Instrumental Variable Treatment of Nonclassical Measurement Error Models

Yingyao Hu
Department of Economics
The University of Texas at Austin
1 University Station C3100
BRB 1.116
Austin, TX 78712
hu@eco.utexas.edu

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

First version: December 2004; This version: March 2007.

Abstract

While the literature on nonclassical measurement error traditionally relies on the availability of an auxiliary dataset containing correctly measured observations, we establish that the availability of instruments enables the identification of a large class of nonclassical nonlinear errors-in-variables models with continuously distributed variables. Our main identifying assumption is that, conditional on the value of the true regressors, some "measure of location" of the distribution of the measurement error (e.g. its mean, mode or median) is equal to zero. The proposed approach relies on the eigenvalue-eigenfunction decomposition of an integral operator associated with specific joint probability densities. The main identifying assumption is used to "index" the eigenfunctions so that the decomposition is unique. We propose a convenient sieve-based estimator, derive its asymptotic properties and investigate its finite-sample behavior through Monte Carlo simulations.

Keywords: Nonclassical measurement error, nonlinear errors-in-variables model, instrumental variable, operator, semiparametric estimator, sieve maximum likelihood.

*Corresponding author. S. M. Schennach acknowledges support from the National Science Foundation via grant SES-0452089. The authors would like to thank Lars Hansen, James Heckman, Marine Carrasco, Maxwell Stinchcombe and Xiaohong Chen, as well as seminar audiences at various universities, at the Cemmap/ESRC Econometric Study Group Workshop on Semiparametric Methods and at the Econometric Society 2006 Winter meetings for helpful comments.

1 Introduction

In recent years, there has been considerable progress in the development of inference methods that account for the presence of measurement error in the explanatory variables in nonlinear models (see, for instance, Chesher (1991), Lewbel (1996), Chesher (1998), Lewbel (1998), Hausman (2001), Chesher (2001), Chesher, Dumangane, and Smith (2002), Hong and Tamer (2003), Carrasco and Florens (2005)). The case of classical measurement errors, in which the measurement error is either independent from the true value of the mismeasured variable or has zero mean conditional on it, has been thoroughly studied. In this context, approaches that establish identifiability of the model, and provide estimators that are either consistent or root n consistent and asymptotically normal have been devised when either instruments (Hausman, Newey, Ichimura, and Powell (1991), Newey (2001), Schennach (2007)), repeated measurements (Hausman, Newey, Ichimura, and Powell (1991), Li (2002), Schennach (2004a), Schennach (2004b)) or validation data (Hu and Ridder (2004)) are available.

However, the are a number of practical applications where the assumption of classical measurement error is not appropriate (Bound, Brown, and Mathiowetz (2001)). In the case of discretely distributed regressors, instrumental variable estimators that are robust to the presence of such "nonclassical" measurement error have been developed for binary regressors (Mahajan (2006), Lewbel (2007)) and general discrete regressors (Hu (2007)). Unfortunately, these results cannot trivially be extended to continuously distributed variables, because the number of nuisance parameters needed to describe the measurement error distribution (conditional on given values of the observable variables) becomes infinite. Identifying these parameters thus involves solving operator equations that exhibit potential ill-defined inverse problems (similar to those discussed in Carrasco, Florens, and Renault (2005), Darolles, Florens, and Renault (2002), and Newey and Powell (2003)).

In the case of continuously distributed variables (in both linear or nonlinear models), the only approach capable of handling nonclassical measurement errors proposed so far has been the use of an auxiliary dataset containing correctly measured observations (Chen, Hong, and Tamer (2005), Chen, Hong, and Tarozzi (2005)). Unfortunately, the availability of such a

clean data set is the exception rather than the rule. Our interest in instrumental variables is driven by the fact that instruments suitable for the proposed approach are conceptually similar to the ones used in conventional instrumental variable methods and researchers will have little difficulty identifying appropriate instrumental variables in typical datasets.

Our approach relies on the observation that, even though the measurement error may not have zero mean conditional on the true value of the regressor, perhaps some other measure of location, such as the median or the mode, could still be zero. This type of nonclassical measurement error has been observed, for instance, in the self-reported income found in the Current Population Survey (CPS).[1] Thanks to the availability of validation data for one of the years of the survey, it was found that, although measurement error is correlated with true income, the median of misreported income conditional on true income is in fact equal to the true income (Bollinger (1998)). In another study on the same dataset, it was found that the mode of misreported income conditional on true income is also equal to the true income (see Bound and Krueger (1991) and Figure 1 in Chen, Hong, and Tarozzi (2005)).

There are numerous plausible settings where the conditional mode, median, or some other quantile, of the error could be zero even though its conditional mean is not. First, if respondents are more likely to report values close to the truth than any particular value far from the truth, then the mode of the measurement error would be zero. This is a very plausible form of measurement error that even allows for systematic over- or underreporting. Intuitively, since there is only one way to report the truth, while there are an infinite number of alternative ways to misreport, respondents would literally have to collude on misreporting in a similar way in order to violate the mode assumption. In addition, data truncation usually preserves the mode, but not the mean, provided the truncation is not so severe that the mode itself is deleted.

Second, if respondents are equally likely to over- or under-report, but not by the same amounts on average, then the median of the measurement error is zero. This could occur perhaps because the observed regressor is a nonlinear monotonic function (e.g., a logarithm)

[1] Bureau of Labor Statistics and Bureau of Census, http://www.bls.census.gov/cps/cpsmain.htm

of some underlying mismeasured variable with symmetric errors. Such a nonlinear function would preserve the zero median, but not the zero mean of the error. Another important case is data censoring, which also preserves the median, as long as the upper censoring point is above the median and the lower censoring point is below the median.

Third, in some cases, a quantile other than the median might be appropriate. For instance, tobacco consumption is likely to be either truthfully reported or under-reported and, in that case, the topmost quantile of the error conditional on the truth would plausibly equal true consumption.

In order to encompass practically relevant cases such as these, which so far could only have been analyzed in the presence of auxiliary correctly measured data, our approach relies on the general assumption that some given "measure of location" (e.g. the mean, the mode, the median, or some other quantile) characterizing the distribution of the observed regressor conditional on the true regressor is left unaffected by the presence of measurement error. This framework is also sufficiently general to include measurement error models in which the true regressor and the errors enter the model in a nonseparable fashion.

The paper is organized as follows. We first provide a general proof of identification before introducing a semiparametric sieve estimator that is shown to be root n consistent and asymptotically normal. Our identification is fully nonparametric and therefore establishes identification in the presence of measurement error of any model that would be identified in the absence of measurement error. Our estimation framework encompasses models which, when expressed in terms of the measurement error-free variables, take the form of either parametric likelihoods or (conditional or unconditional) moment restrictions and automatically provides a corresponding measurement error-robust semiparametric instrumental variable estimator. This framework therefore addresses nonclassical measurement error issues in most of the widely used models, including probit, logit, tobit and duration models, in addition to conditional mean and quantile regressions, as well as nonseparable models (thanks to their relationship with quantile restrictions). The finite sample properties of the estimator are investigated via Monte Carlo simulations.

2 Identification

The "true" model is defined by the joint distribution of the dependent variable y and the true regressor x^*. However, x^* is not observed, only its error-contaminated counterpart, x, is observed. In this section, we rely on the availability of an instrument (or a repeated measurement) z to show that the joint distribution of x^* and y is identified from the knowledge of the distribution of all observed variables. Our treatment can be straightforwardly extended to allow for the presence of a vector w of additional correctly measured regressors, merely by conditioning all densities on w.

Let $\mathcal{Y}$, $\mathcal{X}$, $\mathcal{X}^*$ and $\mathcal{Z}$ denote the supports of the distributions of the random variables y, x, x^* and z, respectively. We consider x, x^* and z to be continuously distributed ($\mathcal{X}$, $\mathcal{X}^* \subset \mathbb{R}^{n_x}$ and $\mathcal{Z} \subset \mathbb{R}^{n_z}$ with $n_z \geq n_x$) while y can be either continuous or discrete. Accordingly, we assume the following.

Assumption 1 *The joint density of y and x^*, x, z admits a bounded density with respect to the product measure of some dominating measure μ (defined on $\mathcal{Y}$) and the Lebesgue measure on $\mathcal{X}^* \times \mathcal{X} \times \mathcal{Z}$. All marginal and conditional densities are also bounded.*

We use the notation $f_a(a)$ and $f_{a|b}(a|b)$ to denote the density of variable a and the density of a conditional on b, respectively. Implicitly, these densities are relative to the relevant dominating measure, as described above. For simplicity, our notation does not distinguish between a random variable and a specific value it may take.

To state our identification result, we start by making natural assumptions regarding the conditional densities of all the variables of the model.

Assumption 2 *(i) $f_{y|xx^*z}(y|x, x^*, z) = f_{y|x^*}(y|x^*)$ for all $(y, x, x^*, z) \in \mathcal{Y} \times \mathcal{X} \times \mathcal{X}^* \times \mathcal{Z}$ and (ii) $f_{x|x^*z}(x|x^*, z) = f_{x|x^*}(x|x^*)$ for all $(x, x^*, z) \in \mathcal{X} \times \mathcal{X}^* \times \mathcal{Z}$.*

Assumption 2(i) indicates that x and z do not provide any more information about y than x^* already provides, while Assumption 2(ii) specifies that z does not provide any more information about x than x^* already provides. These assumptions can be interpreted

as standard exclusion restrictions. Conditional independence restrictions are widely used in the recent econometrics literature (e.g. Holderlein and Mammen (2006), Heckman and Vytlacil (2005), Altonji and Matzkin (2005)).

Remark: Our assumptions regarding the instrument z are sufficiently general to encompass both the repeated measurement and the instrumental variable cases in a single framework. In the repeated measurement case, having the measurement error on the two measurements z and x be mutually independent conditional on x^* will be sufficient to satisfy Assumption 2. Note that while we will refer to y as the "dependent variable", it should be clear that it could also contain another error-contaminated measurement of x^* or even a type of instrument that is "caused by" x^*, as suggested in Chalak and White (2006). Finally, note that our assumptions allow for the measurement error $(x - x^*)$ to be correlated with x^*, which is crucial in the presence of potentially nonclassical measurement error.

To facilitate the statement of our next assumption, is it useful to note that a function of two variables can be associated with an integral operator.

Definition 1 *For two random variables a and b with support $\mathcal{A}$ and $\mathcal{B}$, respectively, let $L_{b|a}$ denote an operator mapping a function g in some function space $\mathcal{G}(\mathcal{A})$ onto the function $L_{b|a}g$ in some function space $\mathcal{G}(\mathcal{B})$ and such that the value of the function $L_{b|a}g$ at the point $b \in \mathcal{B}$ is given by*

$$\left[L_{b|a}g\right](b) \equiv \int_{\mathcal{A}} f_{b|a}(b|a)\, g(a)\, da,$$

where $f_{b|a}(b|a)$ denotes the conditional density of b given a.

In order for the density $f_{b|a}(b|a)$ to be uniquely determined by the operator $L_{b|a}$, the space $\mathcal{G}(\mathcal{A})$ upon which the operator acts must be sufficiently large so that $f_{b|a}(b|a)$ is "sampled" everywhere. For an integral operator, it is sufficient to consider $\mathcal{G}(\mathcal{A})$ to be $\mathcal{L}^1(\mathcal{A})$, the set of all absolutely integrable functions supported on $\mathcal{A}$ (endowed with the norm $\|g\|_1 = \int_{\mathcal{A}} |g(a)|\, da$). It is even sufficient to limit $\mathcal{G}(\mathcal{A})$ to the set of functions in $\mathcal{L}^1(\mathcal{A})$ that are also bounded ($\sup_{a\in\mathcal{A}} |g(a)| < \infty$), denoted $\mathcal{L}^1_{\text{bnd}}(\mathcal{A})$.[2] In our subsequent

[2] This can be seen from the fact that

$$f_{b|a}(b|a_0) = \lim_{n\to\infty} \left[L_{b|a}g_{n,a_0}\right](b) \tag{1}$$

treatment, we will consider the cases where $\mathcal{G} = \mathcal{L}^1$ or where $\mathcal{G} = \mathcal{L}^1_{\text{bnd}}$. We can then state our next assumption.

Assumption 3 *The operators $L_{x|x^*}$ and $L_{z|x}$ are injective (for either $\mathcal{G} = \mathcal{L}^1$ or $\mathcal{G} = \mathcal{L}^1_{bnd}$).*

An operator $L_{b|a}$ is said to be *injective* if its inverse $L^{-1}_{b|a}$ is defined over the range of the operator $L_{b|a}$ (see Section 3.1 in Carrasco, Florens, and Renault (2005)). The qualification on the range is needed to account for the fact that inverses are often defined only over a restricted domain in infinite-dimensional spaces. Assumption 3 could also be stated in terms of the injectivity of $L_{z|x^*}$ and $L_{x|x^*}$, since it can be shown that injectivity of $L_{z|x^*}$ and $L_{x|x^*}$ implies injectivity of $L_{z|x}$.

Intuitively, an operator $L_{b|a}$ will be injective if there is enough variation in the density of b for different values of a. For instance, a simple case where $L_{b|a}$ is not injective is when $f_{b|a}(b|a)$ is a uniform density on $\mathcal{B}$ for any $a \in \mathcal{A}$. In general, however, injectivity assumptions are quite weak and are commonly made in the literature on nonparametric instrumental variable methods. They are sometimes invoked by assuming that an operator $L_{b|a}$ admits a singular value decomposition with nonzero singular values (Darolles, Florens, and Renault (2002)) or by stating that an operator is "nonsingular" (Horowitz (2006), Hall and Horowitz (2005)).

Injectivity assumptions are often phrased in terms of *completeness* (or *bounded completeness*) of the family of distributions playing the role of the kernel of the integral operator considered (Newey and Powell (2003), Blundell, Chen, and Kristensen (2003), Chernozhukov and Hansen (2005), Chernozhukov, Imbens, and Newey (2006)). This characterization is worth explaining in more detail, as it leads to primitive sufficient conditions. Formally, a

where $g_{n,a_0}(a) = n1\left(|a - a_0| \leq n^{-1}\right)$, a sequence of absolutely integrable and bounded functions (the limit of that sequence does not need to belong to $\mathcal{G}(\mathcal{A})$, since we are *not* calculating $L_{b|a} \lim_{n\to\infty} g_{n,a_0}$). The so-called kernel $f_{b|a}(b|a_0)$ of the integral operator $L_{b|a}$ is therefore uniquely determined by evaluating the limit (1) for all values of $a_0 \in \mathcal{A}$. It is also straightforward to check that, for a bounded $f_{b|a}(b|a)$, $g \in \mathcal{L}^1(\mathcal{A})$ implies $L_{b|a}g \in \mathcal{L}^1(\mathcal{B})$ and that $g \in \mathcal{L}^1_{\text{bnd}}(\mathcal{A})$ implies $L_{b|a}g \in \mathcal{L}^1_{\text{bnd}}(\mathcal{B})$. Indeed, $\left\|L_{b|a}g\right\|_1 \leq \int\int f_{b|a}(b|a)\,db\,|g(a)|\,da = \int 1\,|g(a)|\,da = \|g\|_1$ and $\sup_{b\in\mathcal{B}}\left|\left[L_{b|a}g\right](b)\right| \leq \int\left(\sup_{b\in\mathcal{B}}\sup_{a\in\mathcal{A}}\left|f_{b|a}(b|a)\right|\right)|g(a)|\,da = \left(\sup_{b\in\mathcal{B}}\sup_{a\in\mathcal{A}}\left|f_{b|a}(b|a)\right|\right)\|g\|_1$.

family of distribution $f_{a|b}(a|b)$ is complete if the only solution $\tilde{g}(a)$ to

$$\int_{\mathcal{A}} \tilde{g}(a) f_{a|b}(a|b)\, da = 0 \text{ for all } b \in \mathcal{B} \tag{2}$$

(among all $\tilde{g}(a)$ such that (2) is defined) is $\tilde{g}(a) = 0$. Under Assumption 1, this condition implies injectivity of $L_{b|a}$ (viewed as a mapping from $\mathcal{L}^1(\mathcal{A})$ to $\mathcal{L}^1(\mathcal{B})$). Indeed, $\int f_{a|b}(a|b)\tilde{g}(a)\, da = (f_b(b))^{-1} \int f_{b|a}(b|a) f_a(a) \tilde{g}(a)\, da$ and since $0 < f_a(a) < \infty$ and $0 < f_b(b) < \infty$ over the interior of their respective supports, having $\tilde{g}(a) = 0$ as the unique solution is equivalent to having $g(a) = 0$ as the unique solution to $\int f_{b|a}(b|a) g(a)\, da = 0$. If $g(a) = 0$ is the unique solution among all $g(a)$ such that the integral is defined, then it is also the unique solution in $\mathcal{L}^1(\mathcal{A})$, which implies that $L_{b|a}$ is injective. Bounded completeness is similarly defined by stating that the only solution to (2) among all *bounded* $\tilde{g}(a)$ is $\tilde{g}(a) = 0$. Analogously, this implies that $L_{b|a}$ is injective, when viewed as a mapping from $\mathcal{L}^1_{\text{bnd}}(\mathcal{A})$ to $\mathcal{L}^1_{\text{bnd}}(\mathcal{B})$.

A nice consequence of the connection between injectivity and (bounded) completeness is that primitive conditions for (bounded) completeness are readily available in the literature. For instance, some very general exponential families of distributions are known to be complete (as invoked in Newey and Powell (2003)). The weaker notion of bounded completeness can also be used to find even more general families of distributions leading to injective operators (as discussed in Blundell, Chen, and Kristensen (2003)). In particular, when $f_{a|b}(a|b)$ can be written in the form $f_\varepsilon(a-b)$, then $L_{b|a}$ is injective if and only if the Fourier transform of f_ε is everywhere nonvanishing (by Theorem 2.1 in Mattner (1993)) and similar results have also been obtained for more general families of distributions that cannot be written as $f_\varepsilon(a-b)$ (d'Haultfoeuille (2006)).

The assumption of injectivity of $L_{x|x^*}$ allows for x^* and x to be multivariate. Injectivity of $L_{z|x}$ in multivariate settings is also natural whenever the dimension of z is greater or equal to the dimension of x. If the dimension of z is less than the dimension of x or if z contains too many colinear elements, identification will not be possible, as expected.

While Assumption 3 places restrictions on the relationships between z, x and x^*, the following assumption places restrictions on the relationship between y and x^*.

Assumption 4 *For all* $x_1^*, x_2^* \in \mathcal{X}^*$*, the set* $\left\{y : f_{y|x^*}\left(y|x_1^*\right) \neq f_{y|x^*}\left(y|x_2^*\right)\right\}$ *has positive probability whenever* $x_1^* \neq x_2^*$.

This assumption is even weaker than injectivity. It is automatically satisfied if $E\left[y|x^*\right]$ is strictly monotone (for univariate x^*), but also holds far more generally. The presence of conditional heteroskedasticity can be sufficient in the absence of monotonicity. Assumption 4 is only violated if the distribution of y conditional on x^* is identical at two values of x^*.

Remark: In the special case of binary y, Assumption 4 amounts to a monotonicity assumption (e.g. $P\left[y = 0|x^*\right]$ is strictly monotone in x^*). When x^* is multivariate, while the outcome variable is still binary (or when $P\left[y = 0|x^*\right]$ is not monotone), it will be necessary to define y to be a vector containing auxiliary variables in addition to the binary outcome, in order to allow for enough variation in the distribution of y conditional on x^* to satisfy Assumption 4. Each of these additional variables need not be part of the model of interest per se, but does need to be affected by x^* is some way. In that sense, such a variable is a type of "instrument", although it differs conceptually from conventional instruments, as it would typically be "caused by x^*" instead of "causing x^*". See Chalak and White (2006) for a discussion of this type of instrument.

We then characterize the nature of measurement error via an assumption that considerably generalizes the case of classical measurement error.

Assumption 5 *There exists a known functional* M *such that* $M\left[f_{x|x^*}\left(\cdot|x^*\right)\right] = x^*$ *for all* $x^* \in \mathcal{X}^*$.

M is a very general functional that maps a density to a real number (or a vector, if x^* is multivariate) and that defines some measure of location. Examples of M include, but are not limited to, the mean, the mode, or the τ quantile, corresponding to the following definitions of M, respectively,

$$M\left[f\right] = \int_{\mathcal{X}} x f(x) dx \tag{3}$$

$$M\left[f\right] = \arg\max_{x \in \mathcal{X}} f(x) \tag{4}$$

$$M\left[f\right] = \inf\left\{x^* \in \mathcal{X}^* : \int 1\left(x \leq x^*\right) f(x) dx \geq \tau\right\}. \tag{5}$$

Case (3) above covers classical measurement error (in which $x = x^* + \varepsilon$, where $E[\varepsilon|x^*] = 0$), since $M\left[f_{x|x^*}(\cdot|x^*)\right] = E[x|x^*] = E[x^* + \varepsilon|x^*] = x^* + E[\varepsilon|x^*] = x^*$ in that case. The other two examples of M cover nonclassical measurement error of various forms. For multivariate x, (3) and (4) apply directly, while (5) could then take the form of a vector of univariate marginal quantiles, for instance. We are now ready to state our main result.

Theorem 1 *Under Assumptions 1-5, given the true observed density $f_{yx|z}$, the equation*

$$f_{yx|z}(y, x|z) = \int_{\mathcal{X}^*} f_{x|x^*}(x|x^*) f_{y|x^*}(y|x^*) f_{x^*|z}(x^*|z)\, dx^* \text{ for all } y \in \mathcal{Y},\ x \in \mathcal{X},\ z \in \mathcal{Z} \quad (6)$$

admits a unique solution[3] $\left(f_{y|x^*}, f_{x|x^*}, f_{x^*|z}\right)$. *A similar result holds for*

$$f_{yxz}(y, x, z) = \int_{\mathcal{X}^*} f_{x|x^*}(x|x^*) f_{yx^*}(y, x^*) f_{z|x^*}(z|x^*)\, dx^* \text{ for all } y \in \mathcal{Y},\ x \in \mathcal{X},\ z \in \mathcal{Z}. \quad (7)$$

The proof can be found in the Appendix and can be outlined as follows. Assumption 2 lets us obtain the integral Equation (6) relating the joint densities of the observable variables to the joint densities of the unobservable variables. This equation is then shown to define the following operator equivalence relationship:

$$L_{y;x|z} = L_{x|x^*} \Delta_{y;x^*} L_{x^*|z}, \quad (8)$$

where $L_{y;x|z}$ is defined analogously to $L_{x|z}$ with $f_{x|z}$ replaced by $f_{y,x|z}$ for a given $y \in \mathcal{Y}$ and where $\Delta_{y;x^*}$ is the "diagonal" operator mapping the function $g(x^*)$ to the function $f_{y|x^*}(y|x^*)\, g(x^*)$, for a given $y \in \mathcal{Y}$. Next, we note that the equivalence $L_{x|z} = L_{x|x^*} L_{x^*|z}$ also holds (by integration of (8) over all $y \in \mathcal{Y}$). Isolating $L_{x^*|z}$ to yield

$$L_{x^*|z} = L_{x|x^*}^{-1} L_{x|z}, \quad (9)$$

substituting it into (8) and rearranging, we obtain:

$$L_{y;x|z} L_{x|z}^{-1} = L_{x|x^*} \Delta_{y;x^*} L_{x|x^*}^{-1}, \quad (10)$$

where all inverses can be shown to exist over suitable domains by Assumption 3 and Lemma 1 in the Appendix. Equation (10) states that the operator $L_{y;x|z} L_{x|z}^{-1}$ admits a spectral decomposition (specifically, an eigenvalue-eigenfunction decomposition in this case). The operator

[3] More formally, if multiple solutions exist, they differ only on a set of zero probability.

to be diagonalized is defined in terms of observable densities, while the resulting eigenvalues $f_{y|x^*}(y|x^*)$ and eigenfunctions $f_{x|x^*}(\cdot|x^*)$ (both indexed by $x^* \in \mathcal{X}^*$) provide the unobserved densities of interest. To ensure uniqueness of this decomposition, we employ four techniques. First, a powerful result from spectral analysis (Theorem XV 4.5 in Dunford and Schwartz (1971)) ensures uniqueness up to some normalizations. Second, the *a priori* arbitrary scale of the eigenfunctions is fixed by the requirement that densities must integrate to one. Third, to avoid any ambiguity in the definition of the eigenfunctions when degenerate eigenvalues are present, we use Assumption 4 and the fact that the eigenfunctions (which do not depend on y, unlike the eigenvalues $f_{y|x^*}(y|x^*)$) must be consistent across different values of the dependent variable y. Finally, in order to uniquely determine the ordering and indexing of the eigenvalues and eigenfunctions, we invoke Assumption 5: If one considers another variable $\tilde{x}^*$ related to x^* through $x^* = R(\tilde{x}^*)$, we have

$$M\left[f_{x|\tilde{x}^*}(\cdot|\tilde{x}^*)\right] = M\left[f_{x|x^*}(\cdot|R(\tilde{x}^*))\right] = R(\tilde{x}^*),$$

which is only equal to $\tilde{x}^*$ if R is the identity function. These four steps ensure that the diagonalization operation uniquely specifies the unobserved densities $f_{y|x^*}(y|x^*)$ and $f_{x|x^*}(x|x^*)$ of interest. Next, Equation (9) implies that $f_{x^*|z}(x^*|z)$ is also identified. Since the identities (10) and (9) use and provide the same information as Equation (6), this establishes uniqueness of the solution to Equation (6). The second conclusion of the Theorem (Equation (7)) follows by similar manipulations.

It should be noted that the assumptions are not mutually contradictory: Models that satisfy all of them can easily be constructed. For instance, one can set $f_{xyzx^*}(x,y,z,x^*) = f_{x|x^*}(x|x^*) f_{y|x^*}(y|x^*) f_{z|x^*}(z|x^*) f_{x^*}(x^*)$ where $f_{x^*}(x^*)$ is a normal and where $f_{x|x^*}(x|x^*)$, $f_{y|x^*}(y|x^*)$, $f_{z|x^*}(z|x^*)$ each are homoskedastic normals whose means depend linearly on x^* (with nonzero slope) and such that $E[x|x^*] = x^*$.

It is possible to replace $f_{y,x|z}(y,x|z)$ by $E[y|x,z] f_{x|z}(x|z)$ and $f_{y|x^*}(y|x^*)$ by $E[y|x^*]$ throughout to obtain an identification result for $E[y|x^*]$ directly, without fully identifying $f_{y|x^*}(y|x^*)$. This would slightly weaken Assumption 2(i) to $E[y|x,x^*,z] = E[y|x^*]$. However, under this approach, the analogues of Assumptions 1 and 4 would become somewhat

restrictive for univariate y and x^*, requiring $E[y|x^*]$ to be strictly monotone in x^* and such that $\sup_{x^* \in \mathcal{X}^*} |E[y|x^*]| < \infty$. These restrictions are avoided if identification of $E[y|x^*]$ is secured through the identification of $f_{y|x^*}(y|x^*)$.

While Theorem 1 establishes identification, we can also show that the model is actually overidentified, thus permitting a test of the model. Equation (6) relates a function of 3 variables to a triplet of functions of 2 variables. Since the set of functions of 3 variables is much "larger" than the set of triplets of functions of 2 variables, there exist densities $f_{yx|z}(y, x|z)$ that cannot be generated by Equation (6), a telltale sign of an overidentifying restriction. The availability of more than one valid instrument offers further opportunities to test the model's assumptions.

3 Estimation using sieve maximum likelihood

As a starting point, we consider a model expressed in terms of the observed variable y and the unobserved mismeasured regressor x^*,

$$f_{y|x^*}(y|x^*; \theta). \tag{11}$$

It is often convenient to decompose the potentially infinite-dimensional parameter θ that we seek to estimate into two subvectors: b, a finite-dimensional parameter vector of interest, and η, a potentially infinite-dimensional nuisance parameter. Naturally, we assume that the parametrization (11) does not include redundant degrees of freedom, i.e., $\theta \equiv (b, \eta)$ is identified if $f_{y|x^*}$ is identified.

This framework nests most commonly used models as subcases. First, setting $\theta \equiv b$ covers the parametric likelihood case (which will then become semiparametric once we account for measurement error). Second, models defined via conditional moment restrictions $E[m(y, x^*, b)|x^*] = 0$ can be considered by defining a family of densities $f_{y|x^*}(y|x^*; b, \eta)$ such that $\int f_{y|x^*}(y|x^*; b, \eta)\, m(y, x^*, b)\, dy = 0$ for all b and η, which is clearly equivalent to imposing a moment condition. For example, in a nonlinear regression model $y = g(x^*, b) + \epsilon$ with $E(\epsilon|x^*) = 0$, we have $f_{y|x^*}(y|x^*; b, \eta) = f_{\epsilon|x^*}(y - g(x^*, b)|x^*)$. The infinite-dimensional

nuisance parameter η is the conditional density $f_{\epsilon|x^*}(\cdot|\cdot)$, constrained to have zero mean. Another important example is the quantile regression case[4] (where the conditional density $f_{\epsilon|x^*}(\cdot|\cdot)$ is constrained to have its conditional τ-quantile equal to 0). Quantile restrictions are useful, as they provide the fundamental concept enabling a natural treatment of non-separable models (e.g. Chernozhukov, Imbens, and Newey (2006), Matzkin (2003), Chesher (2003)). More generally, our framework also covers most semiparametric setups. For instance, one could devise a family of densities $f_{y|x^*}(y|x^*; b, \eta)$ such that b sets the value of the average derivative $\int (dE[y|x^*]/dx^*) w(x^*) dx^*$ (for some weighting function $w(x^*)$), while η controls all remaining degrees of freedom that affect the shape of the density but that do not affect the value of the average derivative. More examples of a partition of θ into b and η can be found in Shen (1997).

Given a model expressed in terms of the true unobserved variables (11), Equation (6) in Theorem 1 suggests a corresponding measurement-error robust sieve maximum likelihood estimator (e.g. Grenander (1981), Shen (1997), Chen and Shen (1998), Ai and Chen (2003)):

$$\left(\hat{\theta}, \hat{f}_1, \hat{f}_2\right) = \arg \max_{(\theta, f_1, f_2) \in \mathcal{A}_n} \frac{1}{n} \sum_{i=1}^{n} \ln \int_{\mathcal{X}^*} f_{y|x^*}(y_i|x^*; \theta) f_1(x_i|x^*) f_2(x^*|z_i) dx^*. \tag{12}$$

Here, $(x_i, y_i, z_i)_{i=1}^n$ is an i.i.d. sample and $\mathcal{A}_n$ is a sequence of approximating sieve spaces that contain progressively more flexible parametric approximations to the densities (as sample size n increases). Functions in $\mathcal{A}_n$ are required to satisfy Assumption 5 as well as normalizations ensuring that the relevant conditional densities suitably integrate to 1. The remaining assumptions made in the identification theory are regularity conditions that the generating process is assumed to satisfy but that do not need to be imposed in the estimation procedure. Typically, the approximating spaces $\mathcal{A}_n$ are generated by the span of series approximations that are linear in the coefficients, such as polynomials, splines, etc. In this case, all restrictions on $\mathcal{A}_n$ imposed by the original semiparametric model or by Assumption 5 can be easily implemented, since they amount to imposing linear restrictions on the coefficients of the sieve representing the unknown densities.

[4]The nonsmoothness of the moment conditions in this case does not pose special problems, because all quantities are effectively smoothed by the truncated series used to represent all densities.

The Supplementary Material available on the Econometrica Web site fully develops the asymptotic theory of the proposed sieve maximum likelihood estimator. A nonparametric consistency result (in a weighted sup norm) is provided as well as a semiparametric root n consistency and asymptotic normality result for the estimated parametric component b of the parameter θ. Our treatment allows for the support of all variables y, x^*, x, z to be unbounded. For the purpose of simplicity and conciseness, our treatment provides primitive sufficient conditions for the independent and identically distributed case. However, since our estimator takes the form of a semiparametric sieve estimator, the very general treatment of Shen (1997) and Chen and Shen (1998) can be used to establish asymptotic normality and root n consistency under a very wide variety of conditions that include dependent and nonidentically distributed data. The regularity conditions invoked for the asymptotic theory fall into three general classes:

1. Smoothness and boundedness restrictions that limit the "size" of the space of functions considered in order to obtain the compactness of the parameter space (where, here, the parameters include functions) that is traditionally invoked to show consistency.
2. Envelope conditions that limit how rapidly the objective function can change in value as the parameters change, in order to help secure stochastic equicontinuity and uniform convergence results.
3. Sieve approximation rates (i.e. at what rate must the number of terms in the series increase to guarantee a given rate of the decay of the approximation error).

The practical implementation of the method requires the selection of the number of terms in the various approximating series. While a formal selection rule for these smoothing parameters (e.g., based on a higher-order asymptotic analysis) would be desirable, it is beyond the scope of the present paper. Some informal guidelines can nevertheless be given. In our semiparametric setting, the selection of the smoothing parameters is somewhat facilitated (relative to fully nonparametric approaches), because semiparametric estimators are known to have the same asymptotic distribution for a wide range of smoothing parameter sequences. This observation suggests that a valid smoothing parameter can be obtained by scanning

a range of values in search of a region where the estimates are not very sensitive to small variations in the smoothing parameter. Typically, for very short series, the smoothing bias dominates and the estimates will exhibit a marked trend as the number of terms is increased. At the other extreme, for very long series, the statistical noise dominates and, although the point estimates vary significantly as additional terms are added, no clear trend should be visible. In between those extremes should lie a region where any clear trend has leveled off and where the random noise in the estimates has not yet grown to an excessive level. The middle of that region points to a suitable value of the smoothing parameters.

A number of straightforward extensions of the above approach are possible. First, the model specified in (11) also could be conditional on any number of other, correctly measured, variables. The same identification proof and estimation method follow, after conditioning all densities on those variables.

The second conclusion of Theorem 1 also suggests an alternative expression for the observed density which proves useful if the model specifies $f_{yx^*}(y, x^*)$ instead of $f_{y|x^*}(y|x^*)$. Our sieve approach, now based on a likelihood expressed in terms of $f_{yxz}(y, x, z)$, covers this case as well. This also enables the treatment of models defined via unconditional moment restrictions (i.e. $E[m(y, x^*, b)] = 0$).

4 Simulations

This section investigates the performance of the proposed estimator with simulated data. We consider a simple parametric probit model

$$f_{y|x^*}(y|x^*) = [\Phi(a + bx^*)]^y [1 - \Phi(a + bx^*)]^{1-y} \text{ for } y \in \mathcal{Y} = \{0, 1\} \tag{13}$$

where (a, b) is the unknown parameter vector and $\Phi(\cdot)$ is the standard normal cdf. In the simulations, we generate the instrumental variable and the latent variable as follows: $z \sim N(1, (0.7)^2)$ and $x^* = z + 0.3(e - z)$ with an independent $e \sim N(1, (0.7)^2)$. The distribution of both z and η are truncated on $[0, 2]$, for simplicity in the implementation. To illustrate our method's ability to handle a variety of assumptions regarding the measurement error,

our examples of generating processes have the general form

$$f_{x|x^*}(x|x^*) = \frac{1}{\sigma(x^*)} f_\nu \left(\frac{x - x^*}{\sigma(x^*)} \right),$$

where f_ν is a density function that will be specified in each example below. We allow for considerable heteroskedasticity, setting $\sigma(x^*) = 0.5 \exp(-x^*)$ in all examples. Sieves for functions of 2 variables are constructed through tensor product bases of univariate trigonometric series. We let i_n and j_n denote the number of terms taken from each of the two series. The smoothing parameters were determined following the guidelines given in the previous section, by locating the middle of a range of values of i_n and j_n over which the point estimates are relatively constant.

We consider three maximum likelihood estimators: (i) the (inconsistent) estimator obtained when ignoring measurement error, (ii) the (infeasible) estimator obtained using error-free data and (iii) the proposed (consistent and feasible) sieve maximum likelihood estimator. We consider models where (i) the mode of f_ν is at zero, (ii) the median of f_ν is at zero and (iii) the 100^{th} percentile of f_ν is at zero. The Supplementary Material presents additional simulation examples.

The simulation results (see Table 1) show that our proposed estimator performs well under a variety of identification conditions. The sieve estimator has a considerably smaller bias than the estimator ignoring the measurement error. As expected, the sieve estimator has a larger variance than the other two estimators, due to the estimation of nonparametric components. However, the sieve estimator still achieves a reduction in the overall root mean square error (RMSE), relative to the other feasible estimator.

5 Conclusion

This paper represents the first treatment of a wide class of nonclassical nonlinear errors-in-variables models with continuously distributed variables using instruments (or repeated measurements). Our main identifying assumption exploits the observation that, even though the measurement error may not have zero mean conditional on the true value of the regressor, perhaps some other measure of location, such as the median or the mode, could still

be zero. We show that the identification problem can be cast into the form of an operator diagonalization problem in which the operator to be diagonalized is defined in terms of observable densities, while the resulting eigenvalues and eigenfunctions provide the unobserved joint densities of the variables of interest.

This nonparametric identification result suggests a natural sieve-based semiparametric maximum likelihood estimator that is relatively simple to implement. Our framework enables the construction of measurement error-robust counterparts of parametric likelihood or moment conditions models, as well as numerous semiparametric models. Our semiparametric estimator is shown to be root n consistent and asymptotically normal.

A Proofs

Proof of Theorem 1. By the definition of conditional densities and Assumption 2,

$$\begin{aligned} f_{yx|z}(y,x|z) &= \int f_{yxx^*|z}(y,x,x^*|z)\,dx^* = \int f_{y|xx^*z}(y|x,x^*,z)\,f_{xx^*|z}(x,x^*|z)\,dx^* \\ &= \int f_{y|x^*}(y|x^*)\,f_{xx^*|z}(x,x^*|z)\,dx^* = \int f_{y|x^*}(y|x^*)\,f_{x|x^*z}(x|x^*,z)\,f_{x^*|z}(x^*|z)\,dx^* \\ &= \int f_{y|x^*}(y|x^*)\,f_{x|x^*}(x|x^*)\,f_{x^*|z}(x^*|z)\,dx^*. \end{aligned}$$

This establishes Equation (6) of the Theorem. We now show the uniqueness of the solution.

Let the operators $L_{x|z}, L_{x|x^*}$ and $L_{x^*|z}$, be given by Definition 1 and let

$$\begin{aligned} L_{y;x|z} &: \mathcal{G}(\mathcal{Z}) \mapsto \mathcal{G}(\mathcal{X}) \text{ with } L_{y;x|z}g \equiv \int f_{yx|z}(y,\cdot|z)\,g(z)\,dz \\ \Delta_{y;x^*} &: \mathcal{G}(\mathcal{X}^*) \mapsto \mathcal{G}(\mathcal{X}^*) \text{ with } \Delta_{y;x^*}g \equiv f_{y|x^*}(y|\cdot)\,g(\cdot). \end{aligned}$$

The notation $L_{y;x|z}$ emphasizes that y is regarded as a parameter on which $L_{y;x|z}$ depends, while the operator itself maps functions of z onto functions of x. The $\Delta_{y;x^*}$ operator is a "diagonal" operator since it is just a multiplication by a function (for a given y), i.e. $[\Delta_{y;x^*}g](x^*) = f_{y|x^*}(y|x^*)\,g(x^*)$. By calculating $L_{y;x|z}g$ for an arbitrary $g \in \mathcal{G}(\mathcal{Z})$, we

rewrite Equation (6) as an operator equivalence relationship:

$$\begin{aligned}
\left[L_{y;x|z}g\right](x) &= \int f_{yx|z}(y,x|z)\,g(z)\,dz = \int\int f_{yx,x^*|z}(y,x,x^*|z)\,dx^*g(z)\,dz \\
&= \int\int f_{x|x^*}(x|x^*)\,f_{y|x^*}(y|x^*)\,f_{x^*|z}(x^*|z)\,dx^*g(z)\,dz \\
&= \int f_{x|x^*}(x|x^*)\,f_{y|x^*}(y|x^*)\int f_{x^*|z}(x^*|z)\,g(z)\,dzdx^* \\
&= \int f_{x|x^*}(x|x^*)\,f_{y|x^*}(y|x^*)\left[L_{x^*|z}g\right](x^*)\,dx^* \\
&= \int f_{x|x^*}(x|x^*)\left[\Delta_{y;x^*}L_{x^*|z}g\right](x^*)\,dx^* = \left[L_{x|x^*}\Delta_{y;x^*}L_{x^*|z}g\right](x), \qquad (14)
\end{aligned}$$

where we have used, (i) Equation (6), (ii) an interchange of the order of integration (justified by Fubini's Theorem), (iii) the definition of $L_{x^*|z}$, (iv) the definition of $\Delta_{y;x^*}$ operating on the function $\left[L_{x^*|z}g\right]$ and (v) the definition of $L_{x|x^*}$ operating on the function $\left[\Delta_{y;x^*}L_{x^*|z}g\right]$.

Equation (14) thus implies the following operator equivalence (which holds over the domain $\mathcal{G}(\mathcal{Z})$)

$$L_{y;x|z} = L_{x|x^*}\Delta_{y;x^*}L_{x^*|z}. \qquad (15)$$

By integration over y and noting that $\int_{\mathcal{Y}} L_{y;x|z}\mu(dy) = L_{x|z}$ and $\int_{\mathcal{Y}} \Delta_{y;x^*}\mu(dy) = I$, the identity operator, we similarly get

$$L_{x|z} = L_{x|x^*}L_{x^*|z}. \qquad (16)$$

Since $L_{x|x^*}$ is injective (by Assumption 3), Equation (16) can be written as

$$L_{x^*|z} = L_{x|x^*}^{-1}L_{x|z}. \qquad (17)$$

The inverse is guaranteed to be defined over a sufficiently large domain because the results of the inversion $L_{x|x^*}^{-1}L_{x|z}$ yields a well-defined integral operator $L_{x^*|z}$. Moreover, the operator equivalence (17) holds for the same domain space $\mathcal{G}(\mathcal{Z})$ as in (16) because the inverse operator was applied to the left. The expression (17) for $L_{x^*|z}$ can be substituted into Equation (15) to yield

$$L_{y;x|z} = L_{x|x^*}\Delta_{y;x^*}L_{x|x^*}^{-1}L_{x|z}. \qquad (18)$$

As shown in Lemma 1 below, the fact that $L_{z|x}$ is injective (by Assumption 3) implies that the inverse $L_{x|z}^{-1}$ can be applied "from the right" on each side of Equation (18) to yield:

$$L_{y;x|z}L_{x|z}^{-1} = L_{x|x^*}\Delta_{y;x^*}L_{x|x^*}^{-1}, \tag{19}$$

where the operator equivalence holds over a dense subset of the domain space $\mathcal{G}(\mathcal{X})$. The equivalence can then be extended to the whole domain space $\mathcal{G}(\mathcal{X})$ by the standard extension procedure for linear operators.

The operator $L_{y;x|z}L_{x|z}^{-1}$ is defined in terms of densities of the observable variables x, y and z and can therefore be considered known. Equation (19) states that the known operator $L_{y;x|z}L_{x|z}^{-1}$ admits a spectral decomposition taking the form of an eigenvalue-eigenfunction decomposition.[5] The eigenvalues of the $L_{y;x|z}L_{x|z}^{-1}$ operator are given by the "diagonal elements" of the $\Delta_{y;x^*}$ operator (i.e. $\left\{f_{y|x^*}(y|x^*)\right\}$ for a given y and for all x^*) while the eigenfunctions of the $L_{y;x|z}L_{x|z}^{-1}$ operator are given by the kernel of the integral operator $L_{x|x^*}$, i.e. $\left\{f_{x|x^*}(\cdot|x^*)\right\}$ for all x^*. In order to establish identification of the unobserved functions of interests $f_{y|x^*}(y|x^*)$ and $f_{x|x^*}(\cdot|x^*)$, we need to show that the decomposition (19) is unique.

Theorem XV.4.5 in Dunford and Schwartz (1971) provides necessary and sufficient conditions for the existence of a unique representation of the so-called spectral decomposition of a linear operator. If a bounded operator T can be written as $T = A + N$ where A is an operator of the form

$$A = \int_\sigma \lambda P(d\lambda) \tag{20}$$

where P is a projection-valued measure[6] supported on the spectrum σ, a subset of the complex plane, and N is a "quasi-nilpotent" operator commuting with A, then this representation is unique.

[5] A spectral decomposition of an operator T takes the form of an eigenvalue-eigenfunction decomposition when $(T - \lambda I)$ is not one-to-one for all eigenvalues λ in the spectrum. This can be verified to be the case here, because all eigenfunctions $f_{x|x^*}(\cdot|x^*)$ belong to $\mathcal{G}(\mathcal{X})$ and are mapped to 0 under $(T - \lambda I)$. An example of a spectral decomposition that is not an eigenvalue-eigenfunction decomposition would be one where some of the eigenfunctions lie outside the space of functions considered (e.g. can only be reached by a limiting process).

[6] Just like a real-valued measure assigns a real number to each set in some field, a projection-valued measure, assigns a projection operator to each set in some field (here, the Borel σ-field). A projection operator Q, is one that is *idempotent*, i.e. $QQ = Q$.

The result is applicable to our situation (with $T = L_{y;x|z}L_{x|z}^{-1}$), in the special case where $N = 0$ and $\sigma \subset \mathbb{R}$. The spectrum σ is simply the range of $f_{y|x^*}(y|x^*)$, that is, $\left\{f_{y|x^*}(y|x^*) : x^* \in \mathcal{X}^*\right\}$. Since largest element of the spectrum is bounded (by Assumption 1), the operator T is indeed bounded in the sense required by Dunford and Schwartz' result.[7]

In our situation, the projection-valued measure P assigned to any subset Λ of $\mathbb{R}$ is

$$P(\Lambda) = L_{x|x^*} I_\Lambda L_{x|x^*}^{-1} \tag{21}$$

where the operator I_Λ is defined via

$$[I_\Lambda g](x^*) = 1\left(f_{y|x^*}(y|x^*) \in \Lambda\right) g(x^*).$$

An equivalent way to define $P(\Lambda)$ is by introducing the subspace

$$\mathcal{S}(\Lambda) = \text{span}\left\{f_{x|x^*}(\cdot|x^*) : x^* \text{ such that } f_{y|x^*}(y|x^*) \in \Lambda\right\} \tag{22}$$

for any subset Λ of the spectrum σ. The projection $P(\Lambda)$ is then uniquely defined by specifying that its range is $\mathcal{S}(\Lambda)$ and that its null space is $\mathcal{S}(\sigma \backslash \Lambda)$.

The fact that $\int_\sigma \lambda P(d\lambda) = L_{x|x^*}\Delta_{y;x^*}L_{x|x^*}^{-1}$, thus connecting Equation (19) with Equation (20), can be shown by noting that

$$\int_\sigma \lambda P(d\lambda) \equiv \int_\sigma \lambda \left(\frac{d}{d\lambda}P([-\infty,\lambda])\right) d\lambda = L_{x|x^*}\left(\int_\sigma \lambda \frac{dI_{[-\infty,\lambda]}}{d\lambda} d\lambda\right) L_{x|x^*}^{-1},$$

where the operator in parenthesis can be obtained by calculating its effect on some function $g(x^*)$:

$$\begin{aligned}\left[\int_\sigma \lambda \frac{dI_{[-\infty,\lambda]}}{d\lambda} d\lambda g\right](x^*) &= \int_\sigma \lambda \frac{d}{d\lambda} 1\left(f_{y|x^*}(y|x^*) \in [-\infty,\lambda]\right) g(x^*)\, d\lambda \\ &= \int_\sigma \lambda \delta\left(\lambda - f_{y|x^*}(y|x^*)\right) g(x^*)\, d\lambda = f_{y|x^*}(y|x^*)\, g(x^*) = [\Delta_{y;x^*} g](x^*).\end{aligned}$$

where we have used that the differential of a step function $1(\lambda \leq 0)$ is a Dirac delta $\delta(\lambda)$, which has the property that $\int \delta(\lambda) h(\lambda) d\lambda = h(0)$ for any function $h(\lambda)$ continuous at $\lambda = 0$, and, in particular, for $h(\lambda) = \lambda$. Hence, we can indeed conclude that $\int_\sigma \lambda P(d\lambda) = L_{x|x^*}\Delta_{y;x^*}L_{x|x^*}^{-1}$.

[7] As explained in Section XV.4 of Dunford and Schwartz (1971).

Having established uniqueness of the decomposition (20) does not yet imply that the representation (19) is unique. The situation is analogous to standard matrix diagonalization:

1. Each eigenvalue λ is associated with a unique subspace $\mathcal{S}(\{\lambda\})$, for $\mathcal{S}(\cdot)$ as defined in Equation (22). However, there are multiple ways to select a basis of functions whose span defines that subspace.
   (a) Each basis function can always be multiplied by a constant.
   (b) Also, if $\mathcal{S}(\{\lambda\})$ has more than one dimension (i.e. if λ is degenerate), a new basis can be defined in terms of linear combinations of functions of the original basis.
2. There is a unique mapping between λ and $\mathcal{S}(\{\lambda\})$, but one is free to index the eigenvalues by some other variable (here x^*) and represent the diagonalization by a function $\lambda(x^*)$ and the family of subspaces $\mathcal{S}(\{\lambda(x^*)\})$. The choice of the mapping $\lambda(x^*)$ is not unique. For matrices, it is sufficient to place the eigenvectors in the correct order. For operators, once the order of the eigenfunctions is set, it is still possible to parametrize them in multiple ways (e.g. index them by x^* or by $(x^*)^3$), as illustrated in the Supplementary Material.

Issue 1a is avoided because the requirement that $\int f_{x|x^*}(x|x^*)\,dx = 1$ sets the scale of the eigenfunctions.

Issue 1b above, is handled via Assumption 4. The idea is that the operator $L_{x|x^*}$ defining the eigenfunctions does not depend on y while the eigenvalues given by $f_{y|x^*}(y|x^*)$ do. Hence, if there is an eigenvalue degeneracy involving two eigenfunctions $f_{x|x^*}(\cdot|x_1^*)$ and $f_{x|x^*}(\cdot|x_2^*)$ for some value of y, we can look for another value of y that does not exhibit this problem to resolve the ambiguity. Formally, this can be shown as follows. Consider a given eigenfunction $f_{x|x^*}(\cdot|x^*)$ and let $D(y,x^*) = \left\{\tilde{x}^* : f_{y|x^*}(y|\tilde{x}^*) = f_{y|x^*}(y|x^*)\right\}$, the set of other values of x^* indexing eigenfunctions sharing the same eigenvalue. Any linear combination of functions $f_{x|x^*}(\cdot|\tilde{x}^*)$ for $\tilde{x}^* \in D(y,x^*)$ is a potential eigenfunction of $L_{y;x|z}L_{x|z}^{-1}$. However, if $v(x^*) \equiv \cap_{y\in\mathcal{Y}}\text{span}\left(\left\{f_{x|x^*}(\cdot|\tilde{x}^*)\right\}_{\tilde{x}^*\in D(y,x^*)}\right)$ is one dimensional, then the set $v(x^*)$ will uniquely specify the eigenfunction $f_{x|x^*}(\cdot|x^*)$ (after normalization to integrate to 1). We now proceed by

contradiction and show that if $v(x^*)$ is not one dimensional, then Assumption 4 is violated. Indeed, if $v(x^*)$ has more than one dimension, it must contain at least two eigenfunctions, say $f_{x|x^*}(\cdot|x^*)$ and $f_{x|x^*}(\cdot|\tilde{x}^*)$. This implies that $\cap_{y\in\mathcal{Y}} D(y,x^*)$ must at least contain the two points x^* and $\tilde{x}^*$. By the definition of $D(y,x^*)$, we must have that $f_{y|x^*}(y|x^*) = f_{y|x^*}(y|\tilde{x}^*)$ for all $y \in \mathcal{Y}$, thus violating Assumption 4. (The qualification that the set on which the densities differ must have positive probability merely accounts for the fact that densities that differ on a set of zero probability actually represent the same density.)

Next, Assumption 5 resolves the ordering/indexing ambiguity (issue 2 above) because, if one considers another variable $\tilde{x}^*$ related to x^* through $x^* = R(\tilde{x}^*)$, we have

$$M\left[f_{x|\tilde{x}^*}(\cdot|\tilde{x}^*)\right] = M\left[f_{x|x^*}(\cdot|R(\tilde{x}^*))\right] = R(\tilde{x}^*),$$

which is only equal to $\tilde{x}^*$ if R is the identity function. Having shown that $f_{y|x^*}(y|x^*)$ and $f_{x|x^*}(x|x^*)$ are uniquely determined, we can then show that $f_{x^*|z}(x^*|z)$ is uniquely determined, since $L_{x^*|z} = L_{x|x^*}^{-1} L_{x|z}$ where $L_{x|x^*}$ is now known and where $L_{x|z}$ is also known because its kernel is an observed density.

The second conclusion of the theorem is obtained by noting that

$$\begin{aligned} f_{yxz}(y,x,z) = f_{yx|z}(y,x|z)\, f_z(z) &= \int f_{x|x^*}(x|x^*)\, f_{y|x^*}(y|x^*)\, f_{x^*|z}(x^*|z)\, dx^*\ f_z(z) \\ = \int f_{x|x^*}(x|x^*)\, f_{y|x^*}(y|x^*)\, f_{x^*z}(x^*,z)\, dx^* &= \int f_{x|x^*}(x|x^*)\, f_{y|x^*}(y|x^*)\, f_{x^*}(x^*)\, f_{z|x^*}(z|x^*)\, dx^* \\ &= \int f_{x|x^*}(x|x^*)\, f_{y,x^*}(y,x^*)\, f_{z|x^*}(z|x^*)\, dx^* \end{aligned}$$

and showing that $f_{x|x^*}$, f_{yx^*} and $f_{z|x^*}$ are uniquely determined from f_{yxz}. First, we have already shown that $f_{x|x^*}(x|x^*)$ is identified from $f_{yx|z}(y,x|z)$ (and therefore from $f_{yxz}(y,x,z)$). By Equation (17), $f_{x^*|z}(x^*|z)$ is also identified. Next, $f_{x^*}(x^*) = \int f_{x^*|z}(x^*|z)\, f_z(z)\, dz$ where $f_z(z)$ is observed. Then $f_{z|x^*}(z|x^*) = f_{x^*|z}(x^*|z)\, f_z(z) / f_{x^*}(x^*)$ and, finally, $f_{y,x^*}(y,x^*) = f_{y|x^*}(y|x^*)\, f_{x^*}(x^*)$. Hence the solution to Equation (7) is unique. ■

Lemma 1 *Under Assumption 1, if $L_{z|x}$ is injective, then $L_{x|z}^{-1}$ exists and is densely defined over $\mathcal{G}(\mathcal{X})$ (for $\mathcal{G} = \mathcal{L}^1$, $\mathcal{L}^1_{bnd}$).*

Proof. Under Assumption 1, injectivity of $L_{z|x}$ implies injectivity of $L^{\dagger}_{x|z}$, the adjoint of $L_{x|z}$, by arguments similar to the ones given after Equation (2) and the fact that $g(\cdot)/f_x(\cdot) \in \mathcal{G}(\mathcal{X})$ implies that $g \in \mathcal{G}(\mathcal{X})$.

Next, $L_{x|z}$ can be shown to be injective when viewed as a mapping of $\overline{\mathcal{R}\left(L^{\dagger}_{x|z}\right)}$ into $\mathcal{G}(\mathcal{X})$, where $\overline{\mathcal{R}\left(L^{\dagger}_{x|z}\right)}$ denotes the closure of the range of $L^{\dagger}_{x|z}$. Indeed, by Lemma VI.2.8 in Dunford and Schwartz (1971), $\overline{\mathcal{R}\left(L^{\dagger}_{x|z}\right)}$ is the orthogonal complement of the null space of $L_{x|z}$, denoted $\mathcal{N}\left(L_{x|z}\right)$. It follows that $L^{-1}_{x|z}$ exists.

By Lemma VI.2.8 in Dunford and Schwartz (1971) again, $\overline{\mathcal{R}\left(L_{x|z}\right)}$ is the orthogonal complement of $\mathcal{N}\left(L^{\dagger}_{x|z}\right)$. But since $L^{\dagger}_{x|z}$ is injective, $\mathcal{N}\left(L^{\dagger}_{x|z}\right) = \{0\}$. Hence, $\overline{\mathcal{R}\left(L_{x|z}\right)} = \mathcal{G}(\mathcal{X})$ and $L^{-1}_{x|z}$ is therefore defined on a dense subset of $\mathcal{G}(\mathcal{X})$. ■

References

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795–1843.

Altonji, J. G., and R. L. Matzkin (2005): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73, 1053–1102.

Blundell, R., X. Chen, and D. Kristensen (2003): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *conditionally accepted in Econometrica*.

Bollinger, C. R. (1998): "Measurement Error in the Current Population Survey: A Nonparametric Look," *Journal of Labor Economics*, 16, 576–594.

Bound, J., C. Brown, and N. Mathiowetz (2001): "Measurement Error in Survey Data," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. Leamer, vol. V. Elsevier Science.

Bound, J., and A. B. Krueger (1991): "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right," *Journal of Labor Economics*, 9, 1–24.

Carrasco, M., and J.-P. Florens (2005): "Spectral method for deconvolving a density," Working Paper, University of Rochester.

Carrasco, M., J.-P. Florens, and E. Renault (2005): "Linear Inverse Problems and Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization," in *Handbook of Econometrics*, vol. Vol. 6. Elsevier Science.

Chalak, K., and H. White (2006): "An Extended Class of Instrumental Variables for the Estimation of Causal Effects," Working Paper, UCSD.

Chen, X., H. Hong, and E. Tamer (2005): "Measurement Error Models with Auxiliary Data," *Review of Economic Studies*, 72, 343–366.

Chen, X., H. Hong, and A. Tarozzi (2005): "Semiparametric Efficiency in GMM Models with Nonclassical Measurement Error," Working Paper, Duke University.

Chen, X., and X. Shen (1998): "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica*, 66(2), 289–314.

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.

——— (2001): "Parameter Approximations for Quantile Regressions with Measurement Error," Working Paper CWP02/01, Department of Economics, University College London.

——— (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 1405–1441.

Chesher, A., M. Dumangane, and R. J. Smith (2002): "Duration response measurement error," *Journal of econometrics*, 111, 169–194.

Darolles, S., J.-P. Florens, and E. Renault (2002): "Nonparametric Instrumental Regression," Working Paper 05-2002, Centre de Recherche et Développement en Économique.

d'Haultfoeuille, X. (2006): "On the Completeness Condition in Nonparametric Instrumental Problems," Working Paper, ENSAE, CREST-INSEE and Université de Paris I.

DUNFORD, N., AND J. T. SCHWARTZ (1971): *Linear Operators*. John Wiley & Sons, NY.

GRENANDER, U. (1981): *Abstract Inference*. Wiley Series, New York.

HALL, P., AND J. L. HOROWITZ (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *Annals of Statistics*, 33, 2904–2929.

HAUSMAN, J. (2001): "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left," *Journal of Economic Perspectives*, 15, 57–67.

HAUSMAN, J., W. NEWEY, H. ICHIMURA, AND J. POWELL (1991): "Measurement Errors in Polynomial Regression Models," *Journal of Econometrics*, 50, 273–295.

HECKMAN, J. J., AND E. VYTLACIL (2005): "Structural Equations, Treatment Effects and Econometric Policy Evaluation," *Econometrica*, 73, 669–738.

HOLDERLEIN, S., AND E. MAMMEN (2006): "Identification of Marginal Effects in Nonseparable Models without Monotonicity," Working Paper, University of Mannheim.

HONG, H., AND E. TAMER (2003): "A Simple Estimator for Nonlinear Error in Variable Models," *Journal of Econometrics*, 117, 1–19.

HOROWITZ, J. L. (2006): "Testing A Parametric Model Against A Nonparametric Alternative With Identification Through Instrumental Variables," *Econometrica*, 74, 521–538.

HU, Y. (2007): "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: A General Solution," Working Paper, University of Texas at Austin.

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.

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.

——— (2007): "Estimation of Average Treatment Effects With Misclassification," *Econometrica*, 75, 537–551.

Li, T. (2002): "Robust and consistent estimation of nonlinear errors-in-variables models," *Journal of Econometrics*, 110, 1–26.

Mahajan, A. (2006): "Identification and Estimation of Single Index Models with Misclassified Regressor," *Econometrica*, 74, 631–665.

Mattner, L. (1993): "Some incomplete but boundedly complete location families," *Annals of Statistics*, 21, 2158–2162.

Matzkin, R. L. (2003): "Nonparametric Estimation of Nonparametric Nonadditive Random Functions," *Econometrica*, 71, 1339–1375.

Newey, W. (2001): "Flexible Simulated Moment Estimation of Nonlinear Errors-in-Variables Models," *Review of Economics and Statistics*, 83, 616–627.

Newey, W. K., and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

Schennach, S. M. (2004a): "Estimation of Nonlinear Models with Measurement Error," *Econometrica*, 72, 33–75.

——— (2004b): "Nonparametric Estimation in the Presence of Measurement Error," *Econometric Theory*, 20, 1046–1093.

——— (2007): "Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models," *Econometrica*, 75, 201–239.

Shen, X. (1997): "On Methods of Sieves and Penalization," *Annals of Statistics*, 25, 2555–2591.

Table 1: Simulation results. For each estimator, we report the mean, the standard deviation (std. dev.) and the square root of the mean squared error (RMSE) of the estimators averaged over all 1000 replications. The sample size is 2000.

Error distribution (zero mode): $f_\nu(\nu) = \exp[\nu - \exp(\nu)]$						
Parameter (=true value)	$a = -1$			$b = 1$		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
Ignoring meas. error	-0.5676	0.0649	0.4372	0.6404	0.0632	0.3651
Accurate data	-1.0010	0.0813	0.0813	1.0030	0.0761	0.0761
Sieve MLE	-0.9575	0.2208	0.2249	0.9825	0.1586	0.1596
Smoothing parameters $i_n = 6, j_n = 3$ in f_1; $i_n = 3, j_n = 2$ in f_2						

Error distribution (zero median): $f_\nu(\nu) = \frac{1}{\pi}\left(1 + \left[\frac{1}{2} + \frac{1}{2}\exp(\nu) - \exp(-\nu)\right]^2\right)^{-1}$						
Parameter (=true value)	$a = -1$			$b = 1$		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
Ignoring meas. error	-0.6514	0.0714	0.3559	0.6375	0.0629	0.3679
Accurate data	-1.0020	0.0796	0.0796	1.0020	0.0747	0.0748
Sieve MLE	-0.9561	0.2982	0.3014	0.9196	0.2734	0.2850
smoothing parameters: $i_n = 8, j_n = 8$ in f_1; $i_n = 3, j_n = 2$ in f_2;						

Error distribution (100^{th} percentile at zero): $f_\nu(\nu) = \exp(\nu)$ for $\nu \in [-\infty, 0]$						
Parameter (=true value)	$a = -1$			$b = 1$		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
Ignoring meas. error	-0.5562	0.0601	0.4478	0.693	0.0632	0.3134
Accurate data	-1.0010	0.0813	0.0813	1.003	0.0761	0.0761
Sieve MLE	-0.9230	0.2389	0.2510	1.071	0.2324	0.2429
Smoothing parameters: $i_n = 4, j_n = 6$ in f_1; $i_n = 3, j_n = 2$ in f_2						