}$ because it brings no additional information for the purpose of identifying θ . Including $\varepsilon_{xy,-1}$ would add an additional equation but also an additional unknown, namely $d^{\bar{k}+1}\phi(0)/d\zeta^{\bar{k}+1}$.

Proof of Lemma 4. We prove the equivalence for Equation (8) — the corresponding result for Equation (9) follows similarly (see Supplementary Material). Substituting Equations (25) through (58) into Equation (8), we obtain

$$\varepsilon_{y,o}(\zeta) + 2\pi \sum_{k=0}^{\bar{k}} \varepsilon_{y,k} (-\mathbf{i})^k \delta^{(k)}(\zeta) = \left(\gamma_o(\zeta, \theta) + 2\pi \sum_{k=0}^{\bar{k}} \gamma_k(\theta) (-\mathbf{i})^k \delta^{(k)}(\zeta) \right) \phi(\zeta).$$

Equating the ordinary functions part of each expression yields

$$\varepsilon_{y,o}(\zeta) = \gamma_o(\zeta, \theta) \phi(\zeta),$$

while equating the singular parts yields

$$\sum_{k=0}^{\bar{k}} \varepsilon_{y,k} (-\mathbf{i})^k \delta^{(k)}(\zeta) = \sum_{k=0}^{\bar{k}} \gamma_k(\theta) (-\mathbf{i})^k \delta^{(k)}(\zeta) \phi(\zeta).$$

By Lemma 2, we have

$$\sum_{k=0}^{\bar{k}} \varepsilon_{y,k} (-\mathbf{i})^k \delta^{(k)}(\zeta) = \sum_{k=0}^{\bar{k}} \gamma_k(\theta) (-\mathbf{i})^k \sum_{j=0}^k \binom{k}{j} \phi^{(k-j)}(0) \delta^{(j)}(\zeta).$$

Simple manipulations then give

$$\sum_{k=0}^{\bar{k}} \varepsilon_{y,k} (-\mathbf{i})^k \delta^{(k)}(\zeta) = \sum_{j=0}^{\bar{k}} \sum_{k=0}^{\bar{k}} \binom{k}{j} \gamma_k(\theta) (-\mathbf{i})^k \mathbf{1}(j \leq k) \phi^{(k-j)}(0) \delta^{(j)}(\zeta)$$

Equating the coefficients of the delta function derivatives of the same order gives

$$\begin{aligned} \varepsilon_{y,j} &= \mathbf{i}^j \sum_{k=0}^{\bar{k}} \binom{k}{j} \gamma_k(\theta) (-\mathbf{i})^k \mathbf{1}(j \leq k) \phi^{(k-j)}(0) \\ &= \mathbf{i}^j \sum_{l=-j}^{\bar{k}-j} \binom{j+l}{j} \gamma_{j+l}(\theta) (-\mathbf{i})^{j+l} \mathbf{1}(0 \leq l) \phi^{(l)}(0) \\ &= \sum_{l=0}^{\bar{k}-j} \binom{j+l}{j} \gamma_{j+l}(\theta) (-\mathbf{i})^l \phi^{(l)}(0) \\ &= \sum_{l=0}^{\bar{k}} \binom{j+l}{j} \gamma_{j+l}(\theta) (-\mathbf{i})^l \mathbf{1}(l \leq \bar{k} - j) \phi^{(l)}(0), \end{aligned}$$

which is equivalent to Equation (61). ■

Lemma 5 Let $\lambda(\zeta)$ be (i) such that $d^k(\zeta^k \lambda(\zeta))/d\zeta^k$ exists for all $k \in \mathbb{N}$, (ii) supported on $[-\eta, \eta]$ for some $\eta \in]0, \infty]$, (iii) such that $\int \lambda(\zeta) d\zeta = c \in \mathbb{R}$ and (iv) $\int |\lambda(\zeta)| d\zeta < \infty$. If the moment generating function of the distribution of U exists over $[-\eta - \epsilon, \eta + \epsilon]$, then $\phi(\zeta)$, the characteristic function of U , is such that $\int \mu(\zeta) \phi(\zeta) d\zeta = c\phi(0)$ for

$$\mu(\zeta) = \sum_{k=0}^{\infty} \frac{1}{k!} \frac{d^k}{d\zeta^k} (\zeta^k \lambda(\zeta)). \quad (66)$$

Proof of Lemma 5. If the moment generating function of the distribution of U exists over an interval $[-\eta - \epsilon, \eta + \epsilon]$, then the characteristic function $\phi(\zeta)$ will be analytic in a strip $|\text{Im } \zeta| \leq \eta + \epsilon$ in the complex plane (Lukacs (1970), Theorem 7.1.1 and Corollary 7.1.1). It follows that the Taylor series of $\phi(0)$, expanded around some $\zeta \in [-\eta, \eta]$,

$$\phi(0) = \sum_{k=0}^{\infty} \frac{(-\zeta)^k}{k!} \frac{d^k \phi(\zeta)}{d\zeta^k},$$

is convergent. After multiplying by $\lambda(\zeta)$ and integrating over ζ , we have

$$\int_{-\eta}^{\eta} \lambda(\zeta) \phi(0) d\zeta = \int_{-\eta}^{\eta} \sum_{k=0}^{\infty} \frac{(-\zeta)^k}{k!} \frac{d^k \phi(\zeta)}{d\zeta^k} \lambda(\zeta) d\zeta, \quad (67)$$

where the left-hand side is $c\phi(0)$ since $\int_{-\eta}^{\eta} \lambda(\zeta) d\zeta = c$. The integral and summation can be interchanged by Fubini's Theorem since $\int |\lambda(\zeta)| d\zeta < \infty$ by assumption and since the Taylor series of $\phi(\zeta)$ is absolutely summable for $\zeta \in [-\eta, \eta]$, by the Moment Theorem ($d^k \phi(\zeta)/d\zeta^k = \mathbf{i}^k E[U^k]$) and Jensen's inequality:

$$\sum_{k=0}^{\infty} \left| \frac{(-\zeta)^k}{k!} \frac{d^k \phi(\zeta)}{d\zeta^k} \right| \leq \sum_{k=0}^{\infty} \frac{\eta^k}{k!} E[|U|^k] = E[\exp(\eta|U|)] \leq E[\exp(\eta U)] + E[\exp(-\eta U)].$$

After integrating the right-hand side by parts, we obtain

$$c\phi(0) = \sum_{k=0}^{\infty} \int_{-\eta}^{\eta} \frac{1}{k!} \left(\frac{d^k}{d\zeta^k} (\zeta^k \lambda(\zeta)) \right) \phi(\zeta) d\zeta,$$

or $\int \mu(\zeta) \phi(\zeta) d\zeta = c\phi(0)$, where $\mu(\zeta)$ is given by Equation (66). ■

Proof of Theorem 2. Lemma 4 shows that Equations (8) and (9) can be separated into Equations (59) and (60), involving the ordinary function components, and Equations (61) and (62), involving the purely singular components. We will use these expressions to construct moment conditions that must hold for $\theta = \theta^*$, the true value of θ .

Calculation of $r_{y,o}(z, \theta)$ and $r_{xy,o}(z, \theta)$. The equations involving the ordinary function components can be combined to eliminate $\phi(\zeta)$ and yield

$$\varepsilon_{y,o}(\zeta) \dot{\gamma}_o(\zeta, \theta) = \mathbf{i} \varepsilon_{xy,o}(\zeta) \gamma_o(\zeta, \theta). \quad (68)$$

Multiplying each side by a vector of weighting functions $\omega(\zeta)$ chosen as specified in Section 3.2.1 and integrating over ζ yields

$$\int \varepsilon_{y,o}(\zeta) \dot{\gamma}_o(\zeta, \theta) \omega(\zeta) d\zeta = \int \mathbf{i} \varepsilon_{xy,o}(\zeta) \gamma_o(\zeta, \theta) \omega(\zeta) d\zeta. \quad (69)$$

Since $\omega(\zeta)$ is chosen such that $\omega(\zeta) \gamma_o(\zeta, \theta)$ has $\bar{k} + 1$ vanishing derivatives at the origin, we have, for $j = 0, \dots, \bar{k} + 1$,

$$\int \delta^{(j)}(\zeta) \gamma_o(\zeta, \theta) \omega(\zeta) d\zeta = 0.$$

It follows that $\varepsilon_{xy,o}(\zeta)$ can be replaced by $\varepsilon_{xy}(\zeta)$ in Equation (69). Using the same reasoning for the left-hand side of Equation (69), we obtain

$$\int \varepsilon_y(\zeta) \dot{\gamma}_o(\zeta, \theta) \omega(\zeta) d\zeta = \int \mathbf{i} \varepsilon_{xy}(\zeta) \gamma_o(\zeta, \theta) \omega(\zeta) d\zeta. \quad (70)$$

Using the same steps as in Section 3.2.1, this equation can be converted, using Parseval's identity, to an inverse-density-weighted expectation of the form

$$E \left[\frac{Y r_{y,o}(Z, \theta) + XY r_{xy,o}(Z, \theta)}{p(Z)} \right] = 0$$

with $r_{y,o}(z, \theta)$ and $r_{xy,o}(z, \theta)$ given by Equations (29) and (30), respectively.

Calculation of $r_{1y,o}(z, \theta)$. Equation (8) can be rewritten as

$$\frac{\varepsilon_{y,o}(\zeta)}{\gamma_o(\zeta, \theta)} = \phi(\zeta) \quad (71)$$

(defining points were $\gamma_o(\zeta, \theta) = 0$ by taking limits). By Lemma 5 (and Assumption 7), for any function $\varpi \in \mathcal{S}_1$ (as chosen in Section 3.2.1) we have that $\int \varpi(\zeta) \phi(\zeta) d\zeta = \phi(0)$, which is equal to 1 by the properties of characteristic functions. Applying this result to Equation (71) yields

$$\int \frac{\varpi(\zeta)}{\gamma_o(\zeta, \theta)} \varepsilon_{y,o}(\zeta) d\zeta = 1.$$

Since $\varpi(\zeta)$ is also chosen so that $\varpi(\zeta)/\gamma_o(\zeta, \theta)$ has \bar{k} vanishing derivatives at the origin, we can replace $\varepsilon_{y,o}(\zeta)$ by $\varepsilon_y(\zeta)$ without changing the value of the integral. Now, using Parseval's identity, we obtain

$$\int r_{1y,o}(z, \theta) E[Y|Z=z] dz = 1 \tag{72}$$

where $r_{1y,o}(z, \theta)$ is given by Equation (31). After using iterated expectations, Equation (72) can be written as $E[Y r_{1y,o}(Z, \theta) / p(Z) - 1] = 0$.

Calculation of $r_{y,s}(z, \theta)$ and $r_{xy,s}(z, \theta)$. Combining Equations (61) and (62) in Lemma 4 to eliminate Φ yields $\Gamma_y^{-1}(\theta) \Sigma_y = \Gamma_{xy}^{-1}(\theta) \Sigma_{xy}$ where the matrices $\Gamma_y(\theta)$ and $\Gamma_{xy}(\theta)$ are invertible, since they have a triangular structure with nonzero elements on the diagonal.³⁰ This expression provides $\bar{k} + 1$ restrictions, but in Section 3.2.2, we allow for the possibility that the number of degrees of freedom (excluding the scale) N_s that the singular component possesses may be less than $\bar{k} + 1$, so all restrictions may not be needed. Hence, we extract N_s equations by multiplying each side by a selection matrix S_{2, N_s+1} (defined in Section 3.2.2) to obtain:

$$S_{2, N_s+1} \Gamma_y^{-1}(\theta) \Sigma_y = S_{2, N_s+1} \Gamma_{xy}^{-1}(\theta) \Sigma_{xy}. \tag{73}$$

Note that the first element is also eliminated in the construction of $r_{y,s}(z, \theta)$ and $r_{xy,s}(z, \theta)$, because it will be used in the construction of $r_{1y,s}(z, \theta)$ below. We now

³⁰The determinant of a triangular matrix is equal to the product of the diagonal elements. Due to our convention of indices, the “diagonal” elements have indices $(\bar{k} + 1, 1), (\bar{k}, 2), \dots, (1, \bar{k} + 1)$ instead of the usual $(1, 1), (2, 2), \dots, (\bar{k} + 1, \bar{k} + 1)$ but reordering the rows does not change the magnitude of the determinant.

show that the functions $r_{y,s}(z, \theta)$ and $r_{xy,s}(z, \theta)$ constructed in Section 3.2.2 define moment conditions imposing the same constraints as Equation (73). First, for $r_{y,s}(z, \theta)$ as defined in Equation (40), we have that $E[Yr_{y,s}(Z, \theta)/p(Z)]$ is equal to

$$\int r_{y,s}(z, \theta) E[Y|Z=z] dz = S_{2,N_s+1} \Gamma_y^{-1}(\theta) M_y^{-1}(\theta) \int V_y(z, \theta) E[Y|Z=z] dz \quad (74)$$

where, by Parseval's identity, the rightmost integral is equal to $(2\pi)^{-1} \int \nu_y(\zeta, \theta) \varepsilon_y(\zeta) d\zeta$.

Replacing the functions $\nu_y(\zeta, \theta)$ and $\varepsilon_y(\zeta)$ by their full expressions from Equations (34) and (57) gives:

$$(2\pi)^{-1} \int \frac{\mu_y(\zeta)}{\gamma_o(\zeta, \theta)} \varepsilon_{y,o}(\zeta) d\zeta + \int \frac{\mu_y(\zeta)}{\gamma_o(\zeta, \theta)} \sum_{k=0}^{\bar{k}} \varepsilon_{y,k}(-\mathbf{i})^k \delta^{(k)}(\zeta) d\zeta \quad (75)$$

The first integral is equal to $\int \mu_y(\zeta) \varepsilon_{y,o}(\zeta) / \gamma_o(\zeta, \theta) d\zeta = \int \mu_y(\zeta) \phi(\zeta) d\zeta$ by Equation (59). Since each element of μ_y is a function in \mathcal{S}_0 (as chosen in Section 3.2.2), we have, by Lemma 5 (and Assumption 7), that $\int \mu_y(\zeta) \phi(\zeta) d\zeta = 0 \cdot \phi(\zeta) = 0$ and the first integral in Equation (75) vanishes. The second integral in Equation (75) is equal to

$$\sum_{k=0}^{\bar{k}} \left[\frac{d^k}{d\zeta^k} \left(\frac{\mu_y(\zeta)}{\gamma_o(\zeta, \theta)} \right) \right]_{\zeta=0} (\mathbf{i})^k \varepsilon_{y,k} \quad (76)$$

by the properties of $\delta^{(k)}(\zeta)$. Using the definition of $M_y(\theta)$ (from Equations (36) and (34)) and Σ_y (from Equation (64)), Equation (76) can be rewritten as $M_y(\theta) \Sigma_y$. Substituting this expression into Equation (74) gives

$$S_{2,N_s+1} \Gamma_y^{-1}(\theta) M_y^{-1}(\theta) (M_y(\theta) \Sigma_y) = S_{2,N_s+1} (\Gamma_y(\theta))^{-1} \Sigma_y \quad (77)$$

which is exactly the left-hand side of Equation (73). Proceeding similarly for the calculation of $E[XYr_{xy,s}(Z, \theta)/p(Z)]$ yields the right-hand side of Equation (73). Indeed, defining $\Sigma_{xy}^* = (\varepsilon_{xy,-1}, \dots, \varepsilon_{xy,\bar{k}})'$, and noting that $S_{2,\bar{k}+2} \Sigma_{xy}^* = \Sigma_{xy}$, we have, by Equation (41),

$$\begin{aligned} -E \left[\frac{XYr_{xy,s}(Z, \theta)}{p(Z)} \right] &= S_{2,N_s+1} \Gamma_{xy}^{-1}(\theta) S_{2,\bar{k}+2} M_{xy}^{-1}(\theta) \int V_{xy}(z, \theta) E[XY|Z=z] dz \\ &= S_{2,N_s+1} \Gamma_{xy}^{-1}(\theta) S_{2,\bar{k}+2} M_{xy}^{-1}(\theta) M_{xy}(\theta) \Sigma_{xy}^* \\ &= S_{2,N_s+1} \Gamma_{xy}^{-1}(\theta) S_{2,\bar{k}+2} \Sigma_{xy}^* = S_{2,N_s+1} \Gamma_{xy}^{-1}(\theta) \Sigma_{xy} \end{aligned}$$

which is the right-hand side of Equation (73).

Calculation of $r_{1y,s}(z, \theta)$. We first note that Equation (61) can be written as $\Phi = \Gamma_y^{-1}(\theta) \Sigma_y$. Since the first elements of Φ is $\phi(0) = 1$, we can write

$$S_{1,1} \Gamma_y^{-1}(\theta) \Sigma_y = 1. \quad (78)$$

We next proceed as in the calculation of $r_{y,s}(z, \theta)$, showing that $E[Y r_{1y,s}(Z, \theta) / p(Z)] = 1$, with $r_{1y,s}(z, \theta)$ given by Equation (42), is equivalent to Equation (78). We have, by Equation (42) and following the same steps as Equations (74)-(77),

$$\begin{aligned} E \left[\frac{Y r_{1y,s}(Z, \theta)}{p(Z)} \right] &= S_{1,1} \Gamma_y^{-1}(\theta) M_y^{-1}(\theta) \int V_y(z, \theta) E[Y|Z=z] dz \\ &= S_{1,1} \Gamma_y^{-1}(\theta) M_y^{-1}(\theta) M_y(\theta) \Sigma_y = S_{1,1} \Gamma_y^{-1}(\theta) \Sigma_y. \end{aligned}$$

We have now shown that, at the true value θ^* of θ , the moment conditions $E[Q(X, Y, Z, \theta, p)] = 0$ defined in Section 3.2 are satisfied and that we can obtain as many equations as there are elements in θ . Under Assumption 6, $E[\partial Q(X, Y, Z, \theta, p) / \partial \theta']$ is non-singular at $\theta = \theta^*$ and θ^* is a “regular” point since $E[\partial Q(X, Y, Z, \theta, p) / \partial \theta']$ has constant (full) rank in a neighborhood of θ^* . Hence, θ^* is the unique solution to $E[Q(X, Y, Z, \theta, p)] = 0$ is some open neighborhood $\tilde{\Theta}$ of θ^* (e.g., see Theorem 5.1.1 in Hsiao (1983)). Hence, taking Θ to be some compact set contained in $\tilde{\Theta}$ and containing θ^* in its interior completes the proof of the Theorem. ■

References

- AI, C., AND X. CHEN (2003): “Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions,” *Econometrica*, 71, 1795–1843.
- AMEMIYA, Y. (1985): “Instrumental Variable Estimator for the Nonlinear Errors-in-Variables Model,” *Journal of Econometrics*, 28, 273–289.
- ANDREWS, D. W. K. (1995): “Nonparametric Kernel Estimation for Semiparametric Models,” *Econometric Theory*, 11, 560–596.

- CARRASCO, M., AND J.-P. FLORENS (2002): “Efficient GMM Estimation Using the Empirical Characteristic Function,” Working Paper, University of Rochester.
- CARROLL, R. J., D. RUPPERT, AND L. A. STEFANSKI (1995): *Measurement Error in Nonlinear Models*. New York: Chapman & Hall.
- CHEN, X. (2005): “Large Sample Sieve Estimation of Semi-Nonparametric Models,” in *Handbook of Econometrics*, vol. Vol. 6. Elsevier Science.
- CHEN, X., H. HONG, AND E. TAMER (2005): “Measurement Error Models with Auxiliary Data,” *Review of Economic Studies*, 72, 343–366.
- CHERNOZHUKOV, V., AND C. HANSEN (2005): “An IV Model of Quantile Treatment Effects,” *Econometrica*, 73, 245–261.
- CHERNOZHUKOV, V., G. W. IMBENS, AND W. K. NEWEY (2006): “Instrumental Variable Identification and Estimation of Nonseparable Models via Quantile Conditions,” *Journal of Econometrics*, forthcoming.
- CHESHER, A. (1991): “The Effect of Measurement Error,” *Biometrika*, 78, 451.
- (1998): “Polynomial Regression with Covariate Measurement Error,” Discussion Paper 98/448, University of Bristol.
- (2003): “Identification in Nonseparable Models,” *Econometrica*, 71, 1405–1441.
- (2005): “Nonparametric Identification under Discrete Variation,” *Econometrica*, 73, 1525–1550.
- DAROLLES, S., J.-P. FLORENS, AND E. RENAULT (2002): “Nonparametric Instrumental Regression,” Working Paper 05-2002, Centre de Recherche et Développement en Économie.
- FAN, J. (1991): “On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems,” *Annals of Statistics*, 19(3), 1257–1272.
- FAN, J., AND Y. K. TRUONG (1993): “Nonparametric Regression with Errors in Variables,” *Annals of Statistics*, 21(4), 1900–1925.

- GEL'FAND, I. M., AND G. E. SHILOV (1964): *Generalized Functions*. Academic Press, New York.
- HARDLE, W., AND T. STOKER (1989): "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, 84, 986–995.
- HAUSMAN, J., W. NEWEY, H. ICHIMURA, AND J. POWELL (1991): "Measurement Errors in Polynomial Regression Models," *Journal of Econometrics*, 50, 273–295.
- HAUSMAN, J., W. NEWEY, AND J. POWELL (1995): "Nonlinear Errors in Variables. Estimation of Some Engel Curves," *Journal of Econometrics*, 65, 205–233.
- HOROWITZ, J. L., AND M. MARKATOU (1996): "Semiparametric Estimation of Regression Models for Panel Data," *Review of Economic Studies*, 63, 145.
- HSIAO, C. (1983): "Identification," in *Handbook of Econometrics*, ed. by Z. Griliches, and M. Intriligator, vol. I. Elsevier Science.
- (1989): "Consistent Estimation for Some Nonlinear Errors-in-Variables Models," *Journal of Econometrics*, 41, 159–185.
- HSIAO, C., AND Q. WANG (2000): "Estimation of Structural Nonlinear Errors-in-Variables Models by Simulated Least Squares Method," *International Economic Review*, 41, 523–542.
- HU, Y., AND G. RIDDER (2004): "Estimation of Nonlinear Models with Measurement Error Using Marginal Information," Working Paper, University of Southern California, Department of Economics.
- HYSLOP, D. R., AND G. W. IMBENS (2001): "Bias from classical and other forms of measurement error," *Journal of Business & Economic Statistics*, 19, 475–481.
- IMBENS, G. W., AND W. K. NEWEY (2003): "Identification and Estimation of Triangular Simultaneous Equations Models without Additivity," Working Paper, MIT, Department of Economics.
- LAVERGE, P., AND Q. H. VUONG (1996): "Nonparametric Selection of Regressors: The Nonnested Case," *Econometrica*, 64, 207–219.

- LEWBEL, A. (1996): “Demand Estimation with Expenditure Measurement Errors on the Left and Right Hand Side,” *The Review of Economics and Statistics*, 78(4), 718–725.
- (1998): “Semiparametric Latent Variable Model Estimation with Endogenous or Mismeasured Regressors,” *Econometrica*, 66, 105–121.
- LI, T. (2002): “Robust and consistent estimation of nonlinear errors-in-variables models,” *Journal of Econometrics*, 110, 1–26.
- LI, T., AND Q. VUONG (1998): “Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators,” *Journal of Multivariate Analysis*, 65, 139–165.
- LIGHTHILL, M. J. (1962): *Introduction to Fourier Analysis and Generalized Function*. London: Cambridge University Press.
- LOÈVE, M. (1977): *Probability Theory I*. New York: Springer.
- LUKACS, E. (1970): *Characteristic Functions*. Griffin, London, second edn.
- NEWBY, W. (1994): “The Asymptotic Variance of Semiparametric Estimators,” *Econometrica*, 62, 1349–1382.
- (2001): “Flexible Simulated Moment Estimation of Nonlinear Errors-in-Variables Models,” *Review of Economics and Statistics*, 83, 616–627.
- NEWBY, W., AND D. MCFADDEN (1994): “Large Sample Estimation and Hypothesis Testing,” in *Handbook of Econometrics*, ed. by R. F. Engel, and D. L. McFadden, vol. IV. Elsevier Science.
- NEWBY, W. K. (1997): “Convergence rates and asymptotic normality of series estimators,” *Journal of Econometrics*, 79, 147–168.
- NEWBY, W. K., AND J. L. POWELL (2003): “Instrumental Variable Estimation of Nonparametric Models,” *Econometrica*, 71, 1565–1578.
- PAGAN, A., AND A. ULLAH (1999): *Nonparametric Econometrics*. Cambridge University Press, Cambridge, UK.

- POWELL, J., J. STOCK, AND T. STOKER (1989): “Semiparametric Estimation of Index Coefficients,” *Econometrica*, 57, 1403–1430.
- RAO, P. (1992): *Identifiability in Stochastic Models*. New York: Academic Press.
- SCHENNACH, S. M. (2004a): “Estimation of Nonlinear Models with Measurement Error,” *Econometrica*, 72, 33–75.
- (2004b): “Exponential specifications and measurement error,” *Economics Letters*, 85, 85–91.
- (2004c): “Nonparametric Estimation in the Presence of Measurement Error,” *Econometric Theory*, 20, 1046–1093.
- (2005): “Quantile regression with mismeasured covariates,” Working Paper, University of Chicago, <http://home.uchicago.edu/~smschenn/qme.pdf>.
- SCHWARTZ, L. (1966): *Théorie des distributions*. Paris, Hermann.
- SINGLETON, K. J. (2001): “Estimation of affine asset pricing models using the empirical characteristic function,” *Journal of Econometrics*, 102, 111–141.
- TAUPIN, M.-L. (1998): “Estimation in the Nonlinear Errors-in-Variables Model,” *C. R. Acad. Sci. Paris*, 326, Serie I, 885–890.
- WANG, L. (2002): “A simple adjustment for measurement errors in some limited dependent variable models,” *Statistics & Probability Letters*, 58, 427–433.
- (2004): “Estimation of nonlinear models with Berkson measurement errors.,” *Annals of Statistics*, forthcoming.
- WANG, L., AND C. HSIAO (1995): “Simulation-Based Semiparametric Estimation of Nonlinear Errors-in-Variables Models,” Working Paper, University of Southern California.

Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models, Supplementary Material: Review, Proofs, Extensions and Example of Application

Susanne M. Schennach*
Department of Economics
University of Chicago
1126 East 59th Street
Chicago IL 60637
smschenn@uchicago.edu

July 2006

Abstract

This Supplementary Material contains some of the more technical details omitted from the paper “Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models”. First, a brief review of the Theory of Generalized Functions is presented. Second, proofs regarding some basic properties of Fourier transforms as well as the asymptotics of the proposed GMM estimator are given. Third, the proposed estimator is compared with the one suggested in Hausman, Newey, Ichimura, Powell (1991). Fourth, an alternative derivation of the moment conditions necessitating weaker regularity conditions is provided. Fifth, the details of the Monte Carlo simulations are described. Sixth, an application of the proposed methodology to the estimation of the black-white income gap is presented. Finally, some computational aspects of the implementation of the estimator are described.

Keywords: errors-in-variables model, Fourier transform, generalized function, semiparametric model.

*This work was made possible in part through financial support from the National Science Foundation via grant SES-0452089. The author would like to thank Jeremy Fox, Ricardo Mayer, Derek Neal and Xiaohong Chen, as well as participants at seminars given at the Universities of Rochester, Chicago, Maryland, Michigan, UCSD and UC-Riverside, the 2004 summer meetings of the Econometric Society and the CIRANO/CIREQ “Operator Methods in Microeconometrics, Time Series and Finance” conference for their helpful comments.

1 Review of the Theory of Generalized Functions

The concept of “generalized functions”, also called “tempered distributions” ((Lighthill 1962), (Gel’fand and Shilov 1964), (Schwartz 1966)), is central to the present paper, because most results will rely on Fourier transforms, which often do not exist within the set of ordinary functions. As generalized functions are not widely used in the econometrics literature ((Phillips 1991) and (Zinde-Walsh and Phillips 2003) are notable exceptions), this section recalls the definitions and known results that are relevant to our problem. Our summary of the theory of generalized functions most closely follows the treatment described in (Lighthill 1962), which is, of course, equivalent to the other treatments found in the literature. We focus on the case of scalar-valued generalized functions of a scalar variable.

In order to define generalized functions,¹ we first need the following definition.

Definition 1 *Let \mathcal{T} be the set of all functions $s : \mathbb{R} \mapsto \mathbb{R}$ that (i) are everywhere differentiable any number of times and (ii) are such that² $\left| \frac{d^k s(t)}{dt^k} \right| = O(|t|^{-\ell})$ as $|t| \rightarrow \infty$ for all $k, \ell \in \mathbb{N}^+$. Functions in \mathcal{T} are called “test” functions.*

Intuitively, functions in \mathcal{T} are both extremely smooth and have extremely thin tails.

Definition 2 *A generalized function³ b is a sequence of functions b_k in \mathcal{T} such that $\lim_{k \rightarrow \infty} \int b_k(t) s(t) dt$ exists for all $s \in \mathcal{T}$.*

Note that the limit of the sequence $b_k(t)$ may not be part of \mathcal{T} , which enables the concept of generalized functions to be more general than a function. The value of the integral $\int b(t) s(t) dt$ for a given $s \in \mathcal{T}$ is then defined as $\lim_{k \rightarrow \infty} \int b_k(t) s(t) dt$.

¹We adopt the term “generalized function” instead of “distribution” to avoid any potential confusion with the concept of probability distribution function.

²By convention $d^k s(t)/dt^k = s(t)$ for $k = 0$.

³Generalized functions can also be defined as bounded linear functionals on \mathcal{T} , but this definition is less convenient for our purposes.

Perhaps the best known example of a generalized function is Dirac's delta function $\delta(t)$, defined, for instance, by the sequence

$$b_k(t) = \sqrt{\frac{k}{2\pi}} \exp\left(-\frac{kt^2}{2}\right). \quad (1)$$

Another important example of a generalized function is the j -th derivative of the delta function, denoted by $\delta^{(j)}(t)$ and defined by the sequence $d^j b_k(t)/dt^j$, where $b_k(t)$ is as in Equation (1). The generalized function $\delta^{(j)}(t)$ has the property that $\delta^{(0)}(t) \equiv \delta(t)$ and

$$\int \delta^{(j)}(t) s(t) dt = (-1)^j \left. \frac{d^j s(t)}{dt^j} \right|_{t=0} \quad \text{for } j \in \mathbb{N}.$$

Definition 3 *Two generalized functions $a(t)$ and $b(t)$ are said to be equal if their associated sequences $a_k(t)$ and $b_k(t)$, respectively, are such that $\lim_{k \rightarrow \infty} \int a_k(t) s(t) dt = \lim_{k \rightarrow \infty} \int b_k(t) s(t) dt$ for all $s \in \mathcal{T}$.*

Note that this definition does not require that $a_k(t) = b_k(t)$ for all k and hence, a given generalized function can be defined in terms of more than one sequence. The set of generalized functions is closed under addition, subtraction and differentiation. The product of a generalized function with an ordinary function is guaranteed to be a generalized function if all of the ordinary function's derivatives exist and diverge no faster than a power of t as $|t| \rightarrow \infty$. However, the product of two generalized functions may not be a generalized function.

Ordinary functions can be viewed as particular cases of generalized functions. For instance, if we let \mathcal{I} be the set of all ordinary functions $c(t)$ such that $\int (1+t^2)^{-\ell} |c(t)| dt$ is finite for some $\ell \in \mathbb{N}$, then all ordinary functions in \mathcal{I} are also generalized functions. A generalized function $b(t)$ is said to be equal to an ordinary function $c(t)$ in an interval I if for all $s \in \mathcal{T}$ that are supported on I , we have $\int b(t) s(t) dt = \int c(t) s(t) dt$. In the case of Dirac's delta function, $\delta(t)$ is equal to the 0 function over any interval that does not contain 0. However, $\delta(t)$ is not equal to any ordinary function over

any interval that includes 0. This concept is important because it will allow us to treat generalized functions as ordinary functions, as long as we stay away from their “singular” points. More generally, two generalized functions $b(t)$ and $c(t)$ are also said to be equal over an interval I if, for all $s \in \mathcal{T}$ that are supported on I , we have $\int b(t) s(t) dt = \int c(t) s(t) dt$.

Perhaps the most important result for our purpose is that the Fourier transform of a generalized function is a generalized function. As a particular case of this result, the Fourier transform of any function in \mathcal{I} is a generalized function. Hence, in general, the Fourier transform of an ordinary function will not necessarily be an ordinary function, but rather a generalized function.

A generalized function $b(t)$ can always be decomposed as

$$b(t) = b_o(t) + b_s(t) \tag{2}$$

where $b_o(t)$ is an ordinary function while $b_s(t)$ is purely singular, consisting solely of a linear combination of delta function derivatives of a finite order.⁴ This result directly follows from the fact that every generalized function can be written as the derivative of order $k \in \mathbb{N}$ of some continuous function $c(t)$ (Theorem III in (Temple 1963) establishes this for a class of generalized functions including the ones considered here as a particular case). At every point t where $c(t)$ is k times differentiable in the usual sense, the generalized function can be written as an ordinary function, while at every point where $c(t)$ is not k times differentiable, a delta function derivative is created in the differentiation process. The fact that the two pieces are additively separable follows from the linear nature of the space of generalized functions. The decomposition (2) is unique because there exists no ordinary function $b_o(t)$ such that, for $k \in \mathbb{N}$, $\int b_o(t) s(t) dt = \int \delta^{(k)}(t) s(t) dt$ for all test function $s(t)$.

Moreover, the product of a generalized function $b(t)$ with an ordinary function

⁴The linear combination can consist of an infinite number of terms (with delta function derivatives at different locations).

$a_o(t)$ can be decomposed as

$$b(t) a_o(t) = b_o(t) a_o(t) + b_s(t) a_o(t) \quad (3)$$

where $b_o(t) a_o(t)$ is an ordinary function and where $b_s(t) a_o(t)$ is purely singular, as implied by Lemma 2 (see Appendix A). Of course, $b(t) a_o(t)$ will only be well-defined if $a_o(t)$ admits a sufficient number of continuous derivatives at the points where the delta functions derivatives contained in $b(t)$ are located.

While this review focuses on so-called *tempered* distributions, there exist more general classes of generalized functions. For instance, as described in (Gel'fand and Shilov 1964), the set \mathcal{T} can be limited to compactly supported infinitely differentiable functions, which expands the set of generalized functions for which the limit $\lim_{k \rightarrow \infty} \int a_k(t) s(t) dt$ exists for any $s \in \mathcal{T}$. However, in this work, we focus on functions $a(t)$ whose Fourier transforms $\alpha(\tau)$ are tempered distributions, therefore limiting ourselves to functions $a(t)$ that diverge no faster than some power of t as $|t| \rightarrow \infty$.

2 Simple Results about Fourier transforms and generalized functions

Definition 4 For some function $\psi(\zeta)$, let $d^{-1}\psi(\zeta)/d\zeta^{-1} \equiv \int_a^\zeta \psi(\xi) d\xi$ for some arbitrary constant a . For $k \geq 1$, define, by recursion,

$$\frac{d^{-k-1}}{d\zeta^{-k-1}}\psi(\zeta) \equiv \frac{d^{-1}}{d\zeta^{-1}} \frac{d^{-k}}{d\zeta^{-k}}\psi(\zeta).$$

Complete proof of Lemma 2. Let ψ be some test function in \mathcal{T} as given in Definition 1. By k repeated integration by parts, we have,

$$\int \left(\delta^{(k)}(\zeta) \phi(\zeta) \right) \psi(\zeta) d\zeta = (-1)^k \int \left(\frac{d^{-k}}{d\zeta^{-k}} \delta^{(k)}(\zeta) \right) \frac{d^k}{d\zeta^k} (\phi(\zeta) \psi(\zeta)) d\zeta,$$

after noting that the boundary terms vanish due to the thin tails of $\psi(\zeta)$ and all of its derivatives. Next,

$$\begin{aligned}
\int \left(\delta^{(k)}(\zeta) \phi(\zeta) \right) \psi(\zeta) d\zeta &= (-1)^k \int \delta(\zeta) \left(\frac{d^k}{d\zeta^k} (\phi(\zeta) \psi(\zeta)) \right) d\zeta \\
&= (-1)^k \int \delta(\zeta) \sum_{j=0}^k \binom{k}{j} \frac{d^{k-j} \phi(\zeta)}{d\zeta^{k-j}} \frac{d^j \psi(\zeta)}{d\zeta^j} d\zeta \\
&= (-1)^k \sum_{j=0}^k \binom{k}{j} \frac{d^{k-j} \phi(0)}{d\zeta^{k-j}} \frac{d^j \psi(0)}{d\zeta^j} \\
&= (-1)^k \sum_{j=0}^k \binom{k}{j} \frac{d^{k-j} \phi(0)}{d\zeta^{k-j}} \int \delta^{(j)}(\zeta) \psi(\zeta) d\zeta \\
&= \int \left((-1)^k \sum_{j=0}^k \binom{k}{j} \frac{d^{k-j} \phi(0)}{d\zeta^{k-j}} \delta^{(j)}(\zeta) \right) \psi(\zeta) d\zeta.
\end{aligned}$$

■

Complete proof of Lemma 4. Substituting Equations (25) through (58) into Equations (8) and (9), we obtain

$$\begin{aligned}
\varepsilon_{y,o}(\zeta) + 2\pi \sum_{k=0}^{\bar{k}} \varepsilon_{y,k} (-\mathbf{i})^k \delta^{(k)}(\zeta) &= \left(\gamma_o(\zeta, \theta) + 2\pi \sum_{k=0}^{\bar{k}} \gamma_k(\theta) (-\mathbf{i})^k \delta^{(k)}(\zeta) \right) \phi(\zeta) \\
\mathbf{i} \varepsilon_{xy,o}(\zeta) + 2\pi \mathbf{i} \sum_{k=-1}^{\bar{k}} \varepsilon_{xy,k} (-\mathbf{i})^{k+1} \delta^{(k+1)}(\zeta) &= \left(\dot{\gamma}_o(\zeta, \theta) + 2\pi \sum_{k=0}^{\bar{k}} \gamma_k(\theta) (-\mathbf{i})^k \delta^{(k+1)}(\zeta) \right) \phi(\zeta).
\end{aligned}$$

Equating the ordinary functions part of each expression yields

$$\begin{aligned}
\varepsilon_{y,o}(\zeta) &= \gamma_o(\zeta, \theta) \phi(\zeta) \\
\mathbf{i} \varepsilon_{xy,o}(\zeta) &= \dot{\gamma}_o(\zeta, \theta) \phi(\zeta),
\end{aligned}$$

while equating the singular parts yields

$$\begin{aligned}
\sum_{k=0}^{\bar{k}} \varepsilon_{y,k} (-\mathbf{i})^k \delta^{(k)}(\zeta) &= \sum_{k=0}^{\bar{k}} \gamma_k(\theta) (-\mathbf{i})^k \delta^{(k)}(\zeta) \phi(\zeta) \\
\sum_{k=-1}^{\bar{k}} \mathbf{i} \varepsilon_{xy,k} (-\mathbf{i})^{k+1} \delta^{(k+1)}(\zeta) &= \sum_{k=0}^{\bar{k}} \gamma_k(\theta) (-\mathbf{i})^k \delta^{(k+1)}(\zeta) \phi(\zeta).
\end{aligned}$$

By Lemma 2, we have

$$\begin{aligned} \sum_{k=0}^{\bar{k}} \varepsilon_{y,k} (-\mathbf{i})^k \delta^{(k)}(\zeta) &= \sum_{k=0}^{\bar{k}} \gamma_k(\theta) (-\mathbf{i})^k \sum_{j=0}^k \binom{k}{j} \phi^{(k-j)}(0) \delta^{(j)}(\zeta) \\ \sum_{k=-1}^{\bar{k}} \mathbf{i} \varepsilon_{xy,k} (-\mathbf{i})^{k+1} \delta^{(k+1)}(\zeta) &= \sum_{k=0}^{\bar{k}} \gamma_k(\theta) (-\mathbf{i})^k \sum_{j=0}^{k+1} \binom{k+1}{j} \phi^{(k+1-j)}(0) \delta^{(j)}(\zeta) \end{aligned}$$

Simple manipulations then give

$$\begin{aligned} \sum_{k=0}^{\bar{k}} \varepsilon_{y,k} (-\mathbf{i})^k \delta^{(k)}(\zeta) &= \sum_{k=0}^{\bar{k}} \gamma_k(\theta) (-\mathbf{i})^k \sum_{j=0}^{\bar{k}} \binom{k}{j} \mathbf{1}(j \leq k) \phi^{(k-j)}(0) \delta^{(j)}(\zeta) \\ \sum_{k=-1}^{\bar{k}} \mathbf{i} \varepsilon_{xy,k} (-\mathbf{i})^{k+1} \delta^{(k+1)}(\zeta) &= \sum_{k=0}^{\bar{k}} \gamma_k(\theta) (-\mathbf{i})^k \sum_{j=0}^{\bar{k}+1} \binom{k+1}{j} \mathbf{1}(j \leq k+1) \phi^{(k+1-j)}(0) \delta^{(j)}(\zeta) \end{aligned}$$

$$\begin{aligned} \sum_{j=0}^{\bar{k}} \varepsilon_{y,j} (-\mathbf{i})^j \delta^{(j)}(\zeta) &= \sum_{j=0}^{\bar{k}} \sum_{k=0}^{\bar{k}} \binom{k}{j} \gamma_k(\theta) (-\mathbf{i})^k \mathbf{1}(j \leq k) \phi^{(k-j)}(0) \delta^{(j)}(\zeta) \\ \sum_{j=-1}^{\bar{k}} \mathbf{i} \varepsilon_{xy,j} (-\mathbf{i})^{j+1} \delta^{(j+1)}(\zeta) &= \sum_{j=0}^{\bar{k}+1} \sum_{k=0}^{\bar{k}} \binom{k+1}{j} \gamma_k(\theta) (-\mathbf{i})^k \mathbf{1}(j \leq k+1) \phi^{(k+1-j)}(0) \delta^{(j)}(\zeta) \\ &= \sum_{j=-1}^{\bar{k}} \sum_{k=0}^{\bar{k}} \binom{k+1}{j+1} \gamma_k(\theta) (-\mathbf{i})^k \mathbf{1}(j \leq k) \phi^{(k-j)}(0) \delta^{(j+1)}(\zeta). \end{aligned}$$

Equating the coefficients of the delta function derivatives of the same order gives

$$\begin{aligned} \varepsilon_{y,j} (-\mathbf{i})^j &= \sum_{k=0}^{\bar{k}} \binom{k}{j} \gamma_k(\theta) (-\mathbf{i})^k \mathbf{1}(j \leq k) \phi^{(k-j)}(0) \\ \mathbf{i} \varepsilon_{xy,j} (-\mathbf{i})^{j+1} &= \sum_{k=0}^{\bar{k}} \binom{k+1}{j+1} \gamma_k(\theta) (-\mathbf{i})^k \mathbf{1}(j \leq k) \phi^{(k-j)}(0) \end{aligned}$$

$$\begin{aligned}
\varepsilon_{y,j} (-\mathbf{i})^j &= \sum_{l=-j}^{\bar{k}-j} \binom{j+l}{j} \gamma_{j+l}(\theta) (-\mathbf{i})^{j+l} \mathbf{1}(j \leq j+l) \phi^{(j+l-j)}(0) \\
&= \sum_{l=-j}^{\bar{k}-j} \binom{j+l}{j} \gamma_{j+l}(\theta) (-\mathbf{i})^{j+l} \mathbf{1}(0 \leq l) \phi^{(l)}(0) \\
&= \sum_{l=0}^{\bar{k}-j} \binom{j+l}{j} \gamma_{j+l}(\theta) (-\mathbf{i})^{j+l} \phi^{(l)}(0) \\
&= \sum_{l=0}^{\bar{k}} \binom{j+l}{j} \gamma_{j+l}(\theta) (-\mathbf{i})^{j+l} \mathbf{1}(l \leq \bar{k}-j) \phi^{(l)}(0) \\
&= \sum_{k=0}^{\bar{k}} \binom{k+j}{j} \gamma_{k+j}(\theta) (-\mathbf{i})^{k+j} \mathbf{1}(k \leq \bar{k}-j) \phi^{(k)}(0)
\end{aligned}$$

$$\begin{aligned}
\mathbf{i}\varepsilon_{xy,j} (-\mathbf{i})^{j+1} &= \sum_{l=-j}^{\bar{k}-j} \binom{j+l+1}{j+1} \gamma_{j+l}(\theta) (-\mathbf{i})^{j+l} \mathbf{1}(j \leq j+l) \phi^{(j+l-j)}(0) \\
&= \sum_{l=0}^{\bar{k}-j} \binom{j+l+1}{j+1} \gamma_{j+l}(\theta) (-\mathbf{i})^{j+l} \phi^{(l)}(0) \\
&= \sum_{l=0}^{\bar{k}+1} \binom{j+l+1}{j+1} \gamma_{j+l}(\theta) (-\mathbf{i})^{j+l} \mathbf{1}(l \leq \bar{k}-j) \phi^{(l)}(0) \\
&= \sum_{k=0}^{\bar{k}} \binom{k+j+1}{j+1} \gamma_{k+j}(\theta) (-\mathbf{i})^{k+j} \mathbf{1}(k \leq \bar{k}-j) \phi^{(k)}(0) \text{ for } j \geq 0.
\end{aligned}$$

$$\begin{aligned}
\varepsilon_{y,j} &= \sum_{k=0}^{\bar{k}} \binom{k+j}{j} \gamma_{k+j}(\theta) \mathbf{1}(k \leq \bar{k}-j) (-\mathbf{i})^k \phi^{(k)}(0) \\
\varepsilon_{xy,j} &= \sum_{k=0}^{\bar{k}} \binom{k+j+1}{j+1} \gamma_{k+j}(\theta) \mathbf{1}(k \leq \bar{k}-j) (-\mathbf{i})^k \phi^{(k)}(0).
\end{aligned}$$

■

3 Asymptotics of the GMM estimator

3.1 Definitions

Let (X_j, Y_j, W_j) for $j = 1, \dots, n$ be a given sample. First, the variable Z_j needs to be constructed from the instruments W_j (see Equation (4) in the main text). To this effect, the parameter vector α in Model (2) is estimated using standard (nonlinear) least-squares on the specification

$$X_j = m(W_j, \alpha) + (\Delta X_j^* + \Delta X_j) \quad (4)$$

where $E[(\Delta X_j^* + \Delta X_j) | W_j] = 0$ by the assumptions of Model (2). The resulting $\hat{\alpha}$ is used to define the variable \hat{Z}_j as

$$\hat{Z}_j = m(W_j, \hat{\alpha}). \quad (5)$$

The variable \hat{Z}_j estimates the true $Z_j = m(W_j, \alpha^*)$, where α^* denotes the true value of α . Let $p(\cdot | \alpha)$ denote the density of the quantity $m(W_j, \alpha)$ for a given α and let $p(z) = p(z | \alpha^*)$. Next, a nonparametric kernel density estimate of $p(\cdot | \hat{\alpha})$ at point \hat{Z}_j can be obtained from

$$\hat{p}(\hat{Z}_j | \hat{\alpha}) = (nh)^{-1} \sum_{i=1, i \neq j}^n K\left(\frac{\hat{Z}_i - \hat{Z}_j}{h}\right)$$

for some kernel $K(\cdot)$ and some bandwidth sequence $h \rightarrow 0$ as $n \rightarrow \infty$.

Finally, $\hat{\theta}$ is defined as the solution to $\hat{Q}(\theta, \hat{\alpha}) = 0$, where

$$\hat{Q}(\theta, \alpha) \equiv n^{-1} \sum_{j=1}^n \left(\frac{\Upsilon(X_j, Y_j, W_j, \theta, \alpha)}{\hat{p}(m(W_j, \alpha) | \alpha)} - \mathbf{e} \right) 1(\hat{p}(m(W_j, \alpha) | \alpha) \geq \tau) \quad (6)$$

$$\Upsilon(x, y, w, \theta, \alpha) = \begin{bmatrix} y r_y(m(w, \alpha), \theta) + xy r_{xy}(m(w, \alpha), \theta) \\ y r_{1y}(m(w, \alpha), \theta) \end{bmatrix} \quad (7)$$

$$\mathbf{e} = \underbrace{(0, \dots, 0)}_{N_\theta - N_s}, \underbrace{(1, \dots, 1)}_{N_s} \quad (8)$$

where $1(\cdot)$ is the indicator function, equal to 1 when the event \cdot occurs and τ is some trimming threshold such that $\tau \rightarrow 0$ as $n \rightarrow \infty$ designed to keep divisions by

zero under control.⁵ The scalar N_s is the dimension of the range of $r_{1y}(z, \theta)$ and can therefore be 0, 1, or 2. The true value of θ , denoted θ^* , is the solution to $Q(\theta, \alpha^*) = 0$, where

$$Q(\theta, \alpha) = E[Q(X, Y, W, \theta, p(\cdot|\alpha))] \quad (9)$$

and where $Q(x, y, w, \theta, p)$ is defined in Equation (16).

3.2 Proofs

While the following Lemma may seem familiar, we were not able to find this result at the required level of generality in the existing literature (Theorem 1 and 3 in (Andrews 1995) and Theorem 2.8 in (Pagan and Ullah 1999) come very close, however).

Lemma 1 *Under Assumptions 8, 14 and 15*

$$\sup_{\alpha \in \mathcal{A}} \sup_{z \in \mathbb{R}} |\tilde{p}(z|\alpha) - p(z|\alpha)| = O_p(n^{-1/2}h^{-1}) + O(h^{N_K})$$

where $\tilde{p}(z|\alpha) = (nh)^{-1} \sum_{j=1}^n K((Z_j - z)/h)$ and $p(z|\alpha)$ is the density of $Z = m(W, \alpha)$ for a given function $m(W, \alpha)$ of some random vector W . The same result holds with $\tilde{p}(z|\alpha)$ replaced by $\hat{p}(z|\alpha) = (nh)^{-1} \sum_{j=1}^n K((Z_j - z)/h) 1(Z_j \neq z)$.

Proof. This proof is based in part on the proof of Theorem 2.8 in (Pagan and Ullah 1999). Note that $\sup_{\alpha \in \mathcal{A}} \sup_{z \in \mathbb{R}} |\tilde{p}(z|\alpha) - p(z|\alpha)| \leq R + B$, where

$$\begin{aligned} R &= \sup_{\alpha \in \mathcal{A}} \sup_{z \in \mathbb{R}} |\tilde{p}(z|\alpha) - E[\tilde{p}(z|\alpha)]|, \\ B &= \sup_{\alpha \in \mathcal{A}} \sup_{z \in \mathbb{R}} |E[\tilde{p}(z|\alpha)] - p(z|\alpha)|. \end{aligned}$$

⁵The trimming is not introduced to ensure that expectations such as $E[Yr_y(Z, \theta)/p(Z)]$ or $E[(Yr_y(Z, \theta)/p(Z))^2]$ exist but rather to show that remainder terms are asymptotically negligible. If $E[(Yr_y(Z, \theta)/p(Z))^2]$, for instance, did not exist, no trimming scheme would restore the root n consistent estimation of the moment $E[Yr_y(Z, \theta)/p(Z)]$.

By the convolution Theorem,

$$\begin{aligned}
R &= \sup_{\alpha \in \mathcal{A}} \sup_{z \in \mathbb{R}} \left| \int \kappa(h\zeta) n^{-1} \sum_{j=1}^n (e^{i\zeta Z_j} - E[e^{i\zeta Z_j}]) e^{-i\zeta z} d\zeta \right| \\
&\leq \sup_{\alpha \in \mathcal{A}} \sup_{z \in \mathbb{R}} \int |\kappa(h\zeta)| \left| n^{-1} \sum_{j=1}^n (e^{i\zeta Z_j} - E[e^{i\zeta Z_j}]) \right| d\zeta \\
&= \sup_{\alpha \in \mathcal{A}} \int |\kappa(h\zeta)| \left| n^{-1} \sum_{j=1}^n (e^{i\zeta Z_j} - E[e^{i\zeta Z_j}]) \right| d\zeta
\end{aligned}$$

where $\kappa(\zeta)$ denotes the Fourier transform of $K(z)$. We then have

$$\begin{aligned}
E[R] &\leq \sup_{\alpha \in \mathcal{A}} \int |\kappa(h\zeta)| E \left[\left| n^{-1} \sum_{j=1}^n (e^{i\zeta Z_j} - E[e^{i\zeta Z_j}]) \right| \right] d\zeta \\
&\leq \sup_{\alpha \in \mathcal{A}} \int |\kappa(h\zeta)| \left(E \left[\left| n^{-1} \sum_{j=1}^n (e^{i\zeta Z_j} - E[e^{i\zeta Z_j}]) \right|^2 \right] \right)^{1/2} d\zeta \\
&= \sup_{\alpha \in \mathcal{A}} \int |\kappa(h\zeta)| (n^{-1} E[(e^{i\zeta Z_j} - E[e^{i\zeta Z_j}]) (e^{-i\zeta Z_j} - E[e^{-i\zeta Z_j}])])^{1/2} d\zeta \\
&= \sup_{\alpha \in \mathcal{A}} n^{-1/2} \int |\kappa(h\zeta)| (E[(e^{i\zeta Z_j} - E[e^{i\zeta Z_j}]) (e^{-i\zeta Z_j} - E[e^{-i\zeta Z_j}])])^{1/2} d\zeta \\
&\leq n^{-1/2} 2^{1/2} \int |\kappa(h\zeta)| d\zeta \\
&= n^{-1/2} h^{-1} 2^{1/2} \int |\kappa(\zeta)| d\zeta \\
&= O(n^{-1/2} h^{-1})
\end{aligned}$$

and $R = O_p(n^{-1/2} h^{-1})$ by Markov's inequality. Next,

$$B = \sup_{\alpha \in \mathcal{A}} \sup_{z \in \mathbb{R}} \left| \int h^{-1} K(h^{-1}v) (p(z+v|\alpha) - p(z|\alpha)) dv \right|.$$

By a Taylor expansion,

$$\begin{aligned}
B &= \sup_{\alpha \in \mathcal{A}} \sup_{z \in \mathbb{R}} \left| \int h^{-1} K(h^{-1}v) \left(\sum_{j=1}^{N_k-1} p^{(j)}(z|\alpha) \frac{v^j}{j!} + p^{(N_k)}(\tilde{z}|\alpha) \frac{v^{N_k}}{N_k!} \right) dv \right| \text{ for } \tilde{z} \in [z, z+v] \\
&= \sup_{\alpha \in \mathcal{A}} \sup_{z \in \mathbb{R}} \left| \int h^{-1} K(h^{-1}v) p^{(N_k)}(\tilde{z}|\alpha) \frac{v^{N_k}}{N_k!} dv \right|
\end{aligned}$$

by Assumption 14(iii). Then, by a change of variable,

$$\begin{aligned} B &= \sup_{\alpha \in \mathcal{A}} \sup_{z \in \mathbb{R}} \left| \int K(u) p^{(N_k)}(\tilde{z}|\alpha) \frac{u^{N_k} h^{N_k}}{N_k!} du \right| \\ &\leq h^{N_k} \left(\sup_{\alpha \in \mathcal{A}} \sup_{z \in \mathbb{R}} |p^{(N_k)}(\tilde{z}|\alpha)| \right) \frac{1}{N_k!} \left| \int |K(u)| |u|^{N_k} du \right| = O(h^{N_k}) \end{aligned}$$

by Assumptions 14(iv) and 15.

The second assertion is shown by noting that the difference between $\tilde{p}(z)$ and $\hat{p}(z)$ is at most $K(0)n^{-1}h^{-1}$ which is of an order less than $n^{-1/2}h^{-1}$ and can therefore be absorbed in the $O_p(n^{-1/2}h^{-1})$ remainder. \blacksquare

Proof of Theorem 3. Let $\Upsilon(x, y, w, \theta, \alpha)$, $\hat{Q}(\theta, \alpha)$ and $Q(\theta, \alpha)$ be as defined in Section 3.1. We first show consistency of $\hat{\theta}$. This involves establishing the uniform convergence of $\hat{Q}(\theta, \hat{\alpha})$ to $Q(\theta, \alpha^*)$ for $\theta \in \Theta$. We first note that $\hat{\alpha} \xrightarrow{p} \alpha^*$, by Lemma 2.4 and Theorem 2.1 in (Newey and McFadden 1994), under Assumptions 8 and 9. Hence $\hat{\alpha} \in \mathcal{A}$ with probability approaching 1 (hereafter w.p.a.1). We can then write, w.p.a.1,

$$\begin{aligned} \sup_{\theta \in \Theta} \left\| \hat{Q}(\theta, \hat{\alpha}) - Q(\theta, \alpha^*) \right\| &\leq \sup_{\theta \in \Theta} \left\| \hat{Q}(\theta, \hat{\alpha}) - Q(\theta, \hat{\alpha}) \right\| + \sup_{\theta \in \Theta} \left\| Q(\theta, \hat{\alpha}) - Q(\theta, \alpha^*) \right\| \\ &\leq \sup_{\alpha \in \mathcal{A}} \sup_{\theta \in \Theta} \left\| \hat{Q}(\theta, \alpha) - Q(\theta, \alpha) \right\| + \sup_{\theta \in \Theta} \left\| Q(\theta, \hat{\alpha}) - Q(\theta, \alpha^*) \right\| \end{aligned}$$

where $\sup_{\theta \in \Theta} \left\| Q(\theta, \hat{\alpha}) - Q(\theta, \alpha^*) \right\| \xrightarrow{p} 0$ by $\hat{\alpha} \xrightarrow{p} \alpha^*$ and Assumption 18. Next,

$$\sup_{\alpha \in \mathcal{A}} \sup_{\theta \in \Theta} \left\| \hat{Q}(\theta, \alpha) - Q(\theta, \alpha) \right\| \leq R_A + R_I + R_D$$

where

$$\begin{aligned} R_A &= \sup_{\alpha \in \mathcal{A}} \sup_{\theta \in \Theta} \left\| n^{-1} \sum_{j=1}^n \frac{\Upsilon(X_j, Y_j, W_j, \theta, \alpha)}{p(m(W_j, \alpha) | \alpha)} - E \left[\frac{\Upsilon(X, Y, W, \theta, \alpha)}{p(m(W, \alpha) | \alpha)} \right] \right\| \\ R_I &= \sup_{\alpha \in \mathcal{A}} \sup_{\theta \in \Theta} \left\| n^{-1} \sum_{j=1}^n \frac{\Upsilon(X_j, Y_j, W_j, \theta, \alpha)}{p(m(W_j, \alpha) | \alpha)} (1(\hat{p}(m(W_j, \alpha) | \alpha) \geq \tau) - 1) \right\| \\ R_D &= \sup_{\alpha \in \mathcal{A}} \sup_{\theta \in \Theta} \left\| n^{-1} \sum_{j=1}^n \Upsilon(X_j, Y_j, W_j, \theta, \alpha) \frac{p(m(W_j, \alpha) | \alpha) - \hat{p}(m(W_j, \alpha) | \alpha)}{\hat{p}(m(W_j, \alpha) | \alpha) p(m(W_j, \alpha) | \alpha)} \times \right. \\ &\quad \left. \times 1(\hat{p}(m(W_j, \alpha) | \alpha) \geq \tau) \right\| \end{aligned}$$

We then have $\sup_{\theta \in \Theta} \|R_A\| \xrightarrow{p} 0$ by Assumptions 8, 10 and 11 and Lemma 2.4 in (Newey and McFadden 1994). Next, by Lemma 1, we have

$$\begin{aligned}
R_I &\leq \sup_{\alpha \in \mathcal{A}} \sup_{\theta \in \Theta} n^{-1} \sum_{j=1}^n \frac{\|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\|}{p(m(W_j, \alpha) | \alpha)} |1(\hat{p}(m(W_j, \alpha) | \alpha) < \tau)| \\
&\leq \sup_{\alpha \in \mathcal{A}} \sup_{\theta \in \Theta} n^{-1} \sum_{j=1}^n \frac{\|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\|}{p(m(W_j, \alpha) | \alpha)} |1(p(m(W_j, \alpha) | \alpha) - Cn^{\epsilon-1/2}h^{-1} < \tau)| \\
&\quad \text{w.p.a. 1 for } \epsilon \in]0, 1/4[\\
&= \sup_{\alpha \in \mathcal{A}} \sup_{\theta \in \Theta} n^{-1} \sum_{j=1}^n \frac{\|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\|}{p(m(W_j, \alpha) | \alpha)} |1(p(m(W_j, \alpha) | \alpha) < \tau(1 + Cn^{\epsilon-1/2}h^{-1}/\tau))| \\
&\leq \sup_{\alpha \in \mathcal{A}} \sup_{\theta \in \Theta} n^{-1} \sum_{j=1}^n \frac{\|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\|}{p(m(W_j, \alpha) | \alpha)} |1(p(m(W_j, \alpha) | \alpha) < 2\tau)| \text{ by Assumption 16}
\end{aligned}$$

and $E[R_I] \leq E[\sup_{\alpha \in \mathcal{A}} \sup_{\theta \in \Theta} \|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\| |1(p(Z_j) < 2\tau)| / p(m(W_j, \alpha) | \alpha)] = o(n^{-1/2})$ by Assumption 17, thus implying that $R_I = o_p(n^{-1/2})$, by Markov's inequality. Next,

$$\begin{aligned}
R_D &\leq \sup_{\alpha \in \mathcal{A}} \sup_{\theta \in \Theta} n^{-1} \sum_{j=1}^n \|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\| \left(\frac{|p(m(W_j, \alpha) | \alpha) - \hat{p}(m(W_j, \alpha) | \alpha)|}{\hat{p}(Z_j) p(m(W_j, \alpha) | \alpha)} \right) \hat{I}_j \\
&\leq \sup_{\alpha \in \mathcal{A}} \sup_{\theta \in \Theta} \tau^{-1} n^{-1} \sum_{j=1}^n \|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\| \left(\frac{|p(m(W_j, \alpha) | \alpha) - \hat{p}(m(W_j, \alpha) | \alpha)|}{p(m(W_j, \alpha) | \alpha)} \right) \hat{I}_j \\
&\leq \sup_{z \in \mathbb{R}} |p(z) - \hat{p}(z)| \tau^{-1} n^{-1} \sum_{j=1}^n \left(\frac{\|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\|}{p(m(W_j, \alpha) | \alpha)} \right) \\
&= (O_p(n^{-1/2}h^{-1}) + O(h^{N_K})) \tau^{-1} O_p(1)
\end{aligned}$$

by Lemma 1, and Lemma 2.4 in (Newey and McFadden 1994) under Assumptions 8, 10 and 11. By Assumption 16, $n^{-1/2}h^{-1}\tau^{-1} \rightarrow 0$ and $h^{N_K} \rightarrow 0$ and it follows that $R_D \xrightarrow{p} 0$.

Having shown that $\sup_{\theta \in \Theta} \|\hat{Q}(\theta, \hat{\alpha}) - Q(\theta, \alpha^*)\| \xrightarrow{p} 0$, we now establish that this implies⁶ that $\hat{\theta}$ converges to θ^* . Since $\hat{Q}(\hat{\theta}, \hat{\alpha}) = 0$ and $\sup_{\theta \in \Theta} \|\hat{Q}(\theta, \hat{\alpha}) - Q(\theta, \alpha^*)\| \xrightarrow{p} 0$

⁶This would be obvious if $\hat{\theta}$ were defined as the maximizer of a random function. Here $\hat{\theta}$ is the solution to a set of equations and the usual consistency result (e.g. Theorem 2.1 in (Newey and McFadden 1994)) does not directly apply.

0 it follows that $\text{plim}_{n \rightarrow \infty} Q(\hat{\theta}, \alpha^*) = 0$. Since $\hat{Q}(\theta, \hat{\alpha})$ is continuous in θ (because $\Upsilon(x_j, y_j, w_j, \theta, \alpha)$ is), and its convergence to $Q(\theta, \alpha^*)$ is uniform in θ , $Q(\theta, \alpha^*)$ must be continuous in θ . Combining these two results yields $\text{plim}_{n \rightarrow \infty} Q(\hat{\theta}, \alpha^*) = Q(\text{plim}_{n \rightarrow \infty} \hat{\theta}, \alpha^*) = 0$. Since $\theta = \theta^*$ is the only solution to $Q(\theta, \alpha^*) = 0$ by Assumption 6, we conclude that $\text{plim}_{n \rightarrow \infty} \hat{\theta} = \theta^*$.

Having shown consistency, we turn to asymptotic normality and root n consistency. By a standard mean value expansion of the first-order conditions $\hat{Q}(\hat{\theta}, \hat{\alpha}) = 0$ around θ^* and the usual manipulations,

$$n^{1/2}(\hat{\theta} - \theta^*) = - \left(\frac{\partial \hat{Q}(\bar{\theta}, \hat{\alpha})}{\partial \theta'} \right)^{-1} n^{1/2} \hat{Q}(\theta^*, \hat{\alpha}), \quad (10)$$

for some mean value $\bar{\theta}$. Following the same steps as used above to show uniform convergence in probability of $\hat{Q}(\theta, \hat{\alpha})$, we can show that $\sup_{\theta \in \mathcal{N}} \left\| \partial \hat{Q}(\theta, \hat{\alpha}) / \partial \theta' - \partial Q(\theta, \alpha^*) / \partial \theta' \right\| \xrightarrow{p} 0$ and $\partial Q(\theta, \alpha^*) / \partial \theta'$ is continuous in θ , by simply replacing Assumption 11 by Assumption 12. Since $\hat{\theta} \xrightarrow{p} \theta^*$ it follows that $\bar{\theta} \xrightarrow{p} \theta^*$ and that $\partial Q(\bar{\theta}, \alpha^*) / \partial \theta' \xrightarrow{p} \partial Q(\theta^*, \alpha^*) / \partial \theta'$, thus implying that

$$\frac{\partial \hat{Q}(\bar{\theta}, \hat{\alpha})}{\partial \theta'} \xrightarrow{p} \frac{\partial Q(\theta^*, \alpha^*)}{\partial \theta'}. \quad (11)$$

Next, we let $\Upsilon_j = \Upsilon(X_j, Y_j, W_j, \theta^*, \alpha^*)$, $Z_j = m(W_j, \alpha^*)$, $\hat{p}(Z_j) = \hat{p}(m(W_j, \alpha^*) | \alpha^*)$, $p(Z_j) = p(m(W_j, \alpha^*) | \alpha^*)$, $\hat{I}_j = 1(\hat{p}(Z_j) \geq \tau)$, $I_j = 1(p(Z_j) \geq \tau)$ and decompose the term $n^{1/2} \hat{Q}(\theta^*, \hat{\alpha})$ in Equation 10 as

$$n^{1/2} \hat{Q}(\theta^*, \hat{\alpha}) = N + N_\alpha + R_{T1} + R_{T2} + R_{T3} + R_L + R_U + R_B + R_{\text{sec}}$$

where the asymptotically normal terms are given by

$$\begin{aligned} N &= n^{-1/2} \sum_{j=1}^n \frac{\Upsilon_j - E[\Upsilon_j | Z_j]}{p(Z_j)} \\ N_\alpha &= n^{1/2} (Q(\theta^*, \hat{\alpha}) - Q(\theta^*, \alpha^*)) \end{aligned}$$

while the remainder terms associated with trimming are

$$\begin{aligned} R_{T1} &= n^{-1/2} \sum_{j=1}^n \frac{\Upsilon_j}{\hat{p}(Z_j)} (\hat{I}_j - I_j) \\ R_{T2} &= n^{1/2} E \left[\frac{\Upsilon_j}{p(Z_j)} (1 - I_j) \right] \\ R_{T3} &= n^{-1/2} \sum_{j=1}^n \frac{(\Upsilon_j - E[\Upsilon_j | Z_j])}{p(Z_j)} (I_j - 1) \end{aligned}$$

the remainder from the linearization is given by

$$R_L = n^{-1/2} \sum_{j=1}^n \frac{\Upsilon_j}{\hat{p}(Z_j) p^2(Z_j)} (\hat{p}(Z_j) - p(Z_j))^2 I_j$$

the “ U -statistic” term is

$$R_U = -n^{-1/2} \sum_{j=1}^n \left(\frac{\Upsilon_j}{p^2(Z_j)} (\hat{p}(Z_j) - E[\hat{p}(Z_j) | Z_j]) I_j - \left(\frac{E[\Upsilon_j | Z_j]}{p(Z_j)} I_j - E \left[\frac{\Upsilon_j}{p(Z_j)} I_j \right] \right) \right),$$

the “bias” term is

$$R_B = n^{-1/2} \sum_{j=1}^n \frac{\Upsilon_j}{p^2(Z_j)} (p(Z_j) - E[\hat{p}(Z_j) | Z_j]) I_j$$

and the “stochastic equicontinuity” remainder term is

$$R_{\text{sec}} = n^{1/2} \left(\left(\hat{Q}(\theta^*, \hat{\alpha}) - Q(\theta^*, \hat{\alpha}) \right) - \left(\hat{Q}(\theta^*, \alpha^*) - Q(\theta^*, \alpha^*) \right) \right).$$

We consider each remainder in turn.

$$\begin{aligned}
|R_{T1}| &\leq n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{\hat{p}(Z_j)} |1(\hat{p}(Z_j) \geq \tau) - 1(p(Z_j) \geq \tau)| \\
&\leq n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{p(Z_j) - Cn^{\epsilon-1/2}h^{-1}} |1(\hat{p}(Z_j) \geq \tau) - 1(p(Z_j) \geq \tau)| \text{ w.p.a. 1 for } \epsilon \in]0, 1/4[\\
&\leq n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{p(Z_j) - \frac{p(Z_j)}{\tau - Cn^{\epsilon-1/2}h^{-1}} Cn^{\epsilon-1/2}h^{-1}} |1(\hat{p}(Z_j) \geq \tau) - 1(p(Z_j) \geq \tau)| \text{ w.p.a. 1} \\
&= \frac{1}{1 - \frac{1}{C\tau n^{1/2-\epsilon}h^{-1}}} n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{p(Z_j)} |1(\hat{p}(Z_j) \geq \tau) - 1(p(Z_j) \geq \tau)| \\
&= O(1) n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{p(Z_j)} |1(\hat{p}(Z_j) \geq \tau \text{ and } p(Z_j) < \tau) - 1(p(Z_j) \geq \tau \text{ and } \hat{p}(Z_j) < \tau)| \\
&= O(1) n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{p(Z_j)} 1(\hat{p}(Z_j) \geq \tau \text{ and } p(Z_j) < \tau) + \\
&\quad + O(1) n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{p(Z_j)} 1(p(Z_j) \geq \tau \text{ and } \hat{p}(Z_j) < \tau) \\
&\leq O(1) n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{p(Z_j)} 1(p(Z_j) < \tau) + O(1) n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{p(Z_j)} 1(\hat{p}(Z_j) < \tau) \\
&\leq O(1) n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{p(Z_j)} 1(p(Z_j) < \tau) + \\
&\quad + O(1) n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{p(Z_j)} 1(p(Z_j) < \tau - Cn^{\epsilon-1/2}h^{-1}) \text{ w.p.a. 1}
\end{aligned}$$

where

$$E \left[n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{p(Z_j)} 1(p(Z_j) < \tau) \right] = n^{1/2} E \left[\frac{|\Upsilon_j|}{p(Z_j)} 1(p(Z_j) < \tau) \right] = o(1)$$

$$\begin{aligned}
& E \left[n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{p(Z_j)} \mathbf{1}(p(Z_j) < \tau - Cn^{\epsilon-1/2}h^{-1}) \right] \\
&= n^{1/2} E \left[\frac{|\Upsilon_j|}{p(Z_j)} \mathbf{1}(p(Z_j) < \tau - Cn^{\epsilon-1/2}h^{-1}) \right] \\
&= n^{1/2} E \left[\frac{|\Upsilon_j|}{p(Z_j)} \mathbf{1}(p(Z_j) < \tau(1 - Cn^{\epsilon-1/2}h^{-1}/\tau)) \right] \\
&= n^{1/2} E \left[\frac{|\Upsilon_j|}{p(Z_j)} \mathbf{1}(p(Z_j) < \tau(1 - o(1))) \right] \\
&\rightarrow n^{1/2} E \left[\frac{|\Upsilon_j|}{p(Z_j)} \mathbf{1}(p(Z_j) < \tau) \right] = o(1)
\end{aligned}$$

and by Markov's inequality $R_{T1} = o_p(1)$. Next,

$$\begin{aligned}
|R_{T2}| &\leq n^{1/2} E \left[\frac{|\Upsilon_j|}{p(Z_j)} |I_j - 1| \right] \\
&= n^{1/2} E \left[\frac{|\Upsilon_j|}{p(Z_j)} \mathbf{1}(p(Z_j) \leq \tau) \right] \\
&= n^{1/2} o(n^{-1/2}) = o(1)
\end{aligned}$$

and

$$\begin{aligned}
E[|R_{T3}|] &= E \left[\left| n^{-1/2} \sum_{j=1}^n \frac{\Upsilon_j - E[\Upsilon_j|Z_j]}{p(Z_j)} (I_j - 1) \right| \right] \\
&\leq n^{1/2} 2E \left[\frac{|\Upsilon_j|}{p(Z_j)} |I_j - 1| \right] \\
&= n^{1/2} o(n^{-1/2}) = o(1)
\end{aligned}$$

implying that $|R_{T3}| = o_p(1)$ as well by the Markov inequality. The linearization remainder is then

$$\begin{aligned}
|R_L| &= \left| n^{-1/2} \sum_{j=1}^n \frac{\Upsilon_j}{\hat{p}(Z_j) p^2(Z_j)} (\hat{p}(Z_j) - p(Z_j))^2 I_j \right| \\
&\leq n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{\hat{p}(Z_j) p^2(Z_j)} |\hat{p}(Z_j) - p(Z_j)|^2 I_j \\
&\leq n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{(p(Z_j) - Cn^{-1/2}h^{-1}) p^2(Z_j)} |\hat{p}(Z_j) - p(Z_j)|^2 I_j \\
&\leq n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{(\tau - Cn^{-1/2}h^{-1}) \tau p(Z_j)} |\hat{p}(Z_j) - p(Z_j)|^2 I_j \\
&= \frac{1}{\tau^2} (1 - Cn^{-1/2}h^{-1}/\tau)^{-1} n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{p(Z_j)} |\hat{p}(Z_j) - p(Z_j)|^2 I_j \\
&\leq \frac{2}{\tau^2} n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{p(Z_j)} |\hat{p}(Z_j) - p(Z_j)|^2 \leq \frac{2Cn^{-1}h^{-2}}{\tau^2} n^{1/2} n^{-1} \sum_{j=1}^n \frac{|\Upsilon_j|}{p(Z_j)} \\
&\leq \frac{2Cn^{-1}h^{-2}}{\tau^2} n^{1/2} \left(n^{-1} \sum_{j=1}^n \frac{|\Upsilon_j|^2}{p^2(Z_j)} \right)^{1/2} = \frac{2Cn^{-1}h^{-2}}{\tau^2} n^{1/2} O_p(1) \\
&= o(n^{-1/2}) n^{1/2} O_p(1) = o_p(1)
\end{aligned}$$

The “ U -statistic” term can be written as

$$\begin{aligned}
-R_U &= n^{-1/2} \sum_{j=1}^n (n-1)^{-1} \times \\
&\quad \times \sum_{i \neq j} \left(\frac{\Upsilon_j I_j}{p^2(Z_j)} (K_h(Z_i - Z_j) - E[K_h(Z_i - Z_j) | Z_j]) - \left(\frac{E[\Upsilon_j | Z_j]}{p(Z_j)} I_j - E \left[\frac{\Upsilon_j}{p(Z_j)} I_j \right] \right) \right) \\
&= n^{-1/2} \sum_{j=1}^n (n-1)^{-1} \times \\
&\quad \times \sum_{i \neq j} \left(\frac{\Upsilon_j I_j}{2p^2(Z_j)} + \frac{\Upsilon_i I_i}{2p^2(Z_i)} \right) (K_h(Z_i - Z_j) - E[K_h(Z_i - Z_j) | Z_j]) + \\
&\quad - \left(\frac{E[\Upsilon_i | Z_i]}{p(Z_i)} I_i - E \left[\frac{\Upsilon_i}{p(Z_i)} I_i \right] \right) \\
&= n^{1/2} \binom{n}{2}^{-1} \sum_{j=1}^n \sum_{i=j+1}^n U((\Upsilon_j, Z_j), (\Upsilon_i, Z_i))
\end{aligned}$$

where $K_h(z) = h^{-1}K(z/h)$ and

$$U((\Upsilon_j, Z_j), (\Upsilon_i, Z_i)) = \left(\frac{\Upsilon_j I_j}{2p^2(Z_j)} + \frac{\Upsilon_i I_i}{2p^2(Z_i)} \right) (K_h(Z_i - Z_j) - E[K_h(Z_i - Z_j) | Z_j]) + \left(\frac{E[\Upsilon_i | Z_i]}{p(Z_i)} I_i - E \left[\frac{\Upsilon_i}{p(Z_i)} I_i \right] \right).$$

Using the “ U -statistic” projection Theorem (e.g. Lemma 3.1 in (Powell, Stock, and Stoker 1989)), standard but tedious manipulations show that $R_U = o_p(1)$ under Assumptions 14 and 16. Finally, the bias term is

$$\begin{aligned} |R_B| &\leq n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{p^2(Z_j)} |p(Z_j) - E[\hat{p}(Z_j) | Z_j]| I_j \\ &\leq \tau^{-1} n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{p(Z_j)} |p(Z_j) - E[\hat{p}(Z_j) | Z_j]| I_j \\ &\leq \tau^{-1} n^{-1/2} \sum_{j=1}^n \frac{|\Upsilon_j|}{p(Z_j)} |p(Z_j) - E[\hat{p}(Z_j) | Z_j]| \\ &\leq \tau^{-1} n^{1/2} n^{-1} \sum_{j=1}^n \frac{|\Upsilon_j|}{p(Z_j)} Ch^{N_K} \text{ by Lemma 1} \end{aligned}$$

and $|R_B| = O_p(n^{1/2}h^{N_K}\tau^{-1}) = o_p(1)$ since $n^{1/2}h^{N_K}\tau^{-1} \rightarrow 0$ by Assumption 16.

To bound the R_{sec} term, let $S_\tau(t)$ be continuously differentiable in t for all $\tau \neq 0$ and such that (i) $1(t \geq \tau) = 0 \Leftrightarrow S_\tau(t) = 0$ (ii) $1(t \geq \tau) = 1 \Leftrightarrow S_{2\tau}(t) = 1$ (iii) $0 \leq S_\tau(t) \leq 1$. (iv) $\sup_{t \in \mathbb{R}} |dS_\tau(t)/dt| = O(\tau)$. We then decompose $\hat{Q}(\theta^*, \alpha)$ as

$$\hat{Q}(\theta^*, \alpha) = \hat{Q}_S(\theta^*, \alpha) + R_S(\alpha)$$

where $\hat{Q}_S(\theta^*, \alpha)$ is continuous in α while $R_S(\alpha)$ may not be and are given by

$$\begin{aligned} \hat{Q}_S(\theta^*, \alpha) &= n^{-1} \sum_{j=1}^n \frac{\Upsilon(X_j, Y_j, W_j, \theta^*, \alpha)}{\hat{p}(m(W_j, \alpha) | \alpha)} S_\tau(\hat{p}(m(W_j, \alpha) | \alpha)) \\ R_S(\alpha) &= n^{-1} \sum_{j=1}^n \frac{\Upsilon(X_j, Y_j, W_j, \theta^*, \alpha)}{\hat{p}(m(W_j, \alpha) | \alpha)} (1(\hat{p}(m(W_j, \alpha) | \alpha) \geq \tau) - S_\tau(\hat{p}(m(W_j, \alpha) | \alpha))). \end{aligned}$$

The remainder $R_S(\alpha)$ satisfies $\sup_{\alpha \in \mathcal{A}} \|R_S(\alpha)\| = o_p(n^{-1/2})$ since

$$\sup_{\alpha \in \mathcal{A}} \|R_S(\alpha)\| \leq \sup_{\alpha \in \mathcal{A}} n^{-1} \sum_{j=1}^n \frac{\|\Upsilon(X_j, Y_j, W_j, \theta^*, \alpha)\|}{p(m(W_j, \alpha) | \alpha) - n^{-1/2}h^{-1}} 1(\hat{p}(m(W_j, \alpha) | \alpha) < 2\tau) = o_p(n^{-1/2})$$

By Assumption 17 and the same arguments as used for R_{T1} . We can then write R_{sec} as

$$\begin{aligned}
R_{\text{sec}} &= n^{1/2} \left(\left(\hat{Q}(\theta^*, \hat{\alpha}) - Q(\theta^*, \hat{\alpha}) \right) - \left(\hat{Q}(\theta^*, \alpha^*) - Q(\theta^*, \alpha^*) \right) \right) \\
&= n^{1/2} \left(\left(\hat{Q}_S(\theta^*, \hat{\alpha}) - Q(\theta^*, \hat{\alpha}) \right) - \left(\hat{Q}_S(\theta^*, \alpha^*) - Q(\theta^*, \alpha^*) \right) \right) + o_p(1) \\
&= \left(\frac{\partial \hat{Q}(\theta^*, \bar{\alpha})}{\partial \alpha'} - \frac{\partial Q(\theta^*, \bar{\alpha})}{\partial \alpha'} \right) n^{1/2} (\hat{\alpha} - \alpha^*) + o_p(1) \tag{12}
\end{aligned}$$

for some mean value $\bar{\alpha}$. We then decompose $\frac{\partial}{\partial \alpha'} \hat{Q}_S(\theta^*, \alpha)$ as

$$\frac{\partial}{\partial \alpha'} \hat{Q}_S(\theta^*, \alpha) = D_1 + D_2 + R_{DS}$$

where

$$\begin{aligned}
D_1 &= n^{-1/2} \sum_{j=1}^n \left(\frac{\frac{\partial}{\partial \alpha'} \Upsilon(X_j, Y_j, W_j, \theta, \alpha)}{\hat{p}(m(W_j, \alpha) | \alpha)} \right) S_\tau(\hat{p}(m(W_j, \alpha) | \alpha)) \\
D_2 &= -n^{-1/2} \sum_{j=1}^n \left(\frac{\Upsilon(X_j, Y_j, W_j, \theta, \alpha)}{\hat{p}^2(m(W_j, \alpha) | \alpha)} \frac{\partial}{\partial \alpha'} \hat{p}(m(W_j, \alpha) | \alpha) \right) S_\tau(\hat{p}(m(W_j, \alpha) | \alpha)) \\
R_{DS} &= n^{-1/2} \sum_{j=1}^n \frac{\Upsilon(X_j, Y_j, W_j, \theta, \alpha)}{\hat{p}(m(W_j, \alpha) | \alpha)} \frac{\partial S_\tau(\hat{p}(m(W_j, \alpha) | \alpha))}{\partial \alpha'}.
\end{aligned}$$

The R_{DS} term is negligible, since

$$\begin{aligned}
&\left\| n^{-1/2} \sum_{j=1}^n \frac{\Upsilon(X_j, Y_j, W_j, \theta, \alpha)}{\hat{p}(m(W_j, \alpha) | \alpha)} \frac{\partial S_\tau(\hat{p}(m(W_j, \alpha) | \alpha))}{\partial \alpha'} \right\| \\
&\leq n^{-1/2} \sum_{j=1}^n \frac{\|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\|}{\tau} \left| \frac{\partial S_\tau(\hat{p}(m(W_j, \alpha) | \alpha))}{\partial \alpha'} \right| \\
&\leq n^{-1/2} \sum_{j=1}^n \frac{\|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\|}{\tau} C\tau 1(p(m(W_j, \alpha) | \alpha) > \tau) \\
&= Cn^{-1/2} \sum_{j=1}^n \|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\| 1(p(m(W_j, \alpha) | \alpha) > \tau) \\
&= Cn^{-1/2} \sum_{j=1}^n \frac{\|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\|}{p(m(W_j, \alpha) | \alpha)} p(m(W_j, \alpha) | \alpha) 1(p(m(W_j, \alpha) | \alpha) > \tau)
\end{aligned}$$

$$\begin{aligned}
&\leq Cn^{1/2}n^{-1}\sum_{j=1}^n\frac{\|\Upsilon(X_j, Y_j, W_j, \theta, \alpha)\|}{p(m(W_j, \alpha)|\alpha)}\mathbb{1}(p(m(W_j, \alpha)|\alpha) > \tau) \\
&\quad \text{since } p(z|\alpha) \text{ is bounded by Assumption 15} \\
&= n^{1/2}o_p(n^{-1/2}) = o_p(1) \text{ by Markov's inequality and Assumption 17.}
\end{aligned}$$

Now, the terms D_1 and D_2 can be handled through the same techniques as the ones used to show uniform convergence of $\hat{Q}(\theta, \hat{\alpha})$ after noting that trimming by $S_\tau(\hat{p}(m(W_j, \alpha)|\alpha))$ is asymptotically equivalent to trimming by $\mathbb{1}(\hat{p}(m(W_j, \alpha)|\alpha) \geq \tau)$.

Under Assumption 19, and by using an expansion of the form

$$\begin{aligned}
D_1 &= n^{-1/2}\sum_{j=1}^n\frac{\frac{\partial}{\partial\alpha'}\Upsilon(X_j, Y_j, W_j, \theta, \alpha)}{p(m(W_j, \alpha)|\alpha)}S_\tau(\hat{p}(m(W_j, \alpha)|\alpha)) + \\
&\quad -n^{-1/2}\sum_{j=1}^n\frac{\frac{\partial}{\partial\alpha'}\Upsilon(X_j, Y_j, W_j, \theta, \alpha)}{p(m(W_j, \alpha)|\alpha)}\frac{(\hat{p}(m(W_j, \alpha)|\alpha) - p(m(W_j, \alpha)|\alpha))}{\hat{p}(m(W_j, \alpha)|\alpha)}S_\tau(\hat{p}(m(W_j, \alpha)|\alpha)) \\
D_2 &= n^{-1/2}\sum_{j=1}^n\frac{\Upsilon(X_j, Y_j, W_j, \theta, \alpha)}{p(m(W_j, \alpha)|\alpha)}\left(1 - \frac{(\hat{p}(m(W_j, \alpha)|\alpha) - p(m(W_j, \alpha)|\alpha))}{\hat{p}(m(W_j, \alpha)|\alpha)}\right) \times \\
&\quad \times \frac{\frac{\partial}{\partial\alpha'}p(m(W_j, \alpha)|\alpha) + (\frac{\partial}{\partial\alpha'}\hat{p}(m(W_j, \alpha)|\alpha) - \frac{\partial}{\partial\alpha'}p(m(W_j, \alpha)|\alpha))}{\hat{p}(m(W_j, \alpha)|\alpha)}
\end{aligned}$$

it can be shown that $D_1 \xrightarrow{p} E[(\partial\Upsilon(X_j, Y_j, W_j, \theta, \alpha)/\partial\alpha')/p(m(W_j, \alpha)|\alpha)]$ and $D_2 \xrightarrow{p} E[(\Upsilon(X_j, Y_j, W_j, \theta, \alpha)/p^2(m(W_j, \alpha)|\alpha))(\partial p(m(W_j, \alpha)|\alpha)/\partial\alpha')]$ uniformly in α for $\alpha \in \mathcal{A}$. (The convergence rate of $\partial\hat{p}(m(W_j, \alpha)|\alpha)/\partial\alpha' - \partial p(m(W_j, \alpha)|\alpha)/\partial\alpha'$ is obtained as in the proof of Lemma 1, with N_K replaced by $N_K - 1$.) This implies, by Assumption 18, that

$$\sup_{\alpha \in \mathcal{A}} \left(\frac{\partial\hat{Q}(\theta^*, \alpha)}{\partial\alpha'} - \frac{\partial Q(\theta^*, \alpha)}{\partial\alpha'} \right) \xrightarrow{p} 0$$

and by Equation (12) and the fact that $\hat{\alpha} - \alpha^* = O_p(n^{-1/2})$, we have that $R_{\text{sec}} = o_p(1)$.

Having bounded all remainder terms, we note that the N term clearly satisfies

$$N = n^{-1/2}\sum_{j=1}^n\psi_\theta(X_j, Y_j, W)$$

where $E[\psi_\theta(X_j, Y_j, W)\psi'_\theta(X_j, Y_j, W)]$ is finite under Assumption 13.

By a mean-value expansion, the N_α term is equal to

$$N_\alpha = \frac{\partial Q(\theta^*, \bar{\alpha})}{\partial \alpha'} n^{1/2} (\hat{\alpha} - \alpha^*)$$

for some mean value $\bar{\alpha}$. Since $\hat{\alpha} \xrightarrow{p} \alpha^*$ and therefore $\bar{\alpha} \xrightarrow{p} \alpha^*$, Assumption 18 implies that $\partial Q(\theta^*, \bar{\alpha}) / \partial \alpha' \xrightarrow{p} \partial Q(\theta^*, \alpha) / \partial \alpha'$.

By standard results (such as Theorem 3.1 in (Newey and McFadden 1994)), Assumptions 8 and 9 imply that the first-step estimate $\hat{\alpha}$ is a root n consistent estimator of α^* with influence function equal to

$$\psi_\alpha(x, w) = - \left(E \left[\frac{\partial m(W, \alpha^*)}{\partial \alpha} \frac{\partial m(W, \alpha^*)}{\partial \alpha'} \right] \right)^{-1} \frac{\partial m(w, \alpha^*)}{\partial \alpha} (x - m(w, \alpha^*))$$

and such that $E[\psi_\alpha(X, W) \psi'_\alpha(X, W)]$ exists. Hence, we can write

$$N_\alpha = n^{-1/2} \sum_{j=1}^n \frac{\partial Q(\theta^*, \alpha)}{\partial \alpha'} \psi_\alpha(X_j, W_j).$$

We have just established that

$$n^{1/2} \hat{Q}(\theta^*, \hat{\alpha}) = n^{-1/2} \sum_{j=1}^n \left(\psi_\theta(X_j, Y_j, W_j) + \frac{\partial Q(\theta^*, \alpha)}{\partial \alpha'} \psi_\alpha(X_j, W_j) \right) + o_p(1)$$

and by the finiteness of $E[\psi_\theta(X_j, Y_j, W_j) \psi'_\theta(X_j, Y_j, W_j)]$ and $E[\psi_\alpha(X_j, W_j) \psi'_\alpha(X_j, W_j)]$, the Cauchy-Schwartz inequality, Assumptions 8 and the Lindeberg-Levy Central Limit Theorem, this sum is asymptotically normal. By Equations (10), (11) and the Slutsky Theorem, the conclusion of the Theorem follows. \blacksquare

4 Comparison with Hausman, Newey, Ichimura and Powell (1991)

In the polynomial case, the proposed estimator can be shown to have the same influence function as the IV estimator described in (Hausman, Newey, Ichimura, and Powell 1991) simply by choosing suitable weighting functions, since both estimators

rely on the same functional equations as a starting point (Equations (6) and (7) in the text). More specifically, in Section 3.2.2, the weighting functions must be selected such that

$$\begin{aligned} V_y(z, \theta) &= p(z) E[\mathbf{Z}\mathbf{Z}']^{-1} \mathbf{z} \\ V_{xy}(z, \theta) &= p(z) E[\mathbf{Z}_+\mathbf{Z}'_+]^{-1} \mathbf{z}_+ \end{aligned}$$

where $\mathbf{Z} = [1, Z, \dots, Z^{\bar{k}}]'$, $\mathbf{z} = [1, z, \dots, z^{\bar{k}}]'$, $\mathbf{Z}_+ = [1, Z, \dots, Z^{\bar{k}+1}]'$ and $\mathbf{z}_+ = [1, z, \dots, z^{\bar{k}+1}]'$. (In that special case, $p(z)$ would not need to be estimated, since it would cancel with the division by $p(z)$ in the moment conditions.) The fact that there exists one choice of weighting functions reaching the same asymptotic variance as in (Hausman, Newey, Ichimura, and Powell 1991) shows that the proposed estimator can be at least as efficient.

5 Root n consistent estimation under more general conditions

The construction of the moment conditions in Section 3.2 uses, as a starting point, a set of smooth and rapidly decaying functions \mathcal{G} , described in Definition 1. This appendix shows how it is possible to define a larger set \mathcal{G}' that enables root n consistent estimation in even more general settings, for instance allowing for distributions of the disturbance U whose moment generating function only exists over a finite interval.

In that case the enlarged set \mathcal{G}' should also contain functions that can be written as linear combinations of products of polynomials and functions of the form $\sigma(a\zeta + b)$ for $a, b \in \mathbb{R}$ and

$$\sigma(\zeta) = \exp(-\cos^{-2}(\zeta\pi/2)) \mathbf{1}(|\zeta| \leq 1). \quad (13)$$

The function $\sigma(\zeta)$ is compactly supported and infinitely many times differentiable (including at $|\zeta| = 1$). It is a refinement over the well-known function $\exp(-(1 - \zeta^2)^{-1}) \mathbf{1}(|\zeta| \leq 1)$ that improves the rate of decay of the inverse Fourier transform of $\sigma(\zeta)$

to $\exp(-c|z|)$ for some $c > 0$ instead of merely faster than $|z|^{-k}$ for any $k \in \mathbb{N}$, as shown in Lemma 2 and Theorem 1 at the end of this section.

The treatment of Section 3.2 carries over with this alternative set \mathcal{G}' with one exception. The fact that $\sigma(\zeta)$ is compactly supported (and therefore that there exists compactly supported $\lambda(\zeta)$ in the set \mathcal{G}') enables the use of Lemma 5 in the case where the moment generating function of U only exists on a finite interval. In that case, the Taylor series of the characteristic function $\phi(\zeta)$ of U only converges in a finite interval and it is crucial that a compactly supported $\lambda(\zeta)$ be used to “eliminate” the region where the Taylor series does not converge. Furthermore, the fact that the inverse Fourier transform of $\sigma(\zeta)$ decays rapidly is helpful to ensure that the functions $r_y(z, \theta)$, $r_{xy}(z, \theta)$ and $r_{1y}(z, \theta)$ are rapidly decaying in z so that expectations of the form $E[Yr_y(Z, \theta)/p(Z)]$, for instance, exist. We can then state the following Corollary to Theorem 1.

Assumption 1 $E[e^{tU}]$ exists for t in some neighborhood of the origin.

Corollary 1 Under Assumptions 1, 2(i), 4, 5-6 and 1, if $Q(x, y, z, \theta, p)$ is as defined in Equation (16) and Section 3.2 (with \mathcal{G} replaced by \mathcal{G}') then there exists a compact set $\Theta \subset \mathbb{R}^{N_\theta}$ containing θ^* in its interior such that $\theta = \theta^*$ is the only solution to $E[Q(X, Y, Z, \theta, p)] = 0$ in Θ . Assumption 1 is unnecessary when $r_{y,o}(z, \theta)$, $r_{xy,o}(z, \theta)$ and $r_{1y,o}(z, \theta)$ are “empty” or when $r_{y,s}(z, \theta)$, $r_{xy,s}(z, \theta)$, $r_{1y,s}(z, \theta)$ and $r_{1y,o}(z, \theta)$ are “empty”.

The set \mathcal{G} can also be enlarged by including functions that have a $(\mathbf{i}\zeta)^{-1}$ prefactor. The treatment of Section 3.2 again carries over to this case, except that the proof of Lemma 5 needs to be adapted (because the absolute integrability assumption made in Lemma 5 does not automatically hold) by employing the following technique. The left-hand side of Equation 67 can be decomposed as

$$\int_{-\eta}^{\eta} \lambda(\zeta) \phi(0) d\zeta = \int_{-\eta}^{\eta} \left(\lambda(\zeta) - \frac{C}{\mathbf{i}\zeta} \right) \phi(0) d\zeta + \int_{-\eta}^{\eta} \frac{C}{\mathbf{i}\zeta} \phi(0) d\zeta$$

where C is a constant such that $(\lambda(\zeta) - C/\mathbf{i}\zeta)$ is absolutely integrable and where $\int_{-\eta}^{\eta} (C\phi(0) / (\mathbf{i}\zeta)) d\zeta = 0$ in the Cauchy principal value sense. Next, $\lambda(\zeta)$ can be replaced by $(\lambda(\zeta) - C/\mathbf{i}\zeta)$ in the right-hand side of Equation 67. The remainder of the proof is unchanged.

Lemma 2 *Let $\sigma(\zeta)$ be the Fourier transform of $s(z)$. For $\alpha \in \mathbb{R}^+$ and $\gamma \in \mathbb{N}$, if*

$$\sum_{t=0}^{\infty} \frac{\alpha^t}{t!} \int \left| \frac{d^{\gamma t} \sigma(\zeta)}{d\zeta^{\gamma t}} \right| d\zeta < \infty$$

then, for some $C > 0$,

$$|s(z)| \leq C \exp(-\alpha |z|^\gamma).$$

Proof. Let $T(z) = \exp(\alpha z^\gamma)$. Since the radius of convergence of the Taylor series of the exponential function is infinite, we can also write $T(z) = \sum_{t=0}^{\infty} \alpha^t z^{\gamma t} / t!$ for all $z \in \mathbb{R}$. Let Θ denote the linear operator defined by

$$\Theta \sigma(\zeta) = \sum_{t=0}^{\infty} \frac{\alpha^t (-\mathbf{i})^{\gamma t} d^{\gamma t} \sigma(\zeta)}{t! d\zeta^{\gamma t}}.$$

Since the Fourier transform of $z^t s(z)$ is $(-\mathbf{i})^t d^t \sigma(\zeta) / d\zeta^t$, the Fourier transform of $T(z) s(z)$ is $\Theta \sigma(\zeta)$. We can then write, for $z \geq 0$,

$$\begin{aligned} |s(z)| &= \frac{1}{|T(z)|} |T(z) s(z)| = \frac{1}{|T(z)|} \left| \int \Theta \sigma(\zeta) e^{-\mathbf{i}\zeta z} d\zeta \right| \leq \frac{1}{|T(z)|} \int |\Theta \sigma(\zeta)| d\zeta \\ &= \frac{1}{|T(z)|} \int \left| \sum_{t=0}^{\infty} \frac{\alpha^t (-\mathbf{i})^{\gamma t} d^{\gamma t} \sigma(\zeta)}{t! d\zeta^{\gamma t}} \right| d\zeta \leq \frac{1}{|T(z)|} \sum_{t=0}^{\infty} \frac{\alpha^t}{t!} \int \left| \frac{d^{\gamma t} \sigma(\zeta)}{d\zeta^{\gamma t}} \right| d\zeta \\ &= \frac{C}{|T(z)|} = C \exp(-\alpha |z|^\gamma). \end{aligned}$$

with

$$C = \sum_{t=0}^{\infty} \frac{\alpha^t}{t!} \int \left| \frac{d^{\gamma t} \sigma(\zeta)}{d\zeta^{\gamma t}} \right| d\zeta < \infty.$$

For $z < 0$, we can similarly write

$$\begin{aligned} |s(z)| &= \frac{1}{|T(-z)|} |T(-z) s(z)| = \frac{1}{|T(-z)|} \left| \int \Theta \sigma(\zeta) e^{\mathbf{i}\zeta z} d\zeta \right| \\ &\leq \frac{1}{|T(-z)|} \int |\Theta \sigma(\zeta)| d\zeta \leq \frac{C}{|T(|z|)|} = C \exp(-\alpha |z|^\gamma). \end{aligned}$$

■

Theorem 1 *The inverse Fourier transform $s(\zeta)$ of the function*

$$\sigma(\zeta) = \exp(-\cos^{-2}(\zeta)) \mathbf{1}_{(|\zeta| \leq \pi/2)}$$

is such that $|s(z)| \leq C \exp(-\alpha|z|)$ for $\alpha \in [0, 1/3[$ and some positive $C < \infty$.

Proof. The proof consists of verifying that $\sigma(\zeta)$ satisfies the hypothesis of Lemma 2. The t -th derivative of $\exp(-\cos^{-2}(\zeta))$ consists of a sum of at most 3^t terms of the form

$$C \exp(-\cos^{-2}(\zeta)) \cos^{-p}(\zeta) \sin^q(\zeta), \quad (14)$$

where $q \geq 0$, $0 \leq p \leq 2t$, and $|C| \leq 1 + t$. Since $p \leq 2t$, $|\sin(\zeta)| \leq 1$ and $X^t \exp(-X) \leq t^t \exp(-t)$ for all $X \in \mathbb{R}^+$ and all $t \in \mathbb{N}$, we have

$$\begin{aligned} & \left| \exp(-\cos^{-2}(\zeta)) \cos^{-p}(\zeta) \sin^q(\zeta) \right| \\ & \leq \exp(-\cos^{-2}(\zeta)) \cos^{-2t}(\zeta) \\ & \leq t^t \exp(-t). \end{aligned}$$

Consequently, for some $C > 0$,

$$\begin{aligned} \sum_{t=0}^{\infty} \frac{\alpha^t}{t!} \int \left| \frac{d^t \sigma(\zeta)}{d\zeta^t} \right| d\zeta & \leq C \sum_{t=0}^{\infty} \frac{\alpha^t}{t!} 3^t (1 + 2t) t^t \exp(-t) \\ & \leq C \sum_{t=0}^{\infty} \alpha^t (3 + \varepsilon_1)^t \frac{t^t \exp(-t)}{t!} \text{ for any } \varepsilon_1 > 0 \\ & \leq C \sum_{t=0}^{\infty} ((3 + \varepsilon_2) \alpha)^t, \text{ for any } \varepsilon_2 > 0 \end{aligned}$$

which converges if $\alpha < 1/3$, choosing $\varepsilon_2 < 1/\alpha - 3$.

■

6 Details of the Monte Carlo Simulations

We consider three different specifications, namely, a polynomial, a rational fraction and a probit model. In all cases, the mismeasured regressor X is generated from

$$\begin{aligned} X &= X^* + \Delta X \\ X^* &= Z - U \end{aligned}$$

with Z, U and ΔX drawn from the following distributions

$$Z \sim N(0, 1), \quad U \sim N(0, 1/4), \quad \Delta X \sim N(0, 1/4).$$

Note that the ratio of the standard deviation of the measurement error ΔX to the standard deviation of the true regressor X^* is $(1/2) / \sqrt{(1 + 1/4)} \approx 0.45$, so that the measurement error is fairly large. In addition the R^2 of the equation $X = Z - U + \Delta X$ is $2/3$, indicating that the “strength” of the instrument is of a magnitude that is fairly typical for applications.

The dependent variable Y is generated from

$$Y = g(X^*, \theta) + \Delta Y,$$

where the functional form of $g(x^*, \theta)$ and the distribution of ΔY differ for each model.

For the kernel density estimation of the density of Z , an infinite order kernel is used, which has the desirable property that the estimation bias decays faster than any power of the bandwidth h as $h \rightarrow 0$. The specific kernel $K(z)$ used is the inverse Fourier transform of

$$\kappa(\zeta) = \left(\int_{-\infty}^{\infty} \sigma\left(\frac{\xi + 2}{1.9}\right) d\xi \right)^{-1} \int_{-\infty}^{\zeta} \left(\sigma\left(\frac{\xi + 2}{1.9}\right) - \sigma\left(\frac{\xi - 2}{1.9}\right) \right) d\xi \quad (15)$$

where $\sigma(\zeta)$ is given by $\sigma(\zeta) = \exp(-\cos^{-2}(\zeta\pi/2)) 1(|\zeta| \leq 1)$. The prefactor ensures that $\kappa(0) = 1$ and therefore that $\int K(z) dz = 1$, as should be the case for a valid kernel. It is the fact that $\kappa(\zeta)$ is constant over $[-0.1, 0.1]$ which makes $K(z)$ an

infinite order kernel. The function $\kappa(\zeta)$ inherits the smoothness of the function $\sigma(\zeta)$, thus ensuring that $K(z)$ is rapidly decaying.

The “optimal” bandwidth parameter h and trimming parameter τ are chosen so as to minimize the GMM objective function associated with the proposed estimator evaluated at θ^* . In our simulation study, this is achieved by scanning values of h from 0.5 to 1.5 in multiplicative increments of 1.1 and values of τ from 0.005 to 0.05 in multiplicative increments of 1.5. The GMM objective function for the given level of smoothing and trimming is then evaluated for 50 replicated samples of 1000 observations and averaged. The “optimal” bandwidth and trimming parameters are found to be $h = 0.585$ and $\tau = 0.026$. The “optimal” values obtained for all three models considered are the same, within the accuracy implied by the spacings between the consecutive values of h or τ scanned. This is perhaps not surprising since the distribution of Z to be nonparametrically estimated is common across all the models. Although the bandwidth and trimming parameters were optimized using knowledge of the experimental set up (i.e. the true value θ^*), the simulation results should not be overly optimistic. Semiparametric estimators tend to be less sensitive to the exact choice of the bandwidth than fully nonparametric estimators are. Also, keeping the trimming parameter fixed over all replications demand a more aggressive trimming to ensure that all replications give reasonable estimates. Empirical researchers would typically fine-tune the trimming parameter for each given sample and could probably do better, on average, than the current simulations show.

The finite sample properties of the proposed estimator (for the given values of h and τ) are studied by drawing 5000 samples of 1000 independent observations. As a point of comparison, we also calculate the standard instrumental variable estimator using $\partial g(Z, \theta) / \partial \theta$ as a vector of instruments and X as the regressor in addition to a standard (nonlinear) least squares estimator using X as the regressor, although both of these estimators are clearly biased in the presence of measurement error.

Let $\hat{\theta}_k$ denote any element of $\hat{\theta}$, the parameter vector estimated by any one the three estimators, and let θ_k^* denote any element of θ^* , the true value of the parameter vector. The three estimators are compared on the basis of their bias, standard deviation, root mean square error and overall root mean square error, given, respectively, by

$$\begin{aligned} \text{Bias} &= E[\hat{\theta}_k] - \theta_k^*, \\ \text{Std. Dev.} &= \left(E \left[\left(\hat{\theta}_k - E[\hat{\theta}_k] \right)^2 \right] \right)^{1/2}, \\ \text{RMSE} &= \left(E \left[\left(\hat{\theta}_k - \theta_k^* \right)^2 \right] \right)^{1/2} \\ \text{RMSE}_{\text{all}} &= \left(\text{tr} E \left[\left(\hat{\theta} - \theta^* \right) \left(\hat{\theta} - \theta^* \right)' \right] \right)^{1/2}. \end{aligned}$$

Note that the last quantity is a convenient summary measure of the overall performance of an estimator.

Although our estimator is based on moment conditions which have zero expectation at the true value of the parameter vector, it is perfectly normal that it could be biased in a finite sample. First, the moment conditions used for estimation are non-linear in θ , and it is well-known that, in this context, just identified GMM exhibits a bias of order n^{-1} , where n is sample size (see, for instance, (Newey and Smith 2004)). Second, the implementation of the estimator relies on kernel smoothing and trimming, two techniques which introduce their own bias. Simulations prove to be a helpful tool to verify that the potential presence of such biases does not overcome the benefits of the elimination of the measurement error-induced bias. We now describe the specifics of each simulation.

6.1 Polynomial Model

This model is defined by

$$g(x^*, \theta) = \theta_1 + \theta_2 x^* + \theta_3 (x^*)^2 + \theta_4 (x^*)^3$$

$$\Delta Y \sim N(0, 1/4)$$

where $\theta_1 = 1$, $\theta_2 = 1$, $\theta_3 = 0$ and $\theta_4 = -0.5$. The Fourier transform of this polynomial contains no ordinary function component, but only a linear combination of delta function derivatives, and therefore the weighting functions $\omega(\zeta)$ and $\varpi(\zeta)$ do not need to be introduced. Following the discussion of Section 3.2.2, the weighting functions $\nu_{y,j}(\zeta, \theta)$ and $\nu_{xy,j}(\zeta, \theta)$ for $j = 0, \dots, \bar{k}$ (where $\bar{k} = 3$) are chosen to be of the form⁷

$$\nu_{y,j}(\zeta, \theta) = (\mathbf{i}\zeta)^j \exp\left(-\frac{1}{2} \left(\frac{\zeta}{(1.1)\pi/2}\right)^2\right) \quad (16)$$

$$\nu_{xy,j}(\zeta, \theta) = (\mathbf{i}\zeta)^j \exp\left(-\frac{1}{2} \left(\frac{\zeta}{(1.1)\pi/2}\right)^2\right) \quad (17)$$

Table 1 compares the performance of the proposed estimator relative to IV and OLS. Although the bias of the proposed estimator is slightly larger than the one of IV for three of the coefficients (θ_1, θ_3 and θ_4), the bias of IV for the remaining coefficient (θ_2) is overwhelmingly large, making the overall performance of IV poor. This is best illustrated by substituting the expected values⁸ of the coefficients obtained from each estimator into the polynomial specification and by overlapping the graph of each resulting polynomial over the “true” model specification. As seen in Figure 1a), the proposed estimator is much closer to the true specification than any of the other estimators. While the reduction in bias achieved with our estimator comes at the expense of increased standard errors for some coefficients, the overall RMSE (the

⁷Since the ordinary part of the Fourier transform of $g(x^*, \theta)$ is zero ($\gamma_o(\zeta, \theta) = 0$), there is no need to ensure that the $\nu_{y,j}(\zeta, \theta)$ and $\nu_{xy,j}(\zeta, \theta)$ are orthogonal to the ordinary part. Hence we can specify $\nu_{y,j}(\zeta, \theta)$ and $\nu_{xy,j}(\zeta, \theta)$ directly without first introducing the functions $\mu_{y,j}, \mu_{xy,j} \in \mathcal{S}_0 \cap \mathcal{C}$, as done in Section 3.2.2.

⁸That is, their average over the replications.

column labeled by “all” in Table 1) is still lower for the proposed estimator than for the other two estimators.

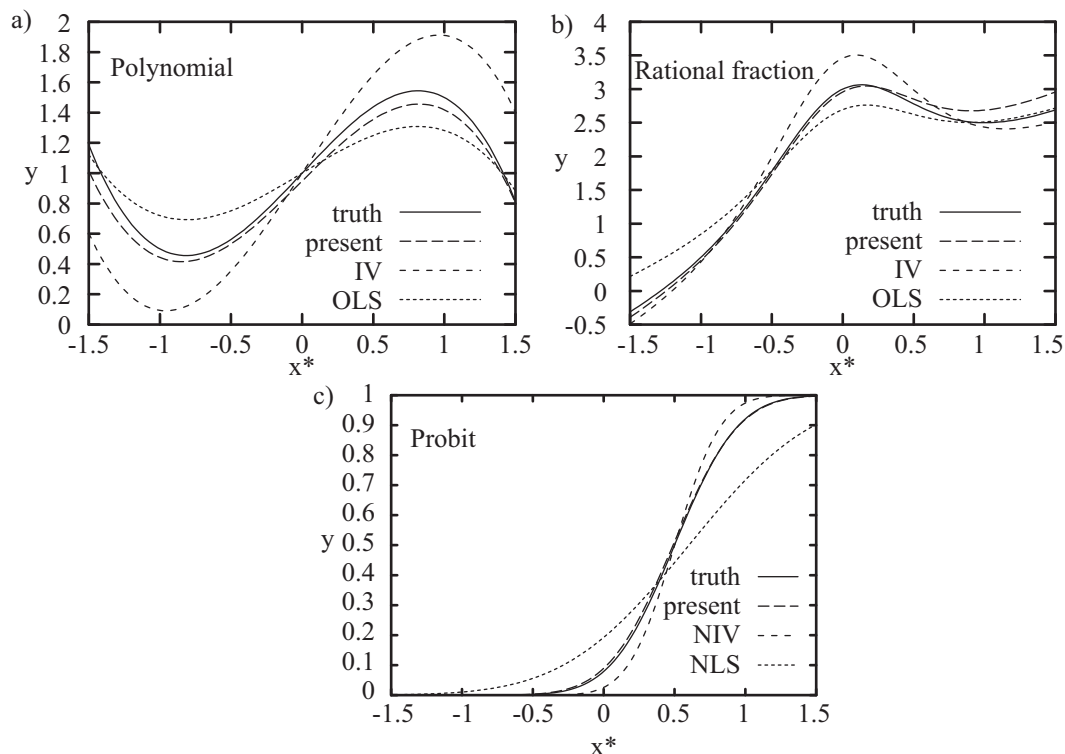


Figure 1: Graphical representation of the bias of each estimator studied. Note that for the probit model in c), the curve for the standard NIV estimator excludes the 50% of the replications that do not yield a finite estimate of θ_2 . The actual performance of NIV is therefore far worse than indicated by the graph.

6.2 Rational fraction

The second example is a specification of the form

$$g(x^*, \theta) = \theta_1 + \theta_2 x^* + \frac{\theta_3}{(1 + (x^*)^2)^2}$$

$$\Delta Y \sim N(0, 1/4)$$

where $\theta_1 = 1$, $\theta_2 = 1$ and $\theta_3 = 2$. The Fourier transform of $g(x^*, \theta)$ in this case contains both an ordinary and a singular component:

$$\gamma(\zeta, \theta) = \theta_1 2\pi \delta(\zeta) - \theta_2 2\pi i \delta^{(1)}(\zeta) + \theta_3 \frac{\pi}{2} (1 + |\zeta|) e^{-|\zeta|}. \quad (18)$$

As discussed in Section 3.2.2, to determine the singular component, we need to construct some functions $\mu_{y,j}, \mu_{xy,j} \in \mathcal{S}_0 \cap \mathcal{C}$. In the case of $\mu_{y,j}$, this is accomplished using Definition 2 with $\lambda(\zeta)$ set to

$$\lambda(\zeta) = (\mathbf{i}\zeta)^j \exp\left(-\frac{1}{2} \left(\frac{\zeta}{(2.1) \pi/2}\right)^2\right) - 2(\mathbf{i}2\zeta)^j \exp\left(-\frac{1}{2} \left(\frac{2\zeta}{(2.1) \pi/2}\right)^2\right) \quad (19)$$

for $j = 1, \dots, 2$. The function $\mu_{xy,j}$ is obtained similarly, with $j = 1, \dots, 3$.

The ordinary part in Equation (18) depends on a single parameter and, consequently, only the scale of the ordinary part needs to be determined. As explained in Section 3.2.1, the vector of weighting function ω associated with the “shape” of the regression function is therefore not needed. Only the weighting function ϖ , associated with the “scale” is. Definition 2 is then used to obtain $\varpi \in \mathcal{S}_1 \cap \mathcal{C}$, with $\lambda(\zeta)$ set to

$$\lambda(\zeta) = (\mathbf{i}\zeta)^2 \exp\left(-\frac{1}{2} \left(\frac{\zeta}{(1.6) \pi/2}\right)^2\right) \times \left(\int (\mathbf{i}\xi)^2 \exp\left(-\frac{1}{2} \left(\frac{\xi}{(1.6) \pi/2}\right)^2\right) d\xi\right)^{-1}.$$

The prefactor $(\mathbf{i}\zeta)^2$ ensures that the singular parts do not affect the estimation of the ordinary part.

Table 2 summarizes the results of the simulations for the rational fraction model and clearly illustrates the bias-correcting power of the proposed estimator. While the IV estimator exhibits a fortuitously low bias on the θ_2 parameter, it clearly fails to produce unbiased estimates of the coefficient on the nonlinear term (θ_3). As is seen in Figure 1b), the proposed estimator provides a nearly unbiased estimate of the height of the nonlinear component of the specification, unlike IV, which overestimates it, and OLS, which underestimates it. The proposed estimator has, overall, a bias of only about 10% for this model. Since our estimator typically exhibits larger standard error than both IV and OLS, it is instructive to verify whether it still comes out ahead when both bias and variance are taken into account. Indeed, the overall RMSE clearly points towards the proposed estimator as the best alternative.

6.3 Probit

The probit model can be written as a regression model with the following specification

$$g(x^*, \theta) = \frac{1}{2} (1 + \operatorname{erf}(\theta_1 + \theta_2 x^*)) \quad (20)$$

where we set $\theta_1 = -1$ and $\theta_2 = 2$ and where the distribution of ΔY conditional on $X^* = x^*$ is given by

$$\Delta Y = \begin{cases} 1 - g(x^*, \theta) & \text{with probability } g(x^*, \theta) \\ -g(x^*, \theta) & \text{with probability } 1 - g(x^*, \theta) \end{cases} .$$

The Fourier transform of $g(x^*, \theta)$ given in Equation (20) is

$$\gamma(\zeta, \theta) = \pi \delta(\zeta) - \frac{1}{\mathbf{i}\zeta} \exp\left(-\mathbf{i}\zeta \frac{\theta_1}{\theta_2} - \frac{\zeta^2}{4\theta_2^2}\right) \quad (21)$$

Since the singular component of Equation (21) does not depend on θ , it provides no information to estimate the model and we therefore only need to consider the ordinary part. In addition, the scale of the regression function is entirely determined by the constraint that it must tend to 1 as $x^* \rightarrow \infty$ and to 0 as $x^* \rightarrow -\infty$ (for $\theta_2 > 0$), so there is no need to estimate the scale. As a result, probit falls into the class of models where the only weighting function needed is $\omega(\zeta)$. As prescribed in Section 3.2.1, the two elements of $\omega(\zeta)$ are chosen to be

$$\omega_j(\zeta) = (\mathbf{i}\zeta)^{j+2} \exp\left(-\frac{1}{2} \left(\frac{\zeta}{(1.5)\pi/2}\right)^2\right) e^{\mathbf{i}\zeta/2} \quad (22)$$

for $j = 1, 2$. Note that the prefactor $(\mathbf{i}\zeta)^{j+2}$ in Equation (22) is chosen to ensure that $\gamma_o(\zeta, \theta) \omega(\zeta)$ and $\dot{\gamma}_o(\zeta, \theta) \omega(\zeta)$ are well-behaved. Indeed, the ordinary part $\gamma_o(\zeta, \theta)$ behaves as ζ^{-1} as $\zeta \rightarrow 0$ (and thus $\dot{\gamma}_o(\zeta, \theta)$ behaves as ζ^{-2}) and the above choice of $\omega(\zeta)$ guarantees that its product with $\gamma_o(\zeta, \theta)$ or $\dot{\gamma}_o(\zeta, \theta)$ is bounded. Finally, the factor $e^{\mathbf{i}\zeta/2}$ simply introduces a shift in $r_y(z, \theta)$ and $r_{xy}(z, \theta)$ so that their respective modes fall within the regions where $E[Y|Z = z]$ and $E[XY|Z = z]$ vary the most rapidly.

The results shown in Table 3 and the graph of Figure 1c) clearly indicate that the proposed estimator is nearly unbiased, unlike nonlinear instrumental variables (NIV) and nonlinear least squares (NLS). Once again, despite its relatively large standard errors, our estimator still outperforms both NIV and NLS in terms of overall RMSE (see last column). It should also be noted that, for the probit model, the NIV estimator using $\partial g(z, \theta) / \partial \theta$ as instruments exhibits the undesirable tendency to give a $\hat{\theta}_2$ that diverges to infinity about 50% of the time. The results for the NIV estimator given in Table 3 and Figure 1c) are averages over only the replications that did converge to a finite value. The actual performance of NIV is therefore far worse than reported in the Table and in the Figure.

7 Application

7.1 Introduction

The estimation of the wage differential between workers of a different race offers an opportunity to assess the presence of discrimination in the labor market and has received considerable attention the economics literature ((Neal and Johnson 1996), (Bollinger 2003), (Carneiro, Heckman, and Masterov 2003), (Card and Lemieux 1994), and many others). A crucial aspect of this estimation problem is the necessity to control for other factors affecting income in order to separate actual labor market discrimination from premarket factors, such as family socioeconomic background or schooling quality. Following (Neal and Johnson 1996), we use the score on a standardized test taken by virtually all respondents prior to job market entry as an explanatory variable controlling for all premarket factors. Neal and Johnson argue that such an approach offers the advantage that the control variable does not suffer from endogeneity, as it is not affected by the respondent's own decisions, unlike other frequently used controls such as years of schooling, occupation, marital status, or geographical location. Neal and Johnson's findings indicate that the apparent black-white male wage gap of 24%

is reduced to only 7% when premarket skills are taken into account.

Although Neal and Johnson’s argument supports the assumption of the exogeneity of skills, it does not rule out that skills may be measured with error, thus making OLS estimates potentially inconsistent. This issue was investigated by (Bollinger 2003), who provides bounds on the wage gap that account for measurement error. While his widely applicable bounding technique does not depend on the availability of instruments, it is only able to consistently estimate an interval containing the true wage gap, instead of providing a consistent point estimate. As a result, the method gives rather wide bounds on the black-white wage gap (with values ranging from 7% to -126%). Surprisingly, this interval mostly contains negative values of the black-white wage gap, apparently suggesting that discrimination is more likely to be against whites. Also, while Neal and Johnson’s study allows for a nonlinear relationship between skill and wages, Bollinger’s analysis focuses on a linear specification, since bounding techniques are not available for nonlinear specifications.⁹

Our proposed estimation strategy offers the opportunity to combine the strengths of both studies, allowing for the presence of both nonlinearity and measurement error. Moreover, our instrumental variable approach makes it possible to obtain consistent point estimates and consequently improves the accuracy of the estimated black-white wage gap relative to a bounding approach. Note that this investigation is mainly intended to briefly describe a relevant example of an application of our proposed estimator and is necessarily less detailed than a thorough study focusing exclusively on the wage gap issue.

7.2 Data and Methodology

The data we use originates from the “young men” survey group obtained from the National Longitudinal Survey (NLS). Our analysis focuses on a subsample containing all

⁹Unless bounds on the magnitude of the measurement error are available (see, e.g., (Stoker, Berndt, Ellerman, and Schennach 2004)).

individuals for which appropriate measures of income, skills and parental characteristics are available. This subsample consists of 2133 white and 333 black respondents.¹⁰ This dataset is different from the one used in the studies of (Neal and Johnson 1996) and (Bollinger 2003), because we found that the measure of ability used in these studies, the Armed Forces Qualification Test (AFQT), has one main limitation. The AFQT score suffers from a significant censoring bias (see Figure 2), which causes two problems. First, this score is less able to distinguish relative abilities among highly skilled people and, secondly, censoring introduces a measurement error that is negatively correlated with true ability, thus violating our conditional mean assumption regarding the measurement error.¹¹ We rely instead on Intellectual Quotient (IQ), as reported in the NLS study.¹² Although IQ is, technically, also a bounded quantity, the probability density of IQ quickly decays away from its mean, so that the upper and lower bounds on IQ are never reached at the sample sizes available in this study (see Figure 3). As a result, the distribution of IQ is virtually indistinguishable from a continuous distribution supported on \mathbb{R} , so that the assumption of classical measurement error is plausible.

Our analysis also requires a measure of permanent income, which we calculate from the NLS's record of the respondents' yearly wage income over the whole the period of the study. We average log wage income over all years for which the respondent was at

¹⁰The 53 respondents who did not belong to either racial groups were omitted.

¹¹The same caveat applies to the work of (Bollinger 2003). Note that, so far, it is not known if it is possible to correct for this type of nonclassical measurement error in the absence of validation data.

¹²The IQ reported in the NLS study actually comes from a variety of different types of IQ tests taken while the respondent was attending highschool. Even if each type of IQ test had a different systematic bias, this would not invalidate our analysis because the heterogeneity in the IQ tests can simply be considered as a form of measurement error, provided it satisfies the appropriate conditional mean restriction. To investigate this potential problem, we have compared summary statistics for IQ score and income within each subgroup sharing the same IQ test type. We have repeated our analysis after excluding each of the groups exhibiting the largest deviations from the overall sample average and obtained similar results which also support our conclusions. The omitted groups were those for which the IQ scores was calculated from the Scholastic Aptitude Test (SAT) or the grade point average (GPA).

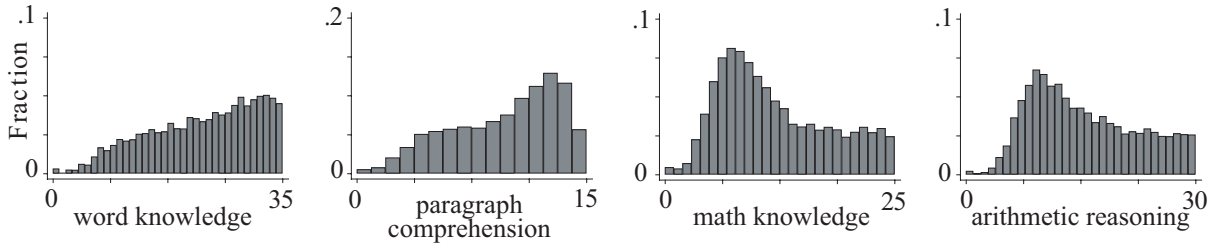


Figure 2: Distribution of scores for each portion of the AFQT test. The thickness of the upper tail of the distributions is indicative that many respondents may have “true” abilities that exceed the maximum possible test score. The absence of a “spike” at the highest score value can be explained by the fact that even very able respondents can make random mistakes.

least 25 years old.¹³ Admittedly, this is far from a perfect measure of log permanent income. However, our setup trivially allows for log permanent income to be error-contaminated since it is the dependent variable. This error is heteroskedastic, since income is available for a different number of years, depending on the respondent, but our setup allows for that possibility as well. Our measure of permanent income may also be biased without invalidating our analysis, as long as the bias does not depend on race or IQ. Although it is common to control for age in wage gap analyses, we favor an approach that avoids an explicit modeling of life cycle effects. We simply consider the effect of age as a random disturbance, since it is reasonable to assume that the age of the respondent at the beginning of the study is independent from race and IQ measured at a fixed age.

Our model is defined by

$$Y_i = \theta_1 + \frac{\theta_2}{2} \operatorname{erf} \left(\frac{X_i^* - \theta_4}{\theta_3} \right) + \Delta Y_i \quad (23)$$

where Y_i is individual’s i log annualized permanent income, X_i^* is his true intellectual quotient and the disturbance ΔY_i is assumed to satisfy the assumptions of Model (2).

A separate model is used for each racial group to allow for a completely general cou-

¹³All amounts were previously deflated using the consumer price index. All incomes reported to be below 3000\$ (in 1984 dollars) were also excluded from the average, since those values are probably the result of temporary loss of work or return to school.

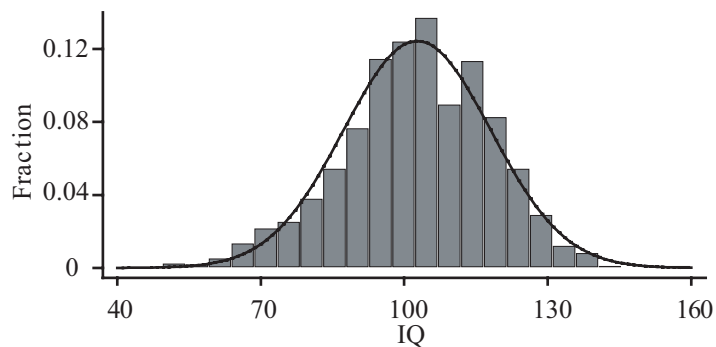


Figure 3: Distribution of IQ.

pling between race and skill response. The parameters of this S-shaped specification have the following meaning. The parameter θ_1 is the overall level of log income, θ_2 the income change from the low end to the high end of the IQ range, θ_3 determines the width of the “S”, while θ_4 indicates the IQ level where the income varies the most rapidly. The choice of the specification is guided by a preliminary analysis based on a nonparametric regression of Y_i on X_i which revealed that the response of income as a function of measured IQ saturates at low and high IQ. Note that the chosen specification reduces to a simple linear specification as $\theta_3 \rightarrow \infty$ and $\theta_2/\theta_3 \rightarrow c \in \mathbb{R}$. Hence, nonlinearity is not imposed in our model — if the response were actually linear, this would be reflected by a very large estimated value of θ_3 .

Since IQ is an error-contaminated measure of true ability, we model observed IQ, X_i , as $X_i = X_i^* + \Delta X_i$ where the measurement error ΔX_i is assumed to fulfill the requirements of Model (2). Our vector of instruments, W_i , is constructed from (i) the respondent’s mother’s highest completed grade¹⁴ and (ii) its square, (iii) the number of siblings the respondent has, (iv) a measure of availability of reading material during the respondent’s childhood and (v) the respondent’s race.¹⁵ This selection of instruments¹⁶ is guided by the predictors of skills identified by (Neal and Johnson 1996).

¹⁴For 73 respondents out of 2466, the mother’s highest completed grade was not available and the father’s highest completed grade was used instead.

¹⁵The instrumental equation $X_i = W_i' \alpha + \Delta X_i^* + \Delta X_i$ is estimated jointly for both racial groups.

¹⁶Although each of these instruments is, strictly speaking, discretely distributed, a linear combi-

A Gaussian kernel¹⁷ was used with a bandwidth (as measured by standard deviation) of 1.8 IQ points for whites and 2.5 IQ points for blacks. Trimming was activated whenever the density of predicted IQ fell below $0.002 (\text{IQ})^{-1}$ for whites and $0.006 (\text{IQ})^{-1}$ for blacks. These settings were determined by gradually varying the bandwidth and trimming parameters in search for the values where the point estimates were the least sensitive to changes in bandwidth and trimming parameters. Our preferred bandwidth and trimming parameters for the two racial groups differ because of their different sample sizes. However, it was verified that our results are robust to setting parameters for the white subsample equal to the ones of the black subsample. The weighting functions used are given in Appendix 7.5. The same weighting functions were used for the two racial groups.

7.3 Diagnostic tests

Before considering the issue of the wage gap, we perform a few diagnostic tests to verify our assumptions, assess the presence of measurement error and verify that the estimator performs as intended.

While most of our assumption take the form of relatively weak conditional mean restrictions, our assumption of independence between the predicted value $Z \equiv E[X|W]$ the prediction error U in the instrument equation of Model (5) warrants verification. Unfortunately, U is not directly observable (as X^* is not observed), and our test of instrument validity will instead rely on the more stringent constraint that $(\Delta X - U)$ is independent from Z . This test can be considered more stringent because, even if it failed, it could be the result of a dependence between ΔX and Z , which would not violate the assumptions of our estimation procedure.¹⁸ This test is also feasible since

nation of them exhibits a distribution whose support consists of such a large number of points that it is virtually indistinguishable from a continuously distributed variable.

¹⁷Bias-reducing kernels were also tried, but were found to require substantial trimming to avoid the “vanishing denominator” problem. A positive second order kernel was found to provide results that were less sensitive to bandwidth and trimming parameter selection.

¹⁸Of course, a dependence between U and ΔX could fortuitously yield a $(\Delta X - U)$ that is inde-

	Fourier (θ_F)				NLS (θ_{LS})			
	θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_4
White	9.97	0.38	2.60	105.1	9.88	0.41	21.18	86.4
	(0.04)	(0.14)	(3.58)	(2.9)	(0.08)	(0.18)	(10.45)	(8.2)
Black	9.74	0.59	4.27	102.8	9.78	0.38	7.12	98.3
	(0.05)	(0.09)	(2.26)	(2.4)	(0.04)	(0.09)	(5.61)	(2.6)

Table 1: Point estimates, with heteroskedasticity-robust standard errors in parenthesis.

$(\Delta X - U)$ can be obtained from the residuals of the regression of X on the instrument vector W . We rely on a Spearman rank correlation test for independence (see, for instance, (van der Vaar 1998), Example 13.22). The test statistic is simply the sample correlation (scaled by \sqrt{n}) between the respective ranks¹⁹ of the two variables of interest (here $(\Delta X - U)^2$ and Z). We use $(\Delta X - U)^2$ instead of simply $(\Delta X - U)$ to improve the power of the test, as the conditional mean restriction on $(\Delta X - U)$ would tend to make the rank correlation between $(\Delta X - U)$ and Z small by construction. (Other powers of $(\Delta X - U)$ and of the instruments yield similar conclusions.) In order to account for the presence of preliminary estimated parameters in the calculations of both $(\Delta X - U)$ and Z , we rely on 1000 bootstrap replications to calibrate the asymptotic variance of this asymptotically normal test statistic. Our conclusion is that, in our application, the null hypothesis of independence is not rejected, as the rank correlation test statistic is 0.63, corresponding to a p -value of 0.53.

We next turn to the issue of testing for the actual presence of measurement error in our data. Table 1 reports the point estimates obtained with our method (labelled by ‘‘Fourier’’) and with conventional Nonlinear Least Squares (NLS) estimates. A consistent test for the presence of measurement error can be constructed by verifying the statistical significance of the difference $(\hat{\theta}_F - \hat{\theta}_{LS})$, where $\hat{\theta}_F$ and $\hat{\theta}_{LS}$ denote the

pendent of Z although U is dependent on Z , but this appears highly unlikely.

¹⁹The rank of a variable is its position in the sample when the sample is sorted according to the value of that variable.

Null Hypothesis	Test Statistic	Degrees of Freedom	<i>P</i> value
No measurement error (white subsample)	15.4	4	0.0039
No measurement error (black subsample)	15.0	4	0.0047
No measurement error (whole sample)	29.9	8	0.0002

Table 2: Testing for the presence of measurement error.

8×1 vectors of all coefficients for the Fourier-based estimator and for nonlinear least squares, respectively. We employ the following test statistic

$$\chi_{\text{rank}(S)}^2 = n \left(\hat{\theta}_F - \hat{\theta}_{LS} \right)' S' \left(S' \hat{E} [(\psi_F - \psi_{LS})(\psi_F - \psi_{LS})'] S \right)^{-1} S \left(\hat{\theta}_F - \hat{\theta}_{LS} \right) \quad (24)$$

where ψ_F and ψ_{LS} denote the influence functions of the corresponding estimator, the \hat{E} operator denotes a sample average operation and S is a rectangular selection matrix extracting the degrees of freedom we wish to test. This type of test statistic reduces to a Hausman test if nonlinear least squares happens to be efficient and has a covariance matrix estimate that is positive definite by construction. As reported in Table 2, our tests clearly reject the null hypothesis of the absence of measurement error for both racial groups.

We now verify that the estimator is effective at capturing the essential features of the data. Figure 4 graphs the returns to IQ implied by specification (23) and our point estimates, both for the Fourier-based and the NLS estimators. Also shown in Figure 4 are the isodensity contours of a nonparametric estimate of the joint density of Y_i and X_i . Our analysis centers on white respondents only, as it is the only subsample that is large enough to obtain a reliable nonparametric bivariate density estimate. The fact that our estimator closely follows the noticeable ridge in the joint density of Y_i and X_i is strongly indicative that the estimator properly identifies the presence of errors in both variables. Its ability to resolve the very sharp marginal returns to IQ in the region of highest density is especially striking. In contrast, the least squares estimator simply tracks the conditional mean of Y_i given X_i and does not detect the sharp increase in income that is clearly noticeable in the nonparametric density plot.

The presence of a region with very sharp marginal returns to IQ has a very plausible explanation. The marginal density of IQ is largest in this region and it follows that a small change in IQ there leads to the largest changes in ranking in the overall population. Assuming that job market outcomes mostly depend on an individual's ranking, a large change in average income would then be expected.

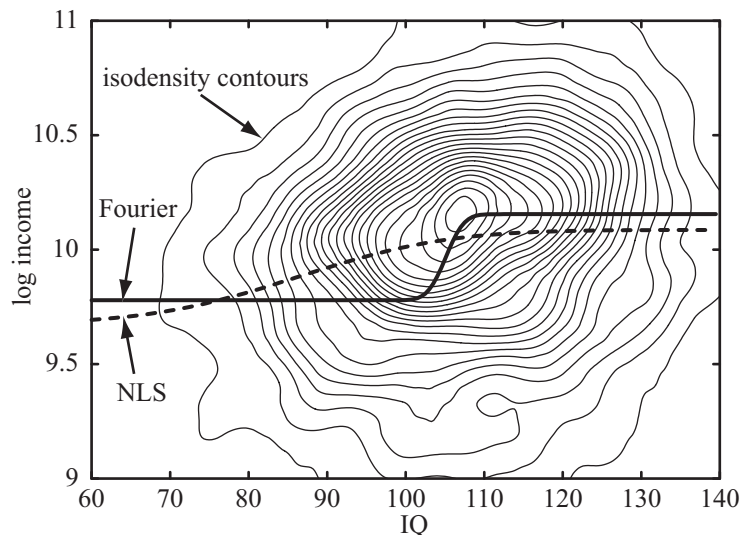


Figure 4: Comparison between IQ response curve obtained with the Fourier-based IV estimator (Fourier) and Nonlinear Least Squares (NLS). Also shown are the isodensity contours of a nonparametric estimate of the joint density of y_i and x_i .

It is interesting to note that, based solely on a conventional least-squares analysis, a linear specification would have appeared to be adequate since the width of the “S” curve obtained with NLS is so large. This application thus provides a clear example where measurement error actually masks the extent of the nonlinearity of the specification and only a nonlinear approach that is robust to measurement error can reliably detect this situation.

The “S” shape of the response also has an unintended advantage in terms of the robustness of our analysis. It has been argued ((Neal and Johnson 1996)) that measures of skills, such as IQ, may be a racially biased measure of skills. However, the only effect of such a bias would be to shift the response horizontally (i.e. bias

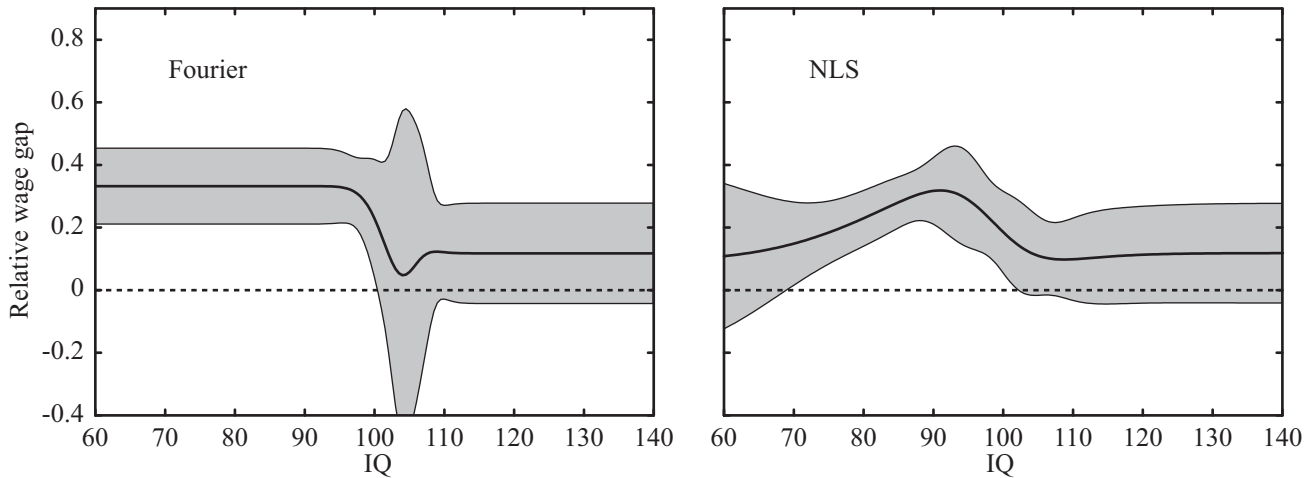


Figure 5: Estimated black-white wage gap as a function of IQ, as estimated with the proposed Fourier based estimator and with NLS. The plotted 95% confidence bands were determined with the delta method from the estimated covariance matrix of the coefficients.

the θ_4 parameter). Hence, as will become evident in the next section, our estimates of the wage gap would be essentially robust to such biases over the relatively wide range of IQ where the income response is flat.

7.4 Results

We now return to the determination of the black-white male wage gap. Since we have allowed the response to IQ to differ between the two racial groups, we are able to determine the wage gap as a function of measurement error-free IQ (see Figure 5), which provides new insight into the issue. The confidence bands in Figure 5 were obtained via the delta method using the estimated covariance matrix of each estimator. The relatively wide confidence bands are mainly attributable to the relatively small size of the black subsample.²⁰

²⁰The “spikes” in the confidence bands around the elbow of the curve are due to the large IQ-dependence of income in this region, which magnifies the noise in the estimated θ_4 coefficient. A similar feature is not clearly visible in the NLS estimates because the estimated IQ-dependence happens to be far weaker for NLS. Fortunately, the very large standard errors in the Fourier-based approach only affects a small portion of the curve and will thus not affect our main findings.

While we have already established that the Fourier-based and NLS results are statistically significantly different, Figure 5 illustrates that the results of the two procedures are also qualitatively very different. Our main findings, based on the measurement error-robust Fourier-based estimates, are two-fold.

1. Below an IQ of about a 100, the wage gap is of the order of 33% and is statistically significant at the 95% level.
2. Above 110 of IQ, the gap shrinks to about 12%, a value which is not statistically significantly different from 0. However, the fact that the wage gap decreases is statistically significant: The χ_1^2 statistic testing that the wage gap is the same below an IQ of 100 and above an IQ of 110 is equal to 5.27, which rejects the null at the 95% level.

It is instructive to compare our findings with the ones of (Neal and Johnson 1996) and (Bollinger 2003). First, it should be noted that differences between our results and these earlier studies can at least in part be traced back to differences in the data used. Repeating Neal and Johnson's main least-squares analysis (using a quadratic dependence on skills and a dummy for race) with our sample yields a wage gap of 33% without controlling for skills, which is reduced to 21% after controlling for skills via IQ (Neal and Johnson found 24% and 7%, respectively). Hence, it should not be surprising that our results more strongly indicate the presence of discrimination. One possible source of the difference is that Neal and Johnson use hourly wages while we use yearly income.²¹ If there is discrimination in the hiring process, black respondents may remain unemployed for longer periods, an effect which would be visible in the reported yearly income but not in the reported hourly wage. Of course, sample selection bias issues may still play a role in our study ((Chandra 2003)).

Our results confirm Neal and Johnson's observation that, when the skill-dependence of income is allowed to differ across racial groups, the gap appears to narrow at the

²¹Our use of yearly income was guided by the fact that hourly wages (calculated or reported) were missing for a large fraction of the respondents in our sample.

higher end of the skill distribution. This trend was not statistically significant in their study, but is clear in Figure 5. The well-known measurement error-induced attenuation phenomenon is a possible source of the lack of significance Neal and Johnson observed, although differences in the data used could also be a factor. Figure 5 also shows that it would be inappropriate to use Bollinger's bounds on the wage gap to conclude that the wage gap is inexistent or negative. The relatively narrow width of our confidence bands enabled by the use of instruments permits us to more precisely pin down the magnitude of the wage gap and show that it is still statistically significant at least over a portion of the skill distribution when measurement error is accounted for. Perhaps our most striking finding is the sharpness of the drop in the wage gap as a function of IQ, a feature which simply cannot be detected in our dataset without properly accounting for both measurement error and nonlinearity.

The results of our analysis are consistent with a number of interpretations. For instance, applicants for jobs requiring low skill levels are typically recruited locally, while more skill-intensive positions are often advertised over a larger geographical area, through newspapers, specialized magazines or recruiting services. Hence, the low-skill wage gap mainly reflects a gap in the prevailing wages in different, segregated, neighborhoods. The gap is smaller among highly skilled individuals, who do not necessarily work in their native neighborhood. An alternative, but related, explanation is that, beyond a certain level of ability, undertaking a college education becomes more likely, which often brings young black men out of their native neighborhood and into other communities where the prevailing wages may be higher. Finally, it is possible that, beyond discrimination in wages, there exists discrimination in the hiring/firing process, which would cause black workers to be employed for a smaller fraction of the year on average than equally qualified white workers, thus resulting in a black-white gap in yearly income. If the turnover rate is higher in occupations demanding lower skills, this would result in a larger income gap between racial groups

Function	Expression
$\omega_j(\zeta)$	$\omega_j(\zeta) = (\mathbf{i}\zeta)^3 \exp(- (1.402) \zeta^2) e^{-10\mathbf{i}\zeta} e^{-\mathbf{i}\zeta(2.75)(j-1)}$ for $j = 1, 2$
$\varpi(\zeta)$	$\lambda(\zeta) = C \frac{\mathbf{i}\sin(27.5\zeta)}{\mathbf{i}\zeta} \exp(- (20.264) \zeta^2) e^{100\mathbf{i}\zeta}$ where $C : \int \lambda(\zeta) d\zeta = 1$
$\nu_{y,0}(\zeta, \theta)$	$\lambda(\zeta) = \frac{\exp(- (0.72)\zeta^2)}{\mathbf{i}\zeta} e^{100\mathbf{i}\zeta}$

Table 3: Weighting functions used for the Fourier-based estimator. The functions $\omega_j(\zeta)$, $\varpi(\zeta)$ and $\nu_{y,0}(\zeta, \theta)$ refer to the functions used in Section 3.2 to construct the moment conditions, using the function $\lambda \in \mathcal{G}$ used as a starting point.

for lower-skilled workers.

7.5 Weighting functions used in the application

We first observe that the Fourier transform of the ‘‘S’’-shaped specification given in Equation (23) is

$$\gamma(\zeta, \theta) = \theta_1 2\pi \delta(\zeta) + \theta_2 (-\mathbf{i}\zeta)^{-1} \exp\left(\mathbf{i}\zeta\theta_4 - \frac{\zeta^2}{4\theta_3^2}\right). \quad (25)$$

Consequently, a weighting function of the form $\nu_{y,0}(\zeta, \theta)$ is needed to extract the magnitude of the singularity θ_1 , a weighting function of the form $\varpi(\zeta)$ is required to determine θ_2 and a two-dimensional vector of weighting functions of the form $\omega(\zeta)$ is needed to obtain θ_3 and θ_4 . (Refer to Section 3.2 for a description of these different types of weighting functions.) The functional forms of these weighting functions, given in Table 3, were derived as follows.

The starting point of the construction of $\omega_j(\zeta)$ (in Section 3.2.1) is a Gaussian function (of z) with a center of mass and a width such that the Gaussian takes a negligible value outside of the range of values of Z actually observed in the sample. After a Fourier transform operation, this yields another Gaussian function (of ζ) multiplied by a phase factor $e^{\mathbf{i}\zeta c}$, where c depends on the center of mass of the original Gaussian. Each element of the vector $\omega(\zeta)$ is obtained from a Gaussian with a slightly different center of mass. Next, this expression is multiplied by a positive power of $(\mathbf{i}\zeta)$ that is (i) sufficiently large to cancel the $(-\mathbf{i}\zeta)^{-1}$ divergence in Equation (25) or

the $(-i\zeta)^{-2}$ divergence in its derivative $\dot{\gamma}(\zeta, \theta)$ and (ii) such that the inner products of $\omega_j(\zeta) \dot{\gamma}(\zeta, \theta^*)$ with $\delta(\zeta)$ and $\omega_j(\zeta) \gamma(\zeta, \theta^*)$ with $\delta^{(1)}(\zeta)$ vanish, thus achieving orthogonality to the singular part.

As described in Section 3.2, the weighting function $\varpi \in \mathcal{S}_1 \cap \mathcal{C}$ is derived from some function $\lambda \in \mathcal{G}'$ (which is the extensions of \mathcal{G} provided in Section 5). The functional form of $\lambda(\zeta)$ (given in Table 3) is obtained by first noting that the weighting function used to identify the height of the “S” function (the θ_2 parameter) should essentially sample the difference between the value of $E[Y|Z=z]$ for values of z before and after the “jump”. Hence, a natural starting point is the difference between two Gaussian functions (of z) centered somewhere before and after the jump. After a Fourier transform operation, we again obtain a Gaussian but the phase factor now includes a multiplicative factor of the form $(e^{i\zeta c} - e^{-i\zeta c})/2 = i \sin(\zeta c)$ due to the presence of two shifted Gaussian with opposite signs. Since our procedure (in Section 3.2.1) requires the resulting function (of ζ) to be divided by $\gamma_o(\zeta, \theta)$, we insert a multiplicative factor of the form $(i\zeta)^{-1}$ designed to cancel a similar divergence in the expression of $\gamma_o(\zeta, \theta)$. No additional step is required to achieve orthogonality to the singular part, since the behavior of the resulting function at the origin already guarantees a vanishing inner product with a delta function. However, we need to introduce a multiplicative constant C determined numerically to ensure that our function is properly normalized to integrate to 1, as required by the constraint that $\varpi \in \mathcal{S}_1 \cap \mathcal{C}$.

As described in Section 3.2, the weighting function $\nu_{y,0}(\zeta, \theta)$ is derived from some function $\mu_{y,0} \in \mathcal{S}_0 \cap \mathcal{C}$, which, in turn is derived from some $\lambda \in \mathcal{G}'$. The expression for $\lambda(\zeta)$ is again based on the Fourier transform of a shifted Gaussian. For the same reason as in the case of the $\varpi(\zeta)$ function, we introduce a $(i\zeta)^{-1}$ factor to cancel a similar divergence in the expression of $\gamma_o(\zeta, \theta)$ when constructing $\mu_{y,0}(\zeta)$. The resulting expression already integrates to 0 (in the Cauchy principal value sense) and

directly satisfies the constraint that $\mu_{y,0} \in \mathcal{S}_0 \cap \mathcal{C}$, which ensures orthogonality to the ordinary part.

These steps provide us with a family of weighting functions with up to two adjustable parameters, typically one for the width of the Gaussian and one for its location (on the z axis). These numerical coefficients were selected by using the estimated asymptotic variance as an informal guide. The point estimates are not very sensitive to the exact values of these coefficients, as long as they are such that the general region where the functions $r_y(z, \theta)$, $r_{xy}(z, \theta)$ and $r_{1y}(z, \theta)$ are the largest in magnitude corresponds to the range of values of Z found in the actual sample.

8 Computational aspects

The implementation of the estimator is considerably simplified by the fact that all the relatively abstract operations requiring Fourier transforms involve nonrandom quantities. The end result of these operations is a vector of nonlinear functions whose expectations are to be evaluated from the observed data.

The first step in the implementation of the estimator is the calculation of the Fourier transform $\gamma(\zeta, \theta)$ of $g(x^*, \theta)$. Symbolic mathematical packages such as Maple and Mathematica are often able to carry out such transforms automatically, even when the answers involve delta function derivatives. When an analytic expression for $\gamma(\zeta, \theta)$ is not available, the following hybrid analytical and numerical approach can be used. The idea is to write $g(x^*, \theta)$ as

$$g(x^*, \theta) = (g(x^*, \theta) - T(x^*, \theta)) + T(x^*, \theta)$$

where $T(x^*, \theta)$ represents the asymptotic behavior of $g(x^*, \theta)$ for large $|x^*|$ and where $(g(x^*, \theta) - T(x^*, \theta))$ is absolutely integrable (with respect to x^*). If the tail $T(x^*, \theta)$ follows a simple behavior such as a linear combination of functions of the form $(x^*)^{k_1} (\ln(x^*))^{k_2}$, then its Fourier transforms $\Theta(\zeta, \theta)$ can be found in standard Fourier

transforms Tables (such as Table I in (Lighthill 1962)). Typically, $\Theta(\zeta, \theta)$ will contain both a sum of delta function derivatives, which will provide the values of $\gamma_j(\theta)$ in Equations (32) and (33), as well as an ordinary function part $\Theta_o(\zeta, \theta)$. The Fourier transform of the remaining absolutely integrable contribution $(g(x^*, \theta) - T(x^*, \theta))$ can then be obtained numerically via

$$\gamma(\zeta, \theta) - \Theta(\zeta, \theta) = \lim_{\substack{t^* \rightarrow \infty \\ b \rightarrow 0}} \sum_{t=-t^*}^{t^*} (g(tb, \theta) - T(tb, \theta)) e^{i\zeta tb}.$$

All the ordinary function contributions, $\gamma_o(\zeta, \theta) = \Theta_o(\zeta, \theta) + \gamma(\zeta, \theta) - \Theta(\zeta, \theta)$, are then added and their value over a grid $\mathbb{G} = \{\zeta \in \mathbb{R} : \zeta = tb, t = -t^*, \dots, 0, \dots, t^*\}$ is stored, while making sure that the grid is sufficiently fine ($b \rightarrow 0$) and extended ($t^* \rightarrow \infty$) to provide an accurate numerical approximation to $\gamma_o(\zeta, \theta)$.

References

- ANDREWS, D. W. K. (1995): “Nonparametric Kernel Estimation for Semiparametric Models,” *Econometric Theory*, 11, 560–596.
- BOLLINGER, C. R. (2003): “Measurement Error In Human Capital And The Black-White Wage Gap,” *Review of Economics and Statistics*, 85, 578–585.
- CARD, D., AND T. LEMIEUX (1994): “Changing Wage Structure and Black-White Wage Differentials among Men and Women: A Longitudinal Analysis,” Working Paper 4755, National Bureau of Economic Research.
- CARNEIRO, P., J. J. HECKMAN, AND D. V. MASTEROV (2003): “Labor Market Discrimination and Racial Differences in Premarket Factors,” Working Paper 10068, National Bureau of Economic Research.
- CHANDRA, A. (2003): “Is the Convergence of the Racial Wage Gap Illusory,” Working Paper 9476, National Bureau of Economic Research.
- GEL’FAND, I. M., AND G. E. SHILOV (1964): *Generalized Functions*. Academic Press, New York.
- HAUSMAN, J., W. NEWEY, H. ICHIMURA, AND J. POWELL (1991): “Measurement Errors in Polynomial Regression Models,” *Journal of Econometrics*, 50, 273–295.
- LIGHTHILL, M. J. (1962): *Introduction to Fourier Analysis and Generalized Function*. London: Cambridge University Press.
- NEAL, D. A., AND W. R. JOHNSON (1996): “The Role of Premarket Factors in Black-White Wage Differences,” *Journal of Political Economy*, 104, 869–895.

- NEWKEY, W., AND D. MCFADDEN (1994): “Large Sample Estimation and Hypothesis Testing,” in *Handbook of Econometrics*, ed. by R. F. Engel, and D. L. McFadden, vol. IV. Elsevier Science.
- NEWKEY, W., AND R. J. SMITH (2004): “Higher-Order Properties of GMM and Generalized Empirical Likelihood Estimators,” *Econometrica*, 72, 219–255.
- PAGAN, A., AND A. ULLAH (1999): *Nonparametric Econometrics*. Cambridge University Press, Cambridge, UK.
- PHILLIPS, P. C. B. (1991): “A Shortcut to LAD Estimator Asymptotics,” *Econometric Theory*, 7, 450–463.
- POWELL, J., J. STOCK, AND T. STOKER (1989): “Semiparametric Estimation of Index Coefficients,” *Econometrica*, 57, 1403–1430.
- SCHWARTZ, L. (1966): *Théorie des distributions*. Paris, Hermann.
- STOKER, T., E. BERNDT, D. ELLERMAN, AND S. M. SCHENNACH (2004): “Panel Data Analysis of U.S. Coal Productivity,” *Journal of Econometrics*, forthcoming.
- TEMPLE, G. (1963): “The Theory of Weak Functions. I,” *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 276, 149–167.
- VAN DER VAAR, A. W. (1998): *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK.
- ZINDE-WALSH, V., AND P. C. B. PHILLIPS (2003): “Fractional Brownian Motion as a Differentiable Generalized Gaussian Process,” Working Paper 1391, Cowles Foundation, Yale University.