Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

Nonparametric Identification and Semiparametric Estimation of Classical Measurement Error Models Without Side Information

S. M. SCHENNACH and Yingyao HU

Virtually all methods aimed at correcting for covariate measurement error in regressions rely on some form of additional information (e.g., validation data, known error distributions, repeated measurements, or instruments). In contrast, we establish that the fully nonparametric classical errors-in-variables model is identifiable from data on the regressor and the dependent variable alone, unless the model takes a very specific parametric form. This parametric family includes (but is not limited to) the linear specification with normally distributed variables as a well-known special case. This result relies on standard primitive regularity conditions taking the form of smoothness constraints and nonvanishing characteristic functions' assumptions. Our approach can handle both monotone and nonmonotone specifications, provided the latter oscillate a finite number of times. Given that the very specific unidentified parametric functional form is arguably the exception rather than the rule, this identification result should have a wide applicability. It leads to a new perspective on handling measurement error in nonlinear and nonparametric models, opening the way to a novel and practical approach to correct for measurement error in datasets where it was previously considered impossible (due to the lack of additional information regarding the measurement error). We suggest an estimator based on non/semiparametric maximum likelihood, derive its asymptotic properties, and illustrate the effectiveness of the method with a simulation study and an application to the relationship between firm investment behavior and market value, the latter being notoriously mismeasured. Supplementary materials for this article are available online.

KEY WORDS: Errors-in-variables; Higher-order moments; Measurement error.

1. INTRODUCTION

Nonlinear regression models in which both the dependent and independent variables are measured with error have received considerable attention over the last few decades (see, e.g., Carroll et al. 2006). This so-called classical nonlinear errors-in-variables model takes the following form.

Model 1. Let $y, x, x^*, \Delta x, \Delta y$ be scalar real-valued random variables related through

$$y = g(x^*) + \Delta y$$
$$x = x^* + \Delta x,$$

where only x and y are observed while all remaining variables are not and satisfy the following assumption:

Assumption 1. The variables x^*, Δx, Δy, are mutually independent, $E[\Delta x] = 0$, and $E[\Delta y] = 0$ (with $E[|\Delta x|] < \infty$ and $E[|\Delta y|] < \infty$).

A well-known result is that when the function $g(x^*)$ is linear while x^*, Δx, and Δy are normal, the model is not identified (see, e.g., Fuller 1987), although the regression coefficients can often be consistently bounded (Klepper and Leamer 1984). This lack of point identification for what is perhaps the most natural regression model has long guided the search for solutions to the errors-in-variables problem toward approaches that rely on additional information (beyond x and y), such as instruments, repeated measurements, validation data, known measurement error distribution, etc. (e.g., Hausman et al. 1991; Li and Vuong 1998; Newey 2001; Wang and Hsiao 2003; Schennach 2004a, b, 2007; Hu and Schennach 2008; Hu and Ridder 2012, among many others).

Nevertheless, since the seminal works of Geary (1942) and Reiersol (1950), a large number of authors (e.g., Kendall and Stuart 1979; Pal 1980; Kapteyn and Wansbeek 1983; Cardoso and Souloumiac 1993; Hyvärinen and Oja 1997; Cragg 1997; Lewbel 1997; Dagenais and Dagenais 1997; Erickson and Whited 2000; Ikeda and Toyama 2000; Erickson and Whited 2002; Beckmann and Smith 2004; Bonhomme and Robin 2009; Bonhomme and Robin 2010, and the many references therein) have exploited independence assumptions (as in Assumption 1 above) to develop alternative methods to identify linear errors-in-variables models and related linear factor models, typically based on the idea that higher-order moments of x and y provide sufficient information to secure identification in the presence of nonnormally distributed variables. Extensions to parametric polynomial models by using selected higher-order moments have also been considered in Chesher (1998) and Kenny and Judd (1984). Some nonlinear factor models have also been considered in Bauer (2005), Yalcin and Amemiya (2001), and Jutten and Karhunen (2003), however, this strand of the literature has largely bypassed the question of identification or has focused on specific cases (such as nonlinear models that can be reduced to linear ones by a suitable transformation). In fact, the question of completely characterizing the set of identifiable models in fully *nonparametric* settings, while fully exploiting the information provided by the joint distribution of all the observable variables to avoid the need for additional information, remains wide open.

We demonstrate that the answer to this long-standing open question turns out to be surprisingly simple, although proving

S. M. Schennach is Professor, Department of Economics, Brown University, Providence, RI 02912 (E-mail: *smschenn@brown.edu*). Yingyao Hu is Associate Professor, Department of Economics, Johns Hopkins University, Baltimore, MD 21218 (E-mail: *yhu@jhu.edu*). The authors thank Daniel Wilhelm and Xavier d'Haultfoeuille for useful comments and Jiaxiong Yao for excellent research assistance. S. M. Schennach acknowledges support from NSF grants SES-0752699 and SES-1061263/1156347.

Journal of the American Statistical Association
March 2013, Vol. 108, No. 501, Theory and Methods
DOI: 10.1080/01621459.2012.751872

so is not. Under fairly simple and natural regularity conditions, a specification of the form $g(x^*) = a + b\ln(e^{cx^*} + d)$ is the *only* functional form that is *not* guaranteed to be identifiable. Even with this specification, the distributions of all the variables must have very specific forms to evade identifiability of the model. As expected, this parametric family includes the well-known linear case (with $d = 0$) with normally distributed variables. Given that this very specific unidentified parametric functional form is arguably the exception rather than the rule, our identification result should have a wide applicability. This leads to a new perspective on handling measurement error in nonlinear and nonparametric models, opening the way to a novel and practical approach to correct for measurement error in datasets where it was previously considered impossible (due to the lack of additional information regarding the measurement error).

Based on this identification result, we suggest a corresponding estimator and derive its asymptotic properties. We illustrate the effectiveness of the method via a simulation study and an application to the relationship between firm investment behavior and market value, the latter being notoriously mismeasured. This application revisits, in general nonlinear settings, the analysis of Erickson and Whited (2000), a well-known successful example of the use of higher-order moments to address measurement error issues in linear models.

2. IDENTIFICATION RESULT

Our identification result will rely on the mutual independence of the model error, the measurement error, and the true regressor (Assumption 1). While such an assumption is arguably strong, it already underlies the extensive and still growing literature on higher-order moments in linear errors-in-variables models (e.g., Reiersol 1950; Kendall and Stuart 1979; Pal 1980; Cragg 1997; Lewbel 1997; Dagenais and Dagenais 1997; Erickson and Whited 2000, 2002; Bonhomme and Robin 2009, 2010). Moreover, even in the measurement error literature that exploits side information, independence assumptions are extremely common (see, e.g., the monograph by Carroll et al. 2006 for a review). On a more fundamental level, the dimensionality of the observables in this problem is only 2 (x and y), while the dimensionality of the unobservables is 3 (Δx, Δy, and x^*). Hence, it is impossible to construct a well-behaved mapping (i.e., other than "fractal" mappings) between the observable and the unobservable distributions without introducing some type of assumption that reduces the dimensionality of the unobservables. Independence achieves this by letting us factor the joint distribution of Δx, Δy, and x^* as products of functions of fewer variables. It is possible that other dimension-reducing assumptions could be concocted, but few, if any, would have the transparency and simplicity of independence assumptions (except perhaps in the case of purely discrete mismeasured regressors (Chen, Hu, and Lewbel 2009), where dimensionality issues can be assumed away with sufficiently strong rank conditions, because all unknown distributions can be characterized by a finite number of unknowns, unlike the continuous case treated in the present article). Independence is also the most logical extension of the existing literature on the topic.

Beyond independence, we also need a few basic regularity conditions.

Assumption 2. $E[e^{i\xi\Delta x}]$ and $E[e^{i\gamma\Delta y}]$ do not vanish for any $\xi, \gamma \in \mathbb{R}$, where $i = \sqrt{-1}$.

The type of assumption regarding the so-called characteristic function has a long history in the deconvolution literature (see, e.g., Fan 1991; Schennach 2004a, and the references therein). The only commonly encountered distributions with a vanishing characteristic function are the uniform and the triangular distributions. We also need a slightly weaker but similar assumption on x and y.

Assumption 3. (i) $E[e^{i\xi x^*}] \neq 0$ for all ξ in a dense subset of $\mathbb{R}$ and (ii) $E[e^{i\gamma g(x^*)}] \neq 0$ for all γ in a dense subset of $\mathbb{R}$ (which may be different than in (i)).

Unlike Assumption 2, this Assumption does allow for these characteristic functions to vanish at points, although not over intervals. This assumption is only needed if one wishes to recover the distribution of the errors (Δx, Δy). Also note that both Assumptions 2 and 3 are implied by the Assumption that $E[e^{i\xi x}] \neq 0$ and $E[e^{i\gamma y}] \neq 0$ everywhere, an assumption that testable, since it involves observables.

Assumption 4. The distribution of x^* admits a uniformly bounded density $f_{x^*}(x^*)$ with respect to the Lebesgue measure that is supported on an interval (which may be infinite).

Assumption 5. The regression function $g(x^*)$ is continuously differentiable over the interior of the support of x^*.

These are standard smoothness constraints.

Assumption 6. The set $\chi = \{x^* : g'(x^*) = 0\}$ has at most a finite number of elements $x_1^*, \ldots, x_m^*$. If χ is nonempty, $f_{x^*}(x^*)$ is continuous and nonvanishing in a neighborhood of each x_k^*, $k = 1, \ldots, m$.

This assumption allows for nonmonotone specifications, but rules out functions that are constant over an interval (not reduced to a point) or that exhibit an *infinite* number of oscillations. This is sufficiently flexible to encompass most specifications of practical interest. Excluding functions that are constant over an interval parallels the assumption of nonzero slope made in linear models (Reiersol 1950) and is therefore difficult to avoid. Without Assumptions 5 and 6, it is difficult to rule out extremely complex and pathological joint distributions of x and y. In particular, one could imagine an extremely rapidly oscillating $g(x^*)$, where nearly undetectable changes in x^* yield changes in y that are virtually observationally indistinguishable from genuine errors in y.

Our main result can then be stated as follows, after we recall the following convenient concept.

Definition 1. We say that a random variable r has an *F factor* if r can be written as the sum of two independent random variables (which may be degenerated), one of which has the distribution F. (This is related to the concept of a decomposable characteristic functions, see Lukacs 1970, sec. 5.1. We allow for degenerate factors here to simplify the statement of the theorem below.)

Theorem 1. Let Assumptions 1–6 hold. There are three mutually exclusive cases.

1. $g(x^*)$ is *not* of the form
$$g(x^*) = a + b\ln(e^{cx^*} + d) \tag{1}$$
for some constants $a, b, c, d \in \mathbb{R}$. Then, $f_{x^*}(x^*)$ and $g(x^*)$ (over the support of $f_{x^*}(x^*)$) and the distributions of Δx and Δy in Model 1 are *identified*.
2. $g(x^*)$ *is* of the form (1) with $d > 0$ (A case where $d < 0$ can be converted into a case with $d > 0$ by permuting the roles of x and y). Then, neither $f_{x^*}(x^*)$ nor $g(x^*)$ in Model 1 are identified iff x^* has a density of the form
$$f_{x^*}(x^*) = A\exp(-Be^{Cx^*} + CDx^*)(e^{Cx^*} + E)^{-F} \tag{2}$$
with $C \in \mathbb{R}$, $A, B, D, E, F \in [0, \infty)$ and Δy has a Type I extreme value factor (whose density has the form $f_u(u) = K_1\exp(K_2\exp(K_3u) + K_4u)$ for some $K_1, K_2, K_3, K_4 \in \mathbb{R}$).
3. $g(x^*)$ *is* linear (i.e., of the form (1) with $d = 0$). Then, neither $f_{x^*}(x^*)$ nor $g(x^*)$ in Model 1 are identified iff x^* is normally distributed and either Δx or Δy has a normal factor.

This identification result establishes when the knowledge of the joint distribution of the observable variables y and x uniquely determines the unobservable quantities of interest: $g(x^*)$ and the distributions of x^*, Δx, and Δy. In other words, it provides conditions under which there cannot be two different models that generate the same joint distribution of the observable variables x and y. Intuitively, this result is made possible by the fact that the observable quantity (the joint density of x and y) is a function of two variables while the unobservable quantities ($g(x^*)$, and the marginal distribution of x^*, Δx, Δy) are all functions of one variable. The former thus "contains" much more information than the latter, so it is intuitively natural that it should be possible to recover the unobservables from the observables alone. The phrasing of Cases 2 and 3 should make it clear that the conclusion of the theorem remains unchanged if one focuses on identifying $g(x^*)$ only and not $f_{x^*}(x^*)$, because the observationally equivalent models ruling out identifiability have different regression functions in all of the unidentified cases.

The proof of this result (outlined in the Appendix and detailed in Section A of the supplementary material) proceeds in five broad steps:

1. We reduce the identification problem of a model with errors along x and y into the equivalent problem of finding two observationally equivalent models, one having errors only along the x-axis and one having errors only along the y-axis.
2. We rule out a number of pathological cases in which the error distributions do not admit densities with respect to the Lebesgue measure by showing that such occurrences would actually imply identification of the model (in essence, any nonsmooth point gives away the shape of the regression function).
3. We show that any point of nonmonotonicity in the regression function makes it impossible to find two distinct but observationally equivalent models, because any extremum in the regression function introduces a nonsmooth point in the density of some observable variables and arguments similar to point 2 can be invoked.
4. We derive necessary conditions for lack of identification that take the form of differential equations involving all densities. This establishes that the large class of models where these equations do not hold are identified.
5. Cases that do satisfy the differential equations are then systematically checked to see if they yield valid densities for all variables, thus pointing toward the only cases that are actually not identified and securing necessary and sufficient conditions for identifiability.

It is somewhat unexpected that in a fully nonparametric setting, the nonidentified family of regression functions would still be parametric with such a low dimension (only four adjustable parameters). It is perhaps not entirely surprising that in the a priori difficult case of normally distributed regressors, most nonlinear specifications are actually identified, since nonlinearity necessarily destroys normality of some of the variables. While our findings regarding linear regressions (Case 3) coincide with Reiersol (1950), the functional forms in the other nonidentified models (Case 2) are hardly trivial and would have been difficult to find without a systematic approach such as ours. Section B of the supplementary material provides independent verification of Case 2 and shows that the constants $a, b, c, d, A, B, C, D, E, F$ can all be set so as to yield two distinct but observationally equivalent models with proper densities.

An interesting feature of Case 2 is that there are only two observationally equivalent models in this case and they are disjoint: one has the form (1) with $d = d_1 < 0$ and the other, $d = d_2 > 0$ but models with $d \in]d_1, d_2[$ are not observationally equivalent. One cannot smoothly go from one model to another observationally equivalent one without going through models that are not observationally equivalent. Hence, Theorem 1 implies that the model is locally identified in Cases 1 and 2. Moreover, in Case 2, it is usually easy to rule out one of the two possible models based on simple considerations regarding the process being studied. One of the two models (with $d < 0$) has a vertical asymptote while the other (with $d > 0$) has a horizontal asymptote. The vertical asymptote is usually incompatible with any reasonable model, since it implies an infinite response to a finite cause. Hence, the only real situation of practical concern could possibly be the linear specification of Case 3. We will return to the linear case when discussing estimation.

In summary, Theorem 1 shows that the errors-in-variables model is identified for virtually all commonly used specifications: exponential, sine, cosine, polynomial (not reduced to a line), logistic, etc. Theorem 1 can be straightforwardly extended to include perfectly observed covariates w, simply by conditioning all densities (and expectations) on these covariates. Theorem 1 then establishes identification of $f_{x^*|w}(x^*|w)$ and $g(x^*, w) \equiv E[y|x^*, w]$ and therefore of $f_{x^*,w}(x^*, w) = f_{x^*|w}(x^*|w) f_w(w)$.

3. ESTIMATION

Assumption 1 implies that the observable density $f_{yx}(y, x)$ is related to the unobservable regression function of interest $g(x^*)$ and the densities of the unobserved variables: $f_{x^*}(x^*)$,

$f_{\Delta y}(\Delta y)$, $f_{\Delta x}(\Delta x)$ via the following integral equation

$$f_{yx}(y, x) = \int f_{\Delta y}(y - g(x^*)) f_{\Delta x}(x - x^*) f_{x^*}(x^*) dx^*. \quad (3)$$

Since our identification result provides conditions under which this equation admits a unique (functional) solution $(g, f_{x^*}, f_{\Delta y}, f_{\Delta x})$, this suggests an analog estimator maximizing the likelihood associated with the density $f_{yx}(y, x)$, in which the shape of all unknown functions on the right-hand side of (3) are jointly optimized. To implement this idea in practice, for a given iid sample $(y_i, x_i)_{i=1}^n$, we employ a sieve maximum likelihood estimator (Shen 1997) based on the equation

$$g = \arg\max_g \sup_{(f_1, f_2, f_3)} \frac{1}{n} \sum_{i=1}^n \ln \int f_1(y_i - g(x^*)) f_2(x_i - x^*) \times f_3(x^*) dx^*, \quad (4)$$

where the max and sup are taken over suitably restricted sets of functions and g is regression function of interest, while f_1, f_2, and f_3, respectively, denote the densities of the model error, the measurement error, and the true regressor. The restrictions include (i) constraints that the densities integrate to one and (ii) zero-mean constraints on the error densities. Also, all four unknown functions g, f_1, f_2, and f_3 are represented by truncated series, with the number of terms in the series increasing with sample size. In our simulations and application below, we rely on a Hermite orthogonal series, which offers the advantage that all required integrals (e.g., in Equation (4)) can be carried out analytically. As is well known in the theory of nonparametric likelihoods, such sample-size-dependent restrictions on the number of terms in the series approximations are necessary to regularize the behavior of the estimator and achieve consistency. These restrictions are detailed in the asymptotic analysis. As an alternative to truncated series, one is free to employ flexible functional forms or even parametric models. Our identification result guarantees that the solution is (asymptotically) unique regardless of the choice of approximation scheme. Given this guarantee of a unique solution, it is not surprising that our likelihood function turns out to be rather well behaved, thus enabling us to employ a standard numerical optimization routine to maximize it: an L-BFGS quasi-Newton algorithm (Nocedal 1980).

In the following examples, we consider the estimation of a parametric regression model, that is, Model 1 with $g(x^*) = m(x^*; \theta)$, where the function $m(x^*; \theta)$ is known up to a parameter vector θ. However, the densities of the unobserved variables Δx, Δy, and x^* are treated nonparametrically. The rationale for this approach is that the convergence rate of a fully nonparametric measurement error model can be very slow, while our semiparametric approach enables root n consistency for the parameter vector of interest, making the approach practical for typically available sample sizes. The use of a parametric regression specification also parallels the focus of the vast majority of the empirical literature, while the nonparametric treatment of the distributions of Δx, Δy, and x^* frees researchers from having to assume specify parametric forms for quantities that do not need to be specified in traditional, measurement error-free, regressions. Hence, the proposed method offers a direct substitute to conventional regression analysis when measurement error is suspected. Our asymptotic theory can be adapted to other semiparametric context, for instance, leaving $g(x^*)$ fully nonparametric but focusing on semiparametric functionals of it (such as average derivatives).

Even though $g(x^*)$ and the densities of the errors are unknown a priori, in practice, there is no real need to worry about checking the functional form restrictions of Theorem 1. First, as explained in Section 2, as soon as a vertical asymptote in $g(x^*)$ can be ruled out, the nonlinear nonidentified case is of little concern and only the linear case remains a potential issue. Next, we observe that if the true model were "too close" to the linear unidentified case for the method to be useful, the likelihood function in (4) would be very flat near its maximum, resulting in very large standard errors. Theorem 1 is nevertheless practically useful: it indicates that the approach is certainly worth trying, since the cases leading to lack of identification are so special and rare. But ultimately, what determines whether this approach leads to useful inference in practice in a given application is the magnitude of the estimated standard errors on the parameters of interest.

Section C of the supplementary material presents a formal asymptotic analysis of this estimator with suitable regularity conditions for consistency as well as root n consistency and asymptotic normality of an estimator of θ. See Newey (2001), Mahajan (2006), Hu and Schennach (2008), and Carroll, Chen, and Hu (2010a), among others, for other examples of the use of sieve maximum likelihood in measurement error models and Schennach (2009) and Carroll, Chen, and Hu (2010b) for further details, and an extensive discussion of the practical use of sieves in this context. It should be noted that root n consistency and asymptotic normality should not be taken for granted in this context—this ideal can only be reached under smoothness, moment existence, and dominance conditions that imply that the estimator admits an asymptotically linear representation in a neighborhood of the truth. Such conditions may be difficult to ascertain formally in applications, because the data generating process is not exactly known. If in doubt, practitioners could try the estimator at a few sample sizes to check if the variances estimates indeed scale as n^{-1}, which would be a good indication that the asymptotic regime has been reached and that it behaves as expected by the theory.

The practical implementation of the method requires the selection of suitable smoothing parameters: the number of terms in each of the truncated series approximations. The construction of a general data-driven smoothing parameter selection procedure and a formal proof of its asymptotic validity is beyond the scope of this article. Nevertheless, our asymptotic theory provides very useful guidance regarding the choice of the smoothing parameters in practice. In a semiparametric context, our asymptotic theory implies that the limiting distribution of the estimator is identical for a wide range of rates of change of the smoothing parameters with sample size (since our assumptions do not require a specific rate but instead take the form of upper and lower bounds on these rates). In fact, not only are the limiting distributions identical, but the difference between two estimators obtained with different choices of smoothing parameters that satisfy our assumptions is asymptotically negligible (relative to the $n^{-1/2}$ leading term of their asymptotic expansion). This suggests that a very direct way to check if a choice of smoothing parameter is appropriate is to simply check the sensitivity of the

results to variations in the smoothing parameters. The region, in smoothing parameter space, yielding estimates that are the least sensitive to given small changes in smoothing parameter values very likely points to valid smoothing parameter choices. Choices outside of that region will tend to either exhibit marked randomness if too many terms in the series are kept (due to increased variance) or exhibit a marked systematic trend if too few terms in the series are kept (due to an increased bias). Insensitivity to smoothing parameter selection (near the optimal choice) represents another advantage of the use of a semiparametric model (instead of fully nonparametric one) and we rely on it in our simulations and empirical application below.

4. SIMULATIONS

We consider a nonlinear regression model as follows:

$$y = m(x^*;\theta) + \Delta y$$
$$x = x^* + \Delta x.$$

The latent variable x^* is drawn from a mixture of two normal distributions $0.6N(0, 1) + 0.4N(0.2, 0.25)$, and the regression error Δy has a normal distribution $N(0, 0.9)$. The measurement error Δx has a de-meaned extreme value distribution $F_{\Delta x}(\Delta x) = 1 - \exp(-\exp(2\Delta x - \gamma_{\Delta x}))$ with $\gamma_{\Delta x} = 0.5772$. Moreover, the right-hand side variables $(\Delta y, \Delta x, x^*)$ are mutually independent. In the simulation, we draw a random sample $\{y_i, x_i, x_i^*\}_{i=1,\ldots,n}$ based on this model for $n = 3000$.

We use the Hermite orthogonal series as our sieve basis functions. Let $p_n(\cdot)$ be the Hermite orthogonal series. We have

$$p_n(x) = \sqrt{\frac{1}{\sqrt{\pi} n! 2^n}} H_n(x) e^{-\frac{x^2}{2}},$$

where $H_0(x) = 1$, $H_1(x) = 2x$, and $H_{n+1}(x) = 2x H_n(x) - 2n H_{n-1}(x)$. The sieve expressions of the nonparametric densities are

$$f_1(\Delta y) = \left[\sum_{i=0}^{k_{\Delta y}} \beta_i^{\Delta y} p_i(\Delta y)\right]^2,$$
$$f_2(\Delta x) = \left[\sum_{i=0}^{k_{\Delta x}} \beta_i^{\Delta x} p_i(\Delta x)\right]^2,$$

and

$$f_3(x^*) = \left[\sum_{i=0}^{k_x} \beta_i^{x} p_i(x^*)\right]^2.$$

The smoothing parameters are $k_{\Delta y}$, $k_{\Delta x}$, and k_x. One can show that the restriction $\int f_1(\Delta y) d\Delta y = 1$ implies $\sum_{i=0}^{k_{\Delta y}} [\beta_i^{\Delta y}]^2 = 1$, and similarly for $\beta_i^{\Delta x}$ and β_i^x. Furthermore, the zero mean assumption $\int \Delta y f(\Delta y) d\Delta y = 0$ implies that $\sum_{i=0}^{k_{\Delta y}-1} \sqrt{2(i+1)} \beta_i^{\Delta y} \beta_{i+1}^{\Delta y} = 0$, and similarly for $\beta_i^{\Delta x}$.

In addition, we consider three related estimators. One is the infeasible nonlinear regression of y on x^*:

$$\widehat{\theta}_{\text{nls}} = \arg\max_{\theta} \sum_{i=1}^{n} -[y_i - m(x_i^*;\theta)]^2,$$

which would be the best estimator, under homoscedasticity, if x^* were hypothetically available in the sample. Another estimator is naive nonlinear least squares (NLS), which ignores the measurement error, as follows:

$$\widehat{\theta}_{\text{nnls}} = \arg\max_{\theta} \sum_{i=1}^{n} -[y_i - m(x_i;\theta)]^2.$$

This estimator should give us the largest bias. Finally, we consider the sieve-based instrumental variable estimator of Hu and Schennach (2008), denoted $\widehat{\theta}_{\text{HS}}$, which is consistent in the presence of measurement error, but requires the availability of an instrument. To ensure a meaningful comparison, we specialized $\widehat{\theta}_{\text{HS}}$ to the case where all the error terms and x^* are mutually independent (as assumed for $\widehat{\theta}_{\text{sieve}}$), for otherwise, allowing for general form of heteroscedasticity in $\widehat{\theta}_{\text{HS}}$ would have caused an efficiency penalty relative to $\widehat{\theta}_{\text{sieve}}$. We would expect $\widehat{\theta}_{\text{HS}}$ to have properties roughly similar to $\widehat{\theta}_{\text{sieve}}$, but probably with smaller standard errors, since it exploits additional information (the instrument). In this case, we use, as an instrument, a repeated measurement with a normally distributed measurement error of variance 0.4. (Note that the variance of the first measurement error is 0.41, so both measurement are about equally informative.)

We consider six specifications of the regression function

$$\begin{aligned}
&\text{case 1:} && m(x;\theta) = \theta_1 x + \theta_2 e^x,\\
&\text{case 2:} && m(x;\theta) = \theta_1 x + \theta_2 x^2,\\
&\text{case 3:} && m(x;\theta) = \theta_1 x + \theta_2/(1+x^2),\\
&\text{case 4:} && m(x;\theta) = (x^2 + \theta_1)(x + \theta_2),\\
&\text{case 5:} && m(x;\theta) = \ln(1 + \theta_1 x + \theta_2 x^2),\\
&\text{case 6:} && m(x;\theta) = \theta_1 x + \theta_2 \ln(1 + x^2).
\end{aligned}$$

For each specification, we estimate the model using the three estimators with 400 randomly generated samples of 3000 observations. We report the mean, standard deviations (std. dev.), and squared root of mean square error (RMSE) of the four estimators $\widehat{\theta}_{\text{sieve}}$, $\widehat{\theta}_{\text{HS}}$, $\widehat{\theta}_{\text{nls}}$, and $\widehat{\theta}_{\text{nnls}}$. The smoothing parameters are chosen, as motivated in the previous section, by identifying a region where the estimates are not very sensitive to variations in the smoothing parameter (i.e., when changes in the means of the point estimates are small relative to their standard deviations, where both quantities are estimated via averages over the randomly generated samples). The smoothing parameters are kept constant across the randomly generated samples. (Section D.1 of the supplementary material reports smoothing parameter sweeps that illustrate this procedure.)

As shown in Table 1, the biases of $\widehat{\theta}_{\text{sieve}}$, $\widehat{\theta}_{\text{nls}}$, and $\widehat{\theta}_{\text{HS}}$ are small compared with $\widehat{\theta}_{\text{nnls}}$, because they are consistent. The variances of $\widehat{\theta}_{\text{sieve}}$ and $\widehat{\theta}_{\text{HS}}$ should be the largest of the four due to the nonparametric approximation. Nevertheless, the sieve estimator $\widehat{\theta}_{\text{sieve}}$ is preferable over the naive estimator in terms of mean squared errors. The comparison between $\widehat{\theta}_{\text{sieve}}$ and $\widehat{\theta}_{\text{HS}}$ is instructive, as it reveals that although $\widehat{\theta}_{\text{sieve}}$ is generally less efficient than $\widehat{\theta}_{\text{HS}}$ (as expected), it is often able to approach the RMSE of $\widehat{\theta}_{\text{HS}}$, even though it relies on less information (no instrument). This indicates that our approach offers a very practical alternative to instrumental variable-based methods. Section D.2 of the supplementary material reports similar results for a smaller sample of only 500 observations that indicate that the bias-reducing power of the method remains down to such sample sizes (although the variance of all estimators obviously increases).

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 400 replications. The sample size is 3000. The selected smoothing parameters are $k_{\Delta y} = 5$, $k_{\Delta x} = 5$, $k_x = 6$

	$\theta_1 = 1$			$\theta_2 = 1$		
Parameter (=true value)	Mean	Std. dev.	RMSE	Mean	Std. dev.	RMSE
Case 1: $m(x;\theta) = \theta_1 x + \theta_2 e^x$						
Ignoring measurement error (meas. error)	0.415	0.081	0.590	0.739	0.053	0.267
Accurate data	1.001	0.028	0.028	1.000	0.010	0.010
Hu and Schennach (2008)	0.883	0.115	0.164	1.037	0.097	0.104
Sieve MLE	1.059	0.213	0.221	0.925	0.145	0.163
Case 2: $m(x;\theta) = \theta_1 x + \theta_2 x^2$						
Ignoring meas. error	0.755	0.031	0.246	0.537	0.028	0.463
Accurate data	1.001	0.021	0.021	1.002	0.011	0.012
Hu and Schennach (2008)	0.942	0.082	0.100	0.926	0.089	0.116
Sieve MLE	0.961	0.062	0.073	0.937	0.060	0.087
Case 3: $m(x;\theta) = \theta_1 x + \theta_2/(1+x^2)$						
Ignoring meas. error	0.631	0.022	0.370	1.037	0.028	0.046
Accurate data	1.000	0.020	0.020	1.000	0.023	0.023
Hu and Schennach (2008)	1.008	0.032	0.033	1.015	0.027	0.031
Sieve MLE	0.959	0.080	0.089	1.053	0.038	0.065
Case 4: $m(x;\theta) = (x^2 + \theta_1)(x + \theta_2)$						
Ignoring meas. error	−0.302	0.079	1.305	1.625	0.284	0.687
Accurate data	1.000	0.017	0.017	1.000	0.012	0.012
Hu and Schennach (2008)	1.055	0.115	0.127	1.077	0.166	0.183
Sieve MLE	1.080	0.145	0.166	1.089	0.150	0.174
Case 5: $m(x;\theta) = \ln(1 + \theta_1 x + \theta_2 x^2)$						
Ignoring meas. error	0.512	0.033	0.489	0.456	0.028	0.545
Accurate data	1.001	0.048	0.048	1.000	0.045	0.045
Hu and Schennach (2008)	1.092	0.083	0.125	1.123	0.130	0.179
Sieve MLE	0.844	0.120	0.197	0.966	0.067	0.075
Case 6: $m(x;\theta) = \theta_1 x + \theta_2 \ln(1 + x^2)$						
Ignoring meas. error	0.662	0.019	0.339	0.722	0.029	0.279
Accurate data	1.000	0.020	0.020	0.997	0.029	0.029
Hu and Schennach (2008)	0.988	0.061	0.062	0.791	0.171	0.270
Sieve MLE	0.915	0.059	0.104	0.979	0.054	0.058

While the nonlinear ($d \neq 0$) nonidentified case poses little problem in practice (as explained at the end of Section 2), it is instructive to investigate how the sieve estimator behaves as one approaches the linear ($d = 0$) unidentified case. Table 2 shows that failure of identification is readily detected via the associated sharp increases in the standard errors, as expected from the fact that for a locally unidentified model, the likelihood function is locally flat.

Table 2. Study of the behavior of the estimator near a nonidentified case

θ_2	Std. dev. θ_1
2.0	0.11
1.5	0.10
1.0	0.09
0.5	0.12
0.0	0.30

NOTE: We use the specification $m(x*, \theta) = \theta_1 x^* + \theta_2 (x^*)^2$ with $x^* \sim N(0.08, 0.4)$, $\Delta x \sim N(0, 0.41)$, $\Delta y \sim N(0, 0.9)$, and $\theta_1 = 1$. We consider a range of values of θ_2 and calculate the corresponding standard errors of estimates of θ_1. Note how the latter increase drastically as we reach the nonidentified case ($\theta_2 = 0$). The sample size is 3000 while the number of replications used to compute the standard errors is 400.

5. APPLICATION

Many studies have followed the seminal works by Brainard and Tobin (1968) and Tobin (1969) on firm investment and the so-called q theory. The theory simply states that a firm will invest if the ratio of the market values of the firm's capital stock to its replacement value, the Tobin's q, is larger than one. The intuition behind the q theory is that a firm should invest when it expects investment to be profitable based on an efficient asset markets' valuation of the firm (Grunfeld 1960). Despite its strong theoretical footing, the Tobin's q theory largely appeared to fail to explain both cross-section and time-series data, until Erickson and Whited (2000) observed that the q theory has, in fact, good explanatory power regarding investments once one allows for the presence of measurement error in q. Our application builds upon Erickson and Whited's notable result, by establishing that the applicability of their finding extends beyond the linear regression model they used. Allowing for nonlinear specifications is an important extension, for two reasons. First, there is clear

Table 3. Investment versus Tobin's q; 400 bootstrap replications used; sample size is 2948. "pctl" denotes percentiles

$m(x;\theta) = \theta_1 x + \theta_2 \ln(1+\theta_3 x)$					
	Point est.	Std. dev.	Median	5th pctl	95th pctl
Parameter			θ_1		
Naive ordinary least squares (OLS)	0.015	0.0015	0.015	0.012	0.017
Erickson and Whited (2000)	0.034	0.005	N/A	N/A	N/A
Ignoring meas. error	−0.021	0.0039	−0.021	−0.028	−0.015
Sieve MLE	−0.031	0.0058	−0.032	−0.042	−0.023
Parameter			θ_2		
Ignoring meas. error	0.55	0.055	0.54	0.47	0.64
Sieve MLE	0.81	0.081	0.81	0.69	0.96
Parameter			θ_3		
Ignoring meas. error	0.098	0.00034	0.098	0.097	0.098
Sieve MLE	0.099	0.00059	0.098	0.097	0.099

evidence of nonlinear response of firm investment to q (e.g., Barnett and Sakellaris 1998). Second, measurement error and nonlinearity (and the associated risk of model misspecification) often manifest themselves in similar ways (Chesher 1991), so that only a method robust to both aspects can disentangle them.

Erickson and Whited (2000) argued that instruments are difficult to find in this application and therefore employ a "higher-order moment" approach in a linear setting. The present article generalizes this approach, thus making it possible to consider q theory in a nonlinear setting with measurement errors. Adopting a nonlinear version of Erickson and Whited's specification, we describe the relationship between investment and Tobin's q as

$$\begin{aligned} y_i &= m(x_i^*, \theta) + z_i'\mu + \Delta y_i \\ x_i &= x_i^* + \varepsilon_i, \end{aligned} \tag{5}$$

where y_i is an investment divided by replacement value of the capital stock, x_i is the mismeasured version of Tobin's q (denoted by x_i^*), and $\Delta y_i, \varepsilon_i$ are disturbances. The variable z_i contains the covariates, specifically, $z_i = (1, z_{1i}, d_i, d_i \times z_{1i})^T$, where z_{1i} is a cash flow divided by replacement value of the capital stock and d_i is a 0–1 indicator of whether firm i is financially constrained. θ is the parameter of interest, while μ is the nuisance parameter associated with the covariates. As in Erickson and Whited (2000), the three variables Δy_i, ε_i, and (x_i^*, z_i) are assumed mutually independent with Δy_i ε_i having zero mean. In this generalized model, only y_i, x_i, and z_i are observed and the regression function $m(\cdot,\cdot)$ is assumed known up to the parameter θ to be estimated. Although our identification theory is fully nonparametric, a parametric estimation strategy is used here, given the size of the available sample. We use the specification

$$m(x^*;\theta) = \theta_1 x^* + \theta_2 \ln(1+\theta_3 x^*), \tag{6}$$

as it nests the linear case and provides flexibility regarding the curvature while maintaining monotonicity, an economically plausible characteristic (unlike a polynomial with the same number of parameters). Specification (6) is also in good agreement with a local nonparametric regression of y_i on x_i (the mismeasured Tobin q) based on the flexible specification $y_i = \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + z_i'\mu_y + \Delta y_i$, which is highly suggestive. Of course, using the mismeasured Tobin q in this preliminary specification analysis assumes that the measurement error is not sufficiently severe to completely alter the shape of the specification (in particular, the presence of a logarithmic tail).

We consider four estimators (see Table 3): the naive linear least squares, Erickson and Whited's "minimum distance GMM4" estimator, naive NLS, and the proposed sieve maximum likelihood estimation (MLE). For the sieve MLE, we use the Hermite polynomial-based sieve described in Section 4, with $k_{\Delta y} = 5, k_{\Delta x} = 6, k_x = 6$. These settings were found by gradually increasing the number of terms in the each series until we found a choice of truncation where the point estimates were the least sensitive to changes of ± 1 in the number of terms in the series. To carry out the search, we initially increased all truncations parameters simultaneously until a preliminary optimum was found. From this preliminary result, we then increased one parameter at a time to find the optimal parameter choice reported here. To save space and avoid confusion, we do not report here the alternative estimates obtained with suboptimal truncation choices.

Standard deviations of all the estimators, as well as the 5th, 50th, and 95th percentiles of their sampling distributions, were obtained using the bootstrap in the usual way: 400 bootstrap samples of a size equal to the original sample ($n = 2948$) were drawn (with replacement) from the original sample. Each bootstrap sample was used to obtain a point estimate and the resulting 400 point estimates were used to compute the relevant statistics (std. dev. and appropriate percentiles). We expect the bootstrap to be applicable in this context, since our semiparametric asymptotic theory establishes that our estimator is asymptotically equivalent to a sample average with finite variance under suitable regularity conditions. The validity of the bootstrap for nonlinear functionals that satisfy this condition has been established previously under quite general conditions (see, e.g., Bickel and Freedman 1981; Politis, Romano, and Wolf 1999, chap. 1.6). In fact, the use of the bootstrap for semiparametric Sieve Maximum Likelihood estimators has precedents in the literature (Chen and Ibrahim 2007).

To account for the presence of covariates, we condition the densities in the sieve on the covariate z_i. By independence, the sieves describing the distributions of the disturbances are unaffected by this conditioning. The distribution of x_i^* conditional

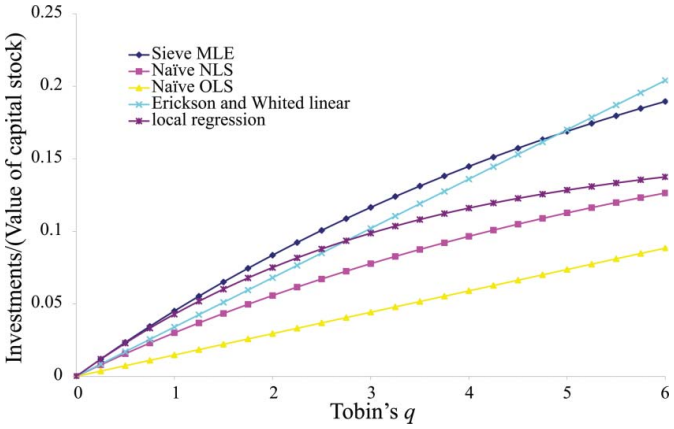

Figure 1. Investment (normalized by replacement value of capital stock) as a function of Tobin's q, as estimated by various techniques. The online version of this figure is in color.

on z_i is modeled as

$$f_{x^*,z}(x_i^*|z_i) = f_e(x_i^* - z_i'\lambda),$$

where λ is a parameter vector and $f_e(\cdot)$ is a univariate density represented by a Hermite polynomial-based sieve. The role of z_i as a Tobin's q shifter parallels its role as an investment shifter in (5). Note that the parameter λ can be straightforwardly estimated via a linear regression on the model

$$x_i = z_i'\lambda + e_i + \Delta x_i,$$

where e_i is disturbance (whose density is f_e) and $e_i + \Delta x_i$ has zero mean and is independent from z_i.

Figure 1 shows that both the naive linear and nonlinear regression considerably underestimate the magnitude of the effect of Tobin's q on investments, relative to the two measurement error-corrected estimators (Erickson and Whited's "minimum distance GMM4" linear estimator and our nonlinear sieve estimator described above).

The magnitude of the effect of Tobin's q according to our nonlinear estimator is broadly comparable to Erickson and Whited's result. Our analysis therefore corroborates their main result under more general conditions. This is an important robustness check, because a nonlinear regression that neglects measurement errors exhibits significant "saturation" at high values of Tobin's q, which seems to indicate that the explanatory power of Tobin's q is not as large as a linear model would suggest. However, our analysis in fact clearly shows that this saturation is not large enough to invalidate Erickson and Whited's result, once we correct for measurement error. Interestingly, the linear and nonlinear result differ more sharply in level before measurement error correction than after. This fact is consistent with the observation by Chesher (1991) that not properly accounting for measurement error can often lead to spurious nonlinearities.

Although accounting for nonlinearity turns out to not affect broad features of the model, such as the explanatory power of Tobin's q, it significantly affects some specific aspects. For instance, the true elasticity $d(\ln y)/d(\ln x^*)$ varies from 0.92 to 0.66 as x^* ranges from 1 to 5 (which roughly represents the range of the bulk of the data). This significant elasticity change cannot be captured with a linear model (whose elasticity remains 1 at all x^*, by construction).

6. CONCLUSION

This article answers the long-standing question of the identifiability of the nonparametric classical errors-in-variables model with a rather encouraging result, namely, that only a specific four-parameter parametric family of regression functions may exhibit lack of identifiability. We show that estimation can be accomplished via a nonparametric maximum likelihood approach and derive a suitable asymptotic theory. The effectiveness of the method is illustrated with a simulation study and an empirical application. We revisit Erickson and Whited's important finding that "Tobin's q" has good explanatory power regarding firm investments when one allows for the presence of measurement error in a linear model. We find that nonlinearities are important in this application but that Erickson and Whited's main conclusions are nevertheless robust to their presence.

APPENDIX: OUTLINE OF PROOF OF THEOREM 1

This Appendix presents a heuristic outline of the arguments leading to Theorem 1. Technical details can be found in the formal proof provided in Section A of the supplementary material.

The joint characteristic function of x and y, defined as $E[e^{i\xi x}e^{i\gamma y}]$, is known to convey the same information as the joint distribution of x and y. Under Model 1, we have

$$E[e^{i\xi x}e^{i\gamma y}] = E[e^{i\xi x^*}e^{i\gamma g(x^*)}e^{i\xi\Delta x}e^{i\gamma\Delta y}].$$

Assumption 1 then implies that

$$E[e^{i\xi x}e^{i\gamma y}] = E[e^{i\xi x^*}e^{i\gamma g(x^*)}]E[e^{i\xi\Delta x}]E[e^{i\gamma\Delta y}]. \tag{A.1}$$

To see when the model is not identified from the observed joint distribution of x and y, we seek an alternative observationally equivalent model (denoted with $\sim$ and satisfying the same assumptions as the original model) that also satisfies:

$$E[e^{i\xi x}e^{i\gamma y}] = E[e^{i\xi\widetilde{x}^*}e^{i\gamma\widetilde{g}(\widetilde{x}^*)}]E[e^{i\xi\Delta\widetilde{x}}]E[e^{i\gamma\Delta\widetilde{y}}]. \tag{A.2}$$

Equating (A.1) and (A.2) and rearranging yields

$$E[e^{i\xi x^*}e^{i\gamma g(x^*)}]\frac{E[e^{i\xi\Delta x}]}{E[e^{i\xi\Delta\widetilde{x}}]} = E[e^{i\xi\widetilde{x}^*}e^{i\gamma\widetilde{g}(\widetilde{x}^*)}]\frac{E[e^{i\gamma\Delta\widetilde{y}}]}{E[e^{i\gamma\Delta y}]},$$

where we have used Assumption 2. In the formal proof, we show that, under our assumptions and if $E[|\Delta x|] \geq E[|\Delta\widetilde{x}|]$, the ratios $E[e^{i\xi\Delta x}]/E[e^{i\xi\Delta\widetilde{x}}]$ and $E[e^{i\gamma\Delta\widetilde{y}}]/E[e^{i\gamma\Delta y}]$ form valid characteristic functions that we denote by $E[e^{i\xi\Delta\bar{x}}]$ and $E[e^{i\xi\Delta\bar{y}}]$, respectively, where $\Delta\bar{x}$ and $\Delta\bar{y}$ are new, implicitly defined, random variables. (The requirement $E[|\Delta x|] \geq E[|\Delta\widetilde{x}|]$ can always be met by permuting the two models if necessary.) The resulting equation

$$E[e^{i\xi x^*}e^{i\gamma g(x^*)}]E[e^{i\xi\Delta\bar{x}}] = E[e^{i\xi\widetilde{x}^*}e^{i\gamma\widetilde{g}(\widetilde{x}^*)}]E[e^{i\xi\Delta\bar{y}}]$$

effectively states the observational equivalence between two models, one with independent errors along "x" only:

$$\bar{y} = g(x^*)$$
$$\bar{x} = x^* + \Delta\bar{x},$$

and one with independent errors along "y" only:

$$\bar{y} = \widetilde{g}(\widetilde{x}^*) + \Delta\bar{y}$$
$$\bar{x} = \widetilde{x}^*$$

(note that, in general, $\bar{x}$ and $\bar{y}$ differ from the original variables x and y).

Next, we impose observational equivalence via the joint density of $\bar{x}$ and $\bar{y}$, expressed in terms of the two alternative models.

Denoting densities by f with appropriate subscripts, we have, by independence between $\Delta\bar{y}$ and $\bar{x}$, $\widetilde{f}_{\bar{x},\bar{y}}(\bar{x},\bar{y}) = \widetilde{f}_{\bar{x},\Delta\bar{y}}(\bar{x},\bar{y}-\widetilde{g}(\bar{x})) = \widetilde{f}_{\Delta\bar{y}|\bar{x}}(\bar{y}-\widetilde{g}(\bar{x})|\bar{x})\,\widetilde{f}_{\bar{x}}(\bar{x}) = \widetilde{f}_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))\,f_{\bar{x}}(\bar{x})$. Proceeding similarly for $f_{\bar{x},\bar{y}}(\bar{x},\bar{y})$, the equality $f_{\bar{x},\bar{y}}(\bar{x},\bar{y}) = \widetilde{f}_{\bar{x},\bar{y}}(\bar{x},\bar{y})$ can be written as

$$f_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))\,f_{\bar{y}}(\bar{y}) = \widetilde{f}_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))\,f_{\bar{x}}(\bar{x}), \tag{A.3}$$

where $h(\bar{y})$ denotes the inverse of $g(\bar{x})$. (In the formal proof, we establish that this inverse exists and that the above densities with respect to the Lebesgue measure exist, posses a sufficient number of derivatives, and are nonvanishing whenever needed. Otherwise, either the assumptions of the model are violated or the lack of regularity actually lead to identification of the model—for instance, a jump in $\widetilde{f}_{\Delta\bar{y}}$ or a point mass in the distribution of $\Delta\bar{y}$ immediately give away the shape of the regression function.) After rearranging and taking logs, we obtain

$$\ln\widetilde{f}_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x})) - \ln f_{\Delta\bar{x}}(\bar{x}-h(\bar{y})) = \ln f_{\bar{y}}(\bar{y}) - \ln f_{\bar{x}}(\bar{x}).$$

Computing the mixed derivative $\partial^2/\partial\bar{x}\partial\bar{y}$ cancels the right-hand-side and yields

$$-\widetilde{g}'(\bar{x})\,\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x})) + h'(\bar{y})\,F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y})) = 0$$

or

$$\frac{\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))}{F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))} = \frac{h'(\bar{y})}{\widetilde{g}'(\bar{x})}, \tag{A.4}$$

where $F \equiv \ln f$ with the corresponding subscripts and arguments while primes denote univariate derivatives. Taking logs again and noting that the right-hand side can again be canceled by applying a mixed derivative, we have

$$\frac{\partial^2}{\partial\bar{x}\partial\bar{y}}\ln\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x})) - \frac{\partial^2}{\partial\bar{x}\partial\bar{y}}\ln F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y})) = 0.$$

(The ln is defined for negative arguments by viewing it as a complex-valued function and selecting the same branch on each side of the equality.) After rearranging, we have

$$\frac{(\ln\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x})))''}{(\ln F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y})))''} = \frac{h'(\bar{y})}{\widetilde{g}'(\bar{x})}, \tag{A.5}$$

where the notation $(\ln\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x})))''$ stands for $(\ln\widetilde{F}''_{\Delta\bar{y}}(u))''|_{u=\bar{y}-\widetilde{g}(\bar{x})}$. Equating (A.4) and (A.5) and rearranging yields

$$\frac{(\ln\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x})))''}{\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))} = \frac{(\ln F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y})))''}{F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))}.$$

Since each side of the equality depends on a different argument ($\bar{y}-\widetilde{g}(\bar{x})$ vs. $\bar{x}-h(\bar{y})$) that can be set to arbitrarily different values, each side must be constant, unless the two models coincide (i.e., $h(\cdot)$ is the inverse of not only $g(\cdot)$ but also $\widetilde{g}(\cdot)$). This fact can be used to set up separate differential equations for $\widetilde{F}_{\Delta\bar{y}}$ and for $F_{\Delta\bar{x}}$ that can be solved analytically. The general solution to this differential equation leads to Case 2 in the theorem while Case 3 arises as a special case when some quantities happen to vanish. These solutions can then be used to recover $h(\cdot)$, $\widetilde{g}'(\cdot)$, $f_{\bar{x}}(\cdot)$, and $f_{\bar{y}}(\cdot)$ via (A.4) and (A.3) and provide the functional forms such that the model is not identified. Case 1 of the theorem covers the situation where the above construction is not possible and there consequently exists no pair of distinct models that are observationally equivalent, thus showing that the model is then identified.

SUPPLEMENTARY MATERIALS

The supplementary material provides a formal proof of Theorem 1 (Section A), an explicit example illustrating Case 2 of Theorem 1 (Section B), a formal asymptotic analysis of this estimator with suitable regularity conditions for consistency as well as root n consistency and asymptotic normality of a semiparametric estimator of the regression function (Section C) and additional simulations illustrating smoothing parameter selection (Section D.1) and performances at smaller sample size (Section D.2).

[Received October 2011. Revised October 2012.]

REFERENCES

Barnett, S. A., and Sakellaris, P. (1998), "Nonlinear Response of Firm Investment to Q: Testing a Model of Convex and Non-Convex Adjustment Costs," *Journal of Monetary Economics*, 42, 261–288. [183]

Bauer, D. J. (2005), "The Role of Nonlinear Factor-to-Indicator Relationships in Tests of Measurement Equivalence," *Psychological Methods*, 10, 305–316. [177]

Beckmann, C. F., and Smith, S. M. (2004), "Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging," *IEEE Transactions on Medical Imaging*, 23, 137–152. [177]

Bickel, P. J., and Freedman, D. A. (1981), "Some Asymptotic Theory for the Bootstrap," *The Annals of Statistics*, 9, 1196–1217. [183]

Bonhomme, S., and Robin, J.-M. (2009), "Consistent Noisy Independent Component Analysis," *Journal of Econometrics*, 149, 12–25. [177,178]

——— (2010), "Generalized Non-Parametric Deconvolution With an Application to Earnings Dynamics," *Review of Economic Studies*, 77, 491–533. [177,178]

Brainard, W. C., and Tobin, J. (1968), "Pitfalls in Financial Model Building," *American Economic Review*, 58, 99–122. [182]

Cardoso, J., and Souloumiac, A. (1993), "Blind Beamforming for Non-Gaussian Signals," *IEE Proceedings-F*, 140, 362–370. [177]

Carroll, R. J., Chen, X., and Hu, Y. (2010a), "Identification and Estimation of Nonlinear Models Using Two Samples With Nonclassical Measurement Errors," *Journal of Nonparametric Statistics*, 22, 379–399. [180]

——— (2010b), "Rejoinder: Identification and Estimation of Nonlinear Models Using Two Samples With Nonclassical Measurement Errors," *Journal of Nonparametric Statistics*, 22, 419–423. [180]

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006), *Measurement Error in Nonlinear Models*, New York: Chapman and Hall. [177,178]

Chen, Q. X., and Ibrahim, D. L. Z. J. G. (2007), "Sieve Maximum Likelihood Estimation for Regression Models With Covariates Missing at Random," *Journal of the American Statistical Association*, 102, 1309–1317. [183]

Chen, X., Hu, Y., and Lewbel, A. (2009), "Nonparametric Identification and Estimation of Nonclassical Errors-in-Variables Models Without Additional Information," *Statistica Sinica*, 19, 949–968. [178]

Chesher, A. (1991), "The Effect of Measurement Error," *Biometrika*, 78, 451–462. [183,184]

——— (1998), "Polynomial Regression With Normal Covariate Measurement Error," Discussion Paper 98/448, University of Bristol. [177]

Cragg, J. C. (1997), "Using Higher Moments to Estimate the Simple Errors-in-Variables Model," *Rand Journal of Economics*, 28, S71–S91. [177,178]

Dagenais, M. G., and Dagenais, D. L. (1997), "Higher Moment Estimators for Linear Regression Models With Errors in Variables," *Journal of Econometrics*, 76, 193–221. [177,178]

Erickson, T., and Whited, T. M. (2000), "Measurement Error and the Relationship Between Investment and 'q'," *Journal of Political Economy*, 108, 1027–1057. [177,178,182,183]

——— (2002), "Two-Step GMM Estimation of the Errors-in-Variables Model Using High-Order Moments," *Econometric Theory*, 18, 776–799. [177,178]

Fan, J. (1991), "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems," *The Annals of Statistics*, 19, 1257–1272. [178]

Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley. [177]

Geary, R. C. (1942), "Inherent Relations Between Random Variables," *Proceedings of the Royal Irish Academy*, 47A, 63–76. [177]

Grunfeld, Y. (1960), "The Determinants of Corporate Investment," in *The Demand for Durable Goods*, ed. A. Harberger, Chicago, IL: University Chicago Press, pp. 211–266. [182]

Hausman, J., Newey, W., Ichimura, H., and Powell, J. (1991), "Measurement Errors in Polynomial Regression Models," *Journal of Econometrics*, 50, 273–295. [177]

Hu, Y., and Ridder, G. (2012), "Estimation of Nonlinear Models with Measurement Error Using Marginal Information," *Journal of Applied Econometrics*, 27, 347–385. [177]

Hu, Y., and Schennach, S. M. (2008), "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 76, 195–216. [177,180,181]

Hyvärinen, A., and Oja, E. (1997), "A Fast Fixed Point Algorithm for Independent Component Analysis," *Neural Computation*, 9, 1483–1492. [177]
Ikeda, S., and Toyama, K. (2000), "Independent Component Analysis for Noisy Data—MEG Data Analysis," *Neural Networks*, 13, 1063–1074. [177]
Jutten, C., and Karhunen, J. (2003), "Advances in Nonlinear Blind Source Separation," *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pp. 245–256. [177]
Kapteyn, A., and Wansbeek, T. (1983), "Identification in the Linear Errors in Variables Model," *Econometrica*, 51, 1847–1849. [177]
Kendall, M. G., and Stuart, A. (1979), *The Advanced Theory of Statistics* (4th ed.), New York: Macmillan. [177,178]
Kenny, D. A., and Judd, C. M. (1984), "Estimating the Nonlinear and Interactive Effects of Latent Variables," *Psychological Bulletin*, 96, 201–210. [177]
Klepper, S., and Leamer, E. E. (1984), "Consistent Sets of Estimates for Regressions With Errors in All Variables," *Econometrica*, 52, 163–183. [177]
Lewbel, A. (1997), "Constructing Instruments for Regressions with Measurement Error When no Additional Data are Available, With an Application to Patents and R&D," *Econometrica*, 65, 1201–1213. [177,178]
Li, T., and Vuong, Q. (1998), "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," *Journal of Multivariate Analysis*, 65, 139–165. [177]
Lukacs, E. (1970), *Characteristic Functions* (2nd ed.), London: Griffin. [178]
Mahajan, A. (2006), "Identification and Estimation of Single Index Models With Misclassified Regressor," *Econometrica*, 74, 631–665. [180]
Newey, W. (2001), "Flexible Simulated Moment Estimation of Nonlinear Errors-in-Variables Models," *Review of Economics and Statistics*, 83, 616–627. [177,180]
Nocedal, J. (1980), "Updating Quasi-Newton Matrices With Limited Storage," *Mathematics of Computation*, 35, 773–782. [180]
Pal, M. (1980), "Consistent Moment Estimators of Regression-Coefficients in the Presence of Errors in Variables," *Journal of Econometrics*, 14, 349–364. [177,178]
Politis, D. N., Romano, J. P., and Wolf, M. (1999), *Subsampling*, New York: Springer. [183]
Reiersol, O. (1950), "Identifiability of a Linear Relation Between Variables Which are Subject to Error," *Econometrica*, 18, 375–389. [177,178,179]
Schennach, S. M. (2004a), "Estimation of Nonlinear Models With Measurement Error," *Econometrica*, 72, 33–75. [177,178]
——— (2004b), "Nonparametric Estimation in the Presence of Measurement Error," *Econometric Theory*, 20, 1046–1093. [177]
——— (2007), "Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models," *Econometrica*, 75, 201–239. [177]
——— (2009), "Instrumental Variable Treatment of the Nonparametric Berkson Measurement Error Model," Working Paper, University of Chicago. [180]
Shen, X. (1997), "On Methods of Sieves and Penalization," *The Annals of Statistics*, 25, 2555–2591. [180]
Tobin, J. (1969), "A General Equilibrium Approach to Monetary Theory," *Journal of Money Credit and Banking*, 1, 15–29. [182]
Wang, L., and Hsiao, C. (2003), "Identification and Estimation of Semiparametric Nonlinear Errors-in-Variables Models," *Journal of Econometrics*, 165, 30–44. [177]
Yalcin, I., and Amemiya, Y. (2001), "Nonlinear Factor Analysis as a Statistical Method," *Statistical Science*, 16, 275–294. [177]