cemmap
centre for microdata methods and practice

Nonparametric identification and semiparametric estimation of classical measurement error models without side information

S. M. Schennach
Yingyao Hu

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP40/12

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Nonparametric identification and semiparametric estimation of classical measurement error models without side information*

S. M. Schennach
Department of Economics
Box B
Brown University
Providence, RI 02912
smschenn@brown.edu

Yingyao Hu
Department of Economics
Johns Hopkins University
3400 N. Charles Street
Baltimore, MD 21218
yhu@jhu.edu

November 26, 2012

*The authors would like thank Daniel Wilhelm and Xavier d'Haultfoeuille for useful comments and Jiaxiong Yao for excellent research assistance. Susanne M. Schennach acknowledges support from NSF grants SES-0752699 and SES-1061263/1156347.

Abstract

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 data sets where it was previously considered impossible (due to the lack of additional information regarding the measurement error). We suggest an estimator based on non/semi-parametric 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.

Keywords: Measurement error, errors-in-variables, higher-order moments, nonparametric, identification.

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, Ruppert, Stefanski, and Crainiceanu (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*

$$\begin{aligned} y &= g(x^*) + \Delta y \\ x &= x^* + \Delta x \end{aligned}$$

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 towards 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, Newey, Ichimura, and Powell (1991), Newey (2001), Li and Vuong (1998), Wang and Hsiao (2003), Schennach (2004a), Schennach (2004b), Schennach (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 so is not. Under fairly simple and natural regularity conditions, a specification of the form $g(x^*) = a + b \ln\left(e^{cx^*} + d\right)$ 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 in order 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 data sets 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 errors 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 above). 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 (such as Reiersol (1950), Kendall and Stuart (1979), Pal (1980), Cragg (1997), Lewbel (1997), Erickson and Whited (2002), Dagenais and Dagenais (1997), Erickson and Whited (2000), Bonhomme and Robin (2009), Bonhomme and Robin (2010)). Moreover, even in the measurement error literature that exploits side information, independence assumptions are extremely common (see, for instance, the monograph by Carroll, Ruppert,

Stefanski, and Crainiceanu (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 unobservables 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 paper.) 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 *$E\left[e^{i\xi\Delta x}\right]$ and $E\left[e^{i\gamma\Delta y}\right]$ 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) and 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) $E\left[e^{i\xi x^*}\right] \neq 0$ for all ξ in a dense subset of $\mathbb{R}$ and (ii) $E\left[e^{i\gamma g(x^*)}\right] \neq 0$ 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 $(\Delta x,\ \Delta y)$. Also note that both Assumptions 2 and 3 are implied by the Assumption that $E\left[e^{i\xi x}\right] \neq 0$ and $E\left[e^{i\gamma y}\right] \neq 0$ 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 a F* factor *if r can be written as the sum of two independent random variables (which may be degenerate), one of which has the distribution F. (This is related to the concept of a decomposable characteristic functions, see Lukacs (1970), Section 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*

$$g(x^*) = a + b \ln\left(e^{cx^*} + d\right) \tag{1}$$

*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*

$$f_{x^*}(x^*) = A \exp\left(-Be^{Cx^*} + CDx^*\right)\left(e^{Cx^*} + E\right)^{-F} \tag{2}$$

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 $f_u(u) = K_1 \exp(K_2 \exp(K_3 u) + K_4 u)$ 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 above 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 towards 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 4 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 model (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 establish 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:

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

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 analogue 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 an i.i.d. sample $(y_i, x_i)_{i=1}^n$, we employ a sieve maximum likelihood estimator (Shen (1997)) based on the following equation:

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

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, (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 (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 surprizing that our likelihood function turns out to be rather well-behaved, thus enabling us to employ a standard numerical optimization routine to maximize it: a L-BFGS quasi-Newton algorithm (Nocedal (1980)).

In our examples below, we consider the estimation of a parametric regression model, i.e. Model 1 with $g\left(x^*\right) = m\left(x^*;\theta\right)$, where the function $m\left(x^*;\theta\right)$ is known up to a parameter vector θ. However, the densities of the unobserved variables $\Delta x, \Delta y$ 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 $\Delta x, \Delta y$ 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 focussing 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 paper. 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 (rel-

ative 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:

$$\begin{aligned} y &= m(x^*;\theta) + \Delta y \\ x &= x^* + \Delta x. \end{aligned}$$

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 $H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x)$. The sieve expression of the nonparametric densities are

$$f_1(\Delta y) = \left[\sum_{i=0}^{k_{\Delta y}} \beta_i^{\Delta y} p_i(\Delta y)\right]^2, \quad f_2(\Delta x) = \left[\sum_{i=0}^{k_{\Delta x}} \beta_i^{\Delta x} p_i(\Delta x)\right]^2, \quad f_3(x^*) = \left[\sum_{i=0}^{k_x} \beta_i^{x} p_i(x^*)\right]^2.$$

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 $\sum_{i=0}^{k_{\Delta y}} \left[\beta_i^{\Delta y}\right]^2 = 1$, 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^*:

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

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

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

This estimator should give us the largest bias. Finally, we consider the sieve-based instrumental variable estimator of Hu and Schennach (2008), denoted $\widehat{\theta}_{HS}$, which is consistent in the presence of measurement error, but requires the availability of an instrument. To ensure a meaningful comparison, we specialized $\widehat{\theta}_{HS}$ to the case where all the error terms and x^* are mutually independent (as assumed for $\widehat{\theta}_{sieve}$), for otherwise, allowing for general form of heteroskedasticity in $\widehat{\theta}_{HS}$ would have caused an

efficiency penalty relative to $\widehat{\theta}_{sieve}$. We would expect $\widehat{\theta}_{HS}$ to have properties roughly similar to $\widehat{\theta}_{sieve}$, 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

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

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 $\widehat{\theta}_{sieve}$, $\widehat{\theta}_{HS}$, $\widehat{\theta}_{nls}$ and $\widehat{\theta}_{nnls}$. 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 $\widehat{\theta}_{sieve}$, $\widehat{\theta}_{nls}$ and $\widehat{\theta}_{HS}$ are small compared with

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 $k_{\Delta y} = 5$, $k_{\Delta x} = 5$, $k_x = 6$.

Case 1: $m(x;\theta) = \theta_1 x + \theta_2 e^x$						
Parameter (=true value)	$\theta_1 = 1$			$\theta_2 = 1$		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
Ignoring 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 &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: $m(x;\theta) = \theta_1 x + \theta_2 x^2$						
Parameter (=true value)	$\theta_1 = 1$			$\theta_2 = 1$		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
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 &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: $m(x;\theta) = \theta_1 x + \theta_2/(1+x^2)$						
Parameter (=true value)	$\theta_1 = 1$			$\theta_2 = 1$		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
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 &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: $m(x;\theta) = \left(x^2 + \theta_1\right)(x + \theta_2)$						
Parameter (=true value)	$\theta_1 = 1$			$\theta_2 = 1$		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
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 &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: $m(x;\theta) = \ln\left(1 + \theta_1 x + \theta_2 x^2\right)$						
Parameter (=true value)	$\theta_1 = 1$			$\theta_2 = 1$		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
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 &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: $m(x;\theta) = \theta_1 x + \theta_2 \ln\left(1 + x^2\right)$						
Parameter (=true value)	$\theta_1 = 1$			$\theta_2 = 1$		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
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 &Schennach (2008)	0.988	0.061	0.062	0.791	0.171	0.270
Sieve MLE	0.915	0.05918	0.104	0.979	0.054	0.058

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

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

$\widehat{\theta}_{nnls}$, because they are consistent. The variances of $\widehat{\theta}_{sieve}$ and $\widehat{\theta}_{HS}$ should be the largest of the four due to the nonparametric approximation. Nevertheless, the sieve estimator $\widehat{\theta}_{sieve}$ is preferable over the naive estimator in terms of mean squared errors. The comparison between $\widehat{\theta}_{sieve}$ and $\widehat{\theta}_{HS}$ is instructive, as it reveals that, although $\widehat{\theta}_{sieve}$ is generally less efficient than $\widehat{\theta}_{HS}$ (as expected), it is often able to approach the RMSE of $\widehat{\theta}_{HS}$, 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).

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.

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 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) argue that instruments are difficult to find in this application and therefore employ a "higher-order moment" approach in a linear setting. The present paper 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 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 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 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 nonlinear least squares and the proposed sieve 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 the 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 (std. dev.) of all the estimators, as well as the 5^{th}, 50^{th} and 95^{th} 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., Politis, Romano, and Wolf (1999), Chapter 1.6 and Bickel and Freedman (1981)). 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

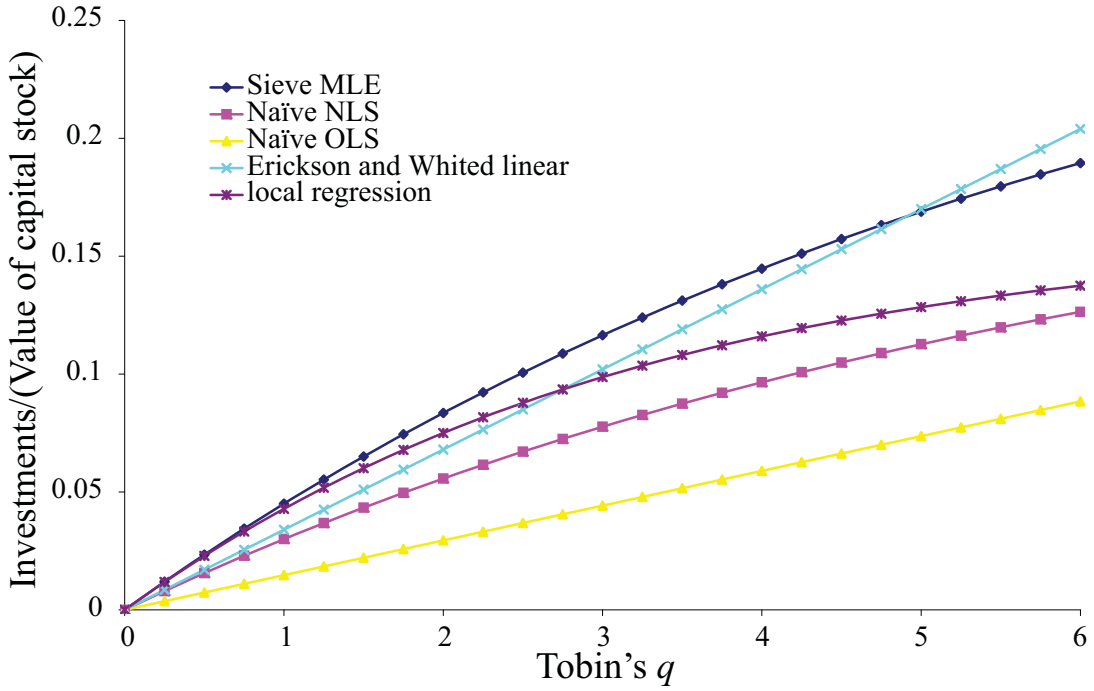

Figure 1: Investment (normalized by replacement value of capital stock) as a function of Tobin's q, as estimated by various techniques.

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 on z_i is modeled as

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

where λ is a parameter vector and $f_e\left(\cdot\right)$ 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).

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

$m(x;\theta) = \theta_1 x + \theta_2 \ln(1+\theta_3 x)$					
Parameter			θ_1		
	point est.	std. dev.	median	5^{th} pctl	95^{th} pctl
Naive OLS	0.015	0.0015	0.015	0.012	0.017
Erickson & Whited	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
			θ_2		
	point est.	std. dev.	median	5^{th} pctl	95^{th} pctl
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
			θ_3		
	point est.	std. dev.	median	5^{th} pctl	95^{th} pctl
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

6 Conclusion

This paper 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 4-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 $E\left[e^{i\xi x}e^{i\gamma y}\right]$, is known to convey the same information as the joint distribution of x and y. Under Model 1, we have

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

Assumption 1 then implies that

$$E\left[e^{i\xi x}e^{i\gamma y}\right] = E\left[e^{i\xi x^*}e^{i\gamma g(x^*)}\right]E\left[e^{i\xi\Delta x}\right]E\left[e^{i\gamma\Delta y}\right]. \tag{7}$$

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:

$$E\left[e^{i\xi x}e^{i\gamma y}\right] = E\left[e^{i\xi \widetilde{x}^*}e^{i\gamma \widetilde{g}(\widetilde{x}^*)}\right]E\left[e^{i\xi\Delta \widetilde{x}}\right]E\left[e^{i\gamma\Delta \widetilde{y}}\right]. \tag{8}$$

Equating (7) and (8) and rearranging yields

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

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

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

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

$$\begin{aligned} \bar{y} &= g(x^*) \\ \bar{x} &= x^* + \Delta\bar{x}. \end{aligned}$$

and one with independent errors along "y" only:

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

(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:

$$f_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))\,f_{\bar{y}}(\bar{y}) = \widetilde{f}_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))\,f_{\bar{x}}(\bar{x}) \tag{9}$$

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

$$\frac{\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))}{F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))}=\frac{h'(\bar{y})}{\widetilde{g}'(\bar{x})} \tag{10}$$

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 cancelled 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

$$\frac{\left(\ln\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))\right)''}{\left(\ln F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))\right)''}=\frac{h'(\bar{y})}{\widetilde{g}'(\bar{x})}, \tag{11}$$

where the notation $\left(\ln\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))\right)''$ stands for $\left(\ln\widetilde{F}''_{\Delta\bar{y}}(u)\right)''|_{u=\bar{y}-\widetilde{g}(\bar{x})}$. Equating (10) and (11) and rearranging yields

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

Since each side of the equality depend on a different argument ($\bar{y}-\widetilde{g}(\bar{x})$ versus $\bar{x}-h(\bar{y})$) 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

$f_{\tilde{y}}(\cdot)$ via (10) and (9) 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.

References

Barnett, S. A., and P. Sakellaris (1998): "Nonlinear Response of Firm Investment to Q: Testing a Model of Convex and Non-convex Adjustment Costs," *J. Monetary Econ.*, 42, 261–88.

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

Beckmann, C. F., and S. M. Smith (2004): "Probabilistic independent component analysis for functional magnetic resonance imaging," *IEEE Transactions on Medical Imaging*, 23, 137–152.

Bickel, P. J., and D. A. Freedman (1981): "Some asymptotic theory for the bootstrap," *Annals of Statistics*, 9, 1196–1217.

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

——— (2010): "Generalized Non-Parametric Deconvolution with an Application to Earnings Dynamics," *Review of Economic Studies*, 77, 491–533.

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

Cardoso, J., and A. Souloumiac (1993): "Blind beamforming for non-gaussian signals," *IEE Proceedings-F*, 140, 362–370.

Carroll, R. J., X. Chen, and Y. Hu (2010a): "Identification and estimation of nonlinear models using two samples with nonclassical measurement errors," *Journal of Nonparametric Statistics*, 22, 379–399.

——— (2010b): "Rejoiner: Identification and estimation of nonlinear models using two samples with nonclassical measurement errors," *Journal of Nonparametric Statistics*, 22, 419–423.

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

Chen, Q. X., and D. L. Z. J. G. Ibrahim (2007): "Sieve maximum likelihood estimation for regression models with covariates missing at random," *Journal of the American Statistical Association*, 102, 1309–1317.

Chen, X., Y. Hu, and A. Lewbel (2009): "Nonparametric identification and estimation of nonclassical errors-in-variables models without additional information," *Statistica Sinica*, 19, 949–968.

Chesher, A. (1991): "The Effect of Measurement Error," *Biometrika*, 78, 451.

——— (1998): "Polynomial Regression with Normal Covariate Measurement Error," Discussion Paper 98/448, University of Bristol.

Cragg, J. C. (1997): "Using higher moments to estimate the simple errors-in-variables model," *Rand Journal of Economics*, 28, S71–S91.

DAGENAIS, M. G., AND D. L. DAGENAIS (1997): "Higher Moment Estimators for Linear Regression Models with Errors in Variables," *Journal of Econometrics*, 76, 193–221.

ERICKSON, T., AND T. M. WHITED (2000): "Measurement Error and the Relationship between Investment and "q"," *Journal of Political Economy*, 108, 1027–1057.

——— (2002): "Two-step GMM estimation of the errors-in-variables model using high-order moments," *Econometric Theory*, 18, 776–799.

FAN, J. (1991): "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems," *Annals of Statistics*, 19(3), 1257–1272.

FULLER, W. A. (1987): *Measurement Error models*. Wiley, New York.

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

GRUNFELD, Y. (1960): "The Determinants of Corporate Investment," in *The Demand for Durable Goods*, ed. by A. Harberger. Univ. Chicago Press.

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

HU, Y., AND G. RIDDER (2012): "Estimation of Nonlinear Models with Measurement Error Using Marginal Information," *Journal of Applied Econometrics*, 27, 347–85.

HU, Y., AND S. M. SCHENNACH (2008): "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 76, 195–216.

HYVÄRINEN, A., AND E. OJA (1997): "A fast fixed point algorithm for independent component analysis," *Neural Computation*, 9, 1483–1492.

Ikeda, S., and K. Toyama (2000): "Independent component analysis for noisy data — MEG data analysis," *Neural Networks*, 13, 1063–1074.

Jutten, C., and J. Karhunen (2003): "Advances in Nonlinear Blind Source Separation," *Proc. of the 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pp. 245–256.

Kapteyn, A., and T. Wansbeek (1983): "Identification in the Linear Errors in Variables Model," *Econometrica*, 51, 1847–1849.

Kendall, M. G., and A. Stuart (1979): *The Advanced Theory of Statistics*. Macmillan, New York, 4th edition edn.

Kenny, D. A., and C. M. Judd (1984): "Estimating the nonlinear and interactive effects of latent variables," *Psychological Bulletin*, 96, 201–210.

Klepper, S., and E. E. Leamer (1984): "Consistent Sets of Estimates for Regressions with Errors in all Variables," *Econometrica*, 52, 163–183.

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(5), 1201–1213.

Li, T., and Q. Vuong (1998): "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," *Journal of Multivariate Analysis*, 65, 139–165.

Lukacs, E. (1970): *Characteristic Functions*. Griffin, London, second edn.

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

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

NOCEDAL, J. (1980): "Updating Quasi-Newton Matrices With Limited Storage," *Math. Comp.*, 35, 773–782.

PAL, M. (1980): "Consistent moment estimators of regression-coefficients in the presence of errors in variables," *Journal of Econometrics*, 14, 349–364.

POLITIS, D. N., J. P. ROMANO, AND M. WOLF (1999): *Subsampling*. Springer, New York.

REIERSOL, O. (1950): "Identifiability of a Linear Relation between Variables Which Are Subject to Error," *Econometrica*, 18, 375–389.

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.

——— (2009): "Instrumental variable treatment of the nonparametric Berkson measurement error model," Working Paper, University of Chicago.

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

TOBIN, J. (1969): "A General Equilibrium Approach to Monetary Theory," *J. Money Credit and Banking*, 1, 15–29.

WANG, L., AND C. HSIAO (2003): "Identification and Estimation of Semiparametric Nonlinear Errors-in-Variables Models," Working Paper, University of Southern California.

YALCIN, I., AND Y. AMEMIYA (2001): "Nonlinear Factor Analysis as a Statistical Method," *Statistical Science*, 16, 275–294.

Supplementary Material to: "Nonparametric identification and semiparametric estimation of classical measurement error models without side information"

S. M. Schennach
Department of Economics
Box B
Brown University
Providence, RI 02912
smschenn@brown.edu

Yingyao Hu
Department of Economics
Johns Hopkins University
3400 N. Charles Street
Baltimore, MD 21218
yhu@jhu.edu

November 26, 2012

A Proofs

A.1 Preliminaries

Throughout this Supplementary Material, references prefixed by a letter (e.g. "Assumption C.1") point to items in the Supplementary material while numbered references (e.g. "Assumption 1") point to the main text.

Definition A.1 *Let $\mathcal{S}_u$ denote the support of the random variable u and let $f_u(u)$ denote its density (and similarly for the multivariate case).*

Remark A.1 *Since two Lebesgue densities can differ on a set of null Lebesgue measure and still represent the same distribution, we use the following convention to prevent countless "almost everywhere" qualifications from obscuring our main argument.*

Definition A.2 *Given a random variable V taking value in $\mathbb{R}^q$ with probability measure μ_V admitting an absolutely continuous density with respect to the Lebesgue measure, we* define *the density $f_V(v)$ of V at the point v as*

$$f_V(v) = \begin{cases} \frac{\partial^q F_V(v)}{\partial v_1 \cdots \partial v_q} \text{ if it exists at } v \\ 0 \text{ otherwise} \end{cases} \tag{A.1}$$

where $F_V(v) = \mu_V((-\infty, v_1] \times \cdots \times (-\infty, v_q])$. *We adopt the same convention for measures such that* $\mu_V(\mathbb{R}^q) < 1$.

Remark A.2 *Note that* $\partial^q F_V(v) / \partial v_1 \cdots \partial v_q$ *exists almost everywhere, since* μ_V *admits a density with respect to the Lebesgue measure by assumption. The second of case of (A.1) applies, for instance, when* $F_V(v)$ *exhibits a kink at a point, so that its derivative does not have a unique value there. Setting* $f_V(v)$ *to zero on a set of null Lebesgue measure has no effect on the corresponding probability distribution. This definition merely selects one of the many possible valid Lebesgue densities defining the same distribution.*

We then define a few convenient concepts related to the divisibility (or decomposability) of characteristic functions (see Lukacs (1970), Section 5.1). "Characteristic functions" will be abreviated "c.f." hereafter.

Definition A.3 *Let* r, s *be random variables. Then,* s *is called a factor of* r *if* r *can be expressed as* $r = s + t$ *for some random variable* t, *with* s *independent from* t. *Any number of these random variables may be constant. (This definition is analogous to Definition* 1, *but without referring to distributions.) For convenience, we also use the term "factor" for the corresponding distributions or corresponding characteristic functions (c.f.), e.g., the distribution the distribution* F_s *is a factor of the distribution* F_r *or the c.f.* $E\left[e^{i\xi s}\right]$ *is a factor of the c.f.* $E\left[e^{i\xi r}\right]$.

Definition A.4 *A random variable* r *is* decomposable *if* r *can be written as the sum of two* ***nonconstant*** *independent random variables.*

Remark A.3 *Note that this definition differs from the notion of* factor *introduced in Definitions 1 and A.3, where constant random variables were allowed in the sum.*

The identification proof will consider an alternative model that is observationally equivalent to Model 1 (in the main text), defined as follows.

Model A.2 *Let* $y, x, \widetilde{x}^*, \Delta\widetilde{x}, \Delta\widetilde{y}$ *be scalar real-valued random variables related through*

$$\begin{aligned} y &= \widetilde{g}(\widetilde{x}^*) + \Delta\widetilde{y} \\ x &= \widetilde{x}^* + \Delta\widetilde{x} \end{aligned}$$

where only x *and* y *are observed (and are equal to the corresponding variables in Model 1) while* $\widetilde{x}^*, \Delta\widetilde{x}, \Delta\widetilde{y}$ *are unobserved (and may differ from the corresponding variables* $x^*, \Delta x, \Delta y$ *in Model 1). However,* $\widetilde{g}(\cdot), \widetilde{x}^*, \Delta\widetilde{x}, \Delta\widetilde{y}$ *satisfy the same Assumptions 1-6 as the corresponding entities* $g(\cdot), x^*, \Delta x, \Delta y$ *in Model 1. (We will invoke Assumptions 1-6 for Model A.2 whenever needed without always explicitly mentioning that the variables have to be replaced by their tilded counterparts in the assumptions.)*

Definition A.5 *Two models are* observationally equivalent *if they generate the same joint distribution of the observable variables.*

Definition A.6 *Two models are* distinct *if their regression functions differ or if the joint distributions of their corresponding unobserved variables differ.*[1]

We first establish a few basic results that will prove useful. We first reduce the identification problem to a simpler but equivalent problem involving only one error term. Consider the following two models:

Model A.3 *Let $\bar{x}, \bar{y}, x^*, \Delta\bar{x}$ be scalar real-valued random variables such that*

$$\begin{aligned} \bar{y} &= g(x^*) \\ \bar{x} &= x^* + \Delta\bar{x} \end{aligned}$$

where $\bar{x}$ and $\bar{y}$ are observable (and may differ from x, y in Model 1 and A.2), where the unobservable x^* and $g(x^*)$ are as in Model 1, and $\Delta\bar{x}$ is independent from x^*, $E[\Delta\bar{x}] = 0$ and the distribution of $\Delta\bar{x}$ is a factor of the distribution of Δx in Model 1.

Model A.4 *Let $\bar{x}, \bar{y}, \widetilde{x}^*, \Delta\bar{y}$ be scalar real-valued random variables such that*

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

where the observables $\bar{x}$ and $\bar{y}$ are as in Model A.3, where the unobservable $\widetilde{x}^*$ and $\widetilde{g}(\widetilde{x}^*)$ are as in Model A.2 and where $\Delta\bar{y}$ is independent from $\widetilde{x}^*$, $E[\Delta\bar{y}] = 0$ and $\Delta\bar{y}$ is a factor of Δy in Model A.2.

Lemma A.1 *If the random variables $\Delta\widetilde{y}$, Δy and $\Delta\bar{y}$ are related trhough $\Delta\widetilde{y} = \Delta y + \Delta\bar{y}$ with $\Delta\bar{y}$ independent from Δy, then $\inf_{c\in\mathbb{R}} E[|\Delta\widetilde{y} - c|] \geq \inf_{c\in\mathbb{R}} E[|\Delta y - c|]$ (assuming the requisite expectations exist).*

Proof. Using, in turn, (i) iterated expectations, (ii) properties of the infimum, (iii) independence of Δy from $\Delta\bar{y}$ and (iv) the fact that an expectation has no effect on a constant, we have:

$$\begin{aligned} \inf_{c\in\mathbb{R}} E[|\Delta\widetilde{y} - c|] &= \inf_{c\in\mathbb{R}} E[|\Delta y + \Delta\bar{y} - c|] = \inf_{c\in\mathbb{R}} E[E[|\Delta y + \Delta\bar{y} - c| \mid \Delta\bar{y}]] \\ &\geq E\left[\inf_{c\in\mathbb{R}} E[|\Delta y + \Delta\bar{y} - c| \mid \Delta\bar{y}]\right] = E\left[\inf_{c\in\mathbb{R}} E[|\Delta y - c| \mid \Delta\bar{y}]\right] \\ &= E\left[\inf_{c\in\mathbb{R}} E[|\Delta y - c|]\right] = \inf_{c\in\mathbb{R}} E[|\Delta y - c|]. \end{aligned}$$

[1] Note that, since the regression function is assumed continuous, if two regression functions differ at a point, they also differ on an interval. Equality of two distributions means that they assign the same probability to measurable sets.

■

Lemma A.2 *If* $E\left[e^{i\xi x}e^{i\gamma y}\right]/E\left[e^{i\gamma\Delta y}\right]$ *(where* $E\left[e^{i\gamma\Delta y}\right]\neq 0$*) is a c.f. then* $E\left[e^{i\gamma y}|x=x_0\right]/E\left[e^{i\gamma\Delta y}\right]$ *is a c.f. for any fixed* x_0 *in the support of* x *(except perhaps for* x_0 *in some null set under the probability measure of* x*).*

Proof. If $E\left[e^{i\xi x}e^{i\gamma y}\right]/E\left[e^{i\gamma\Delta y}\right]$ is a c.f., Theorem 3.3.1 in Lukacs (1970) implies that there exists a random variable y^* such that $y=y^*+\Delta y$ with Δy independent from (y^*,x) and such that $E\left[e^{i\xi x}e^{i\gamma y}\right]/E\left[e^{i\gamma\Delta y}\right]=E\left[e^{i\xi x}e^{i\gamma y^*}\right]$. The associated joint distribution of (x,y^*), admits a conditional distribution of y^* given $x=x_0$ (except perhaps for x_0 in some null set under the probability measure of x). The c.f. of this conditional distribution is $E\left[e^{i\gamma y^*}|x=x_0\right]=E\left[e^{i\gamma y^*}|x=x_0\right]E\left[e^{i\gamma\Delta y}\right]/E\left[e^{i\gamma\Delta y}\right]=E\left[e^{i\gamma(y^*+\Delta y)}|x=x_0\right]/E\left[e^{i\gamma\Delta y}\right]=E\left[e^{i\gamma y}|x=x_0\right]/E\left[e^{i\gamma\Delta y}\right]$ (again by Theorem 3.3.1 in Lukacs (1970), applied to a conditional distribution). ■

Lemma A.3 *Under Assumptions 1, 2, 5 and 6, there exist two distinct observationally equivalent Models 1 and A.2 iff there exist two distinct observationally equivalent models of the form of Models A.3 and A.4.*

Proof. The joint c.f. of x and y, defined as $E\left[e^{i\xi x}e^{i\gamma y}\right]$, conveys the same information as the joint distribution of x and y (by Theorem 3.1.1 in Lukacs (1970)). Under Model 1,

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

Assumption 1 then implies (by Theorem 3.3.1 in Lukacs (1970)) that

$$E\left[e^{i\xi x}e^{i\gamma y}\right]=E\left[e^{i\xi x^*}e^{i\gamma g(x^*)}\right]E\left[e^{i\xi\Delta x}\right]E\left[e^{i\gamma\Delta y}\right]. \tag{A.3}$$

We seek an alternative observationally equivalent model (Model A.2, denoted with $\sim$) also satisfying:

$$E\left[e^{i\xi x}e^{i\gamma y}\right]=E\left[e^{i\xi\widetilde{x}^*}e^{i\gamma\widetilde{g}(\widetilde{x}^*)}\right]E\left[e^{i\xi\Delta\widetilde{x}}\right]E\left[e^{i\gamma\Delta\widetilde{y}}\right]. \tag{A.4}$$

Equating (A.3) and (A.4) yields

$$E\left[e^{i\xi x^*}e^{i\gamma g(x^*)}\right]E\left[e^{i\xi\Delta x}\right]E\left[e^{i\gamma\Delta y}\right]=E\left[e^{i\xi\widetilde{x}^*}e^{i\gamma\widetilde{g}(\widetilde{x}^*)}\right]E\left[e^{i\xi\Delta\widetilde{x}}\right]E\left[e^{i\gamma\Delta\widetilde{y}}\right]. \tag{A.5}$$

Let us assume that $\inf_{c\in\mathbb{R}}E\left[|\Delta\widetilde{y}-c|\right]\geq\inf_{c\in\mathbb{R}}E\left[|\Delta y-c|\right]$ (these expectations exist by the second part of Assumption 1). This is without loss of generality: if this turns out to be untrue, we will merely arrive at a contradiction in the following derivation, forcing us to assume the converse or, equivalently, to permute the role of the observationally equivalent Models 1 and A.2 (as well as Models A.3 and A.4).

Dividing $E\left[e^{i\xi x}e^{i\gamma y}\right]$ by $E\left[e^{i\xi\Delta\widetilde{x}}\right]$ (which is nonvanishing by assumption 2), we obtain (by Equation (A.4)):

$$\frac{E\left[e^{i\xi x}e^{i\gamma y}\right]}{E\left[e^{i\gamma\Delta\widetilde{x}}\right]}=\frac{E\left[e^{i\xi\widetilde{x}^{*}}e^{i\gamma\widetilde{g}(\widetilde{x}^{*})}\right]E\left[e^{i\xi\Delta\widetilde{x}}\right]E\left[e^{i\gamma\Delta\widetilde{y}}\right]}{E\left[e^{i\xi\Delta\widetilde{x}}\right]}=E\left[e^{i\xi\widetilde{x}^{*}}e^{i\gamma\widetilde{g}(\widetilde{x}^{*})}\right]E\left[e^{i\gamma\Delta\widetilde{y}}\right]\quad\text{(A.6)}$$

where the right hand side is the c.f. of the vector $(\widetilde{x}^{*},\widetilde{g}\left(\widetilde{x}^{*}\right)+\Delta\widetilde{y})$ with $\Delta\widetilde{y}$ independent from $\widetilde{x}^{*}$ (and $\widetilde{g}\left(\widetilde{x}^{*}\right)$). Hence $E\left[e^{i\xi\Delta\widetilde{x}}\right]$ is a factor of $E\left[e^{i\xi x}e^{i\gamma y}\right]$ (see Lukacs (1970), Section 5.1). Next, dividing $E\left[e^{i\xi x}e^{i\gamma y}\right]$ instead by $E\left[e^{i\gamma\Delta y}\right]$ (which is nonvanishing by assumption 2), we obtain (by Equation (A.3)):

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

which is the c.f. of the vector $(x^{*}+\Delta x,g\left(x^{*}\right))$ with Δx independent of x^{*} (and $g\left(x^{*}\right)$). Hence $E\left[e^{i\gamma\Delta y}\right]$ is also a factor of $E\left[e^{i\xi x}e^{i\gamma y}\right]$. Since the two factors $E\left[e^{i\xi\Delta\widetilde{x}}\right]$ and $E\left[e^{i\gamma\Delta y}\right]$ involve a different variable (ξ versus γ), they cannot have a factor in common (except for a point mass) and we can conclude that $E\left[e^{i\xi\Delta\widetilde{x}}\right]E\left[e^{i\gamma\Delta y}\right]$ is also a factor of $E\left[e^{i\xi x}e^{i\gamma y}\right]$. It follows that we can divide each side of Equation (A.5) by $E\left[e^{i\xi\Delta\widetilde{x}}\right]E\left[e^{i\gamma\Delta y}\right]$ to obtain:

$$\frac{E\left[e^{i\xi x^{*}}e^{i\gamma g(x^{*})}\right]E\left[e^{i\xi\Delta x}\right]}{E\left[e^{i\xi\Delta\widetilde{x}}\right]}=\frac{E\left[e^{i\xi\widetilde{x}^{*}}e^{i\gamma\widetilde{g}(\widetilde{x}^{*})}\right]E\left[e^{i\gamma\Delta\widetilde{y}}\right]}{E\left[e^{i\gamma\Delta y}\right]}\qquad\text{(A.7)}$$

where each side of the equation is a valid c.f.

Focusing on the right-hand side of (A.7) first, we now show that $E\left[e^{i\gamma\Delta y}\right]$ is a factor of $E\left[e^{i\gamma\Delta\widetilde{y}}\right]$. Note that $E\left[e^{i\xi\widetilde{x}^{*}}e^{i\gamma\widetilde{g}(\widetilde{x}^{*})}\right]E\left[e^{i\gamma\Delta\widetilde{y}}\right]=E\left[e^{i\xi\widetilde{x}^{*}}e^{i\gamma(\widetilde{g}(\widetilde{x}^{*})+\Delta\widetilde{y})}\right]=E\left[e^{i\xi\widetilde{x}^{*}}e^{i\gamma y}\right]$. By Lemma A.2, the fact that $E\left[e^{i\gamma\Delta y}\right]$ is a factor of $E\left[e^{i\xi\widetilde{x}^{*}}e^{i\gamma y}\right]$ implies that it is a factor of $E\left[e^{i\gamma y}|\tilde{x}^{*}=\tilde{x}_{0}^{*}\right]$ for any fixed $\tilde{x}_{0}^{*}$ in the support of $\tilde{x}^{*}$ (except perhaps for x_0 in some null set under the probability measure of $\tilde{x}^{*}$). But $E\left[e^{i\gamma y}|\tilde{x}^{*}=\tilde{x}_{0}^{*}\right]=E\left[e^{i\gamma(\widetilde{g}(\widetilde{x}^{*})+\Delta\widetilde{y})}|\tilde{x}^{*}=\tilde{x}_{0}^{*}\right]=e^{i\gamma\widetilde{g}\left(\widetilde{x}_{0}^{*}\right)}E\left[e^{i\gamma\Delta\widetilde{y}}|\tilde{x}^{*}=\tilde{x}_{0}^{*}\right]=e^{i\gamma\widetilde{g}\left(\widetilde{x}_{0}^{*}\right)}E\left[e^{i\gamma\Delta\widetilde{y}}\right]$ and it follows that $E\left[e^{i\gamma\Delta y}\right]$ is a factor of $E\left[e^{i\gamma\Delta\widetilde{y}}\right]$ (since $e^{i\gamma\widetilde{g}\left(\widetilde{x}_{0}^{*}\right)}$ is the c.f. of a point mass at $\widetilde{g}\left(\widetilde{x}_{0}^{*}\right)$). This then implies that there exists a random variable $\Delta\bar{y}$ independent of Δy such that $\Delta\widetilde{y}=\Delta y+\Delta\bar{y}$. By Lemma A.1, this implies $\inf_{c\in\mathbb{R}}E\left[|\Delta\widetilde{y}-c|\right]\geq\inf_{c\in\mathbb{R}}E\left[|\Delta y-c|\right]$, as assumed (if this had not been the case, we would have been forced to permute the role of the Models 1 and A.2).

Next, turning to the left-hand side of (A.7), we can also show, using a similar argument, that $E\left[e^{i\xi\Delta\widetilde{x}}\right]$ is a factor of $E\left[e^{i\xi\Delta x}\right]$. We have $E\left[e^{i\xi x^{*}}e^{i\gamma g(x^{*})}\right]E\left[e^{i\xi\Delta x}\right]=E\left[e^{i\xi x}e^{i\gamma g(x^{*})}\right]$ and by Lemma A.2, the fact that $E\left[e^{i\gamma\Delta\tilde{x}}\right]$ is a factor of $E\left[e^{i\xi x}e^{i\gamma g(x^{*})}\right]$ implies that it is a factor of $E\left[e^{i\xi x}|g\left(x^{*}\right)=y_{0}^{*}\right]$ for almost any fixed y_0^{*} in the support of the random variable $y^{*}\equiv g\left(x^{*}\right)$. To proceed as before we must show that there exists a y_0^{*} such that $g\left(x^{*}\right)=y_{0}^{*}$ has a unique root. If the set of stationary points χ (from Assumption 6) is empty, then the root is always unique and the problem is solved. If χ is nonempty, we note that, by Assumptions 5 and 6, the range

of values of y_0^* where multiple roots occur must be contained in the finite interval $[\min_{x^* \in \chi} g(x^*), \max_{x^* \in \chi} g(x^*)]$. It thus suffice to show that one can find a y_0^* sufficiently large to be outside of that interval, that is, the range of $g(x^*)$ is sufficiently large. In fact, we can show by contradiction that the range of $g(x^*)$ must be $\mathbb{R}$. Consider the inverse Fourier transform of Equation (A.7) with respect to the variables ξ, γ and denote it by $\mu_{\bar{x}\bar{y}}$. This quantity is a well-defined probability measure since we had shown that the right-hand side of (A.7) is a c.f.. If the range of $g(x^*)$ were not $\mathbb{R}$, the left-hand side of Equation (A.7) implies that the support of $\mu_{\bar{x}\bar{y}}$ would exhibit at least some boundaries that are horizontal (i.e. of constant $\bar{y}$ for a range of values of $\bar{x}$). On the other hand, the right-hand side of (A.7) and the fact that $E\left[e^{i\gamma\Delta\widetilde{y}}\right]/E\left[e^{i\gamma\Delta y}\right]$ is a c.f. implies that horizontal boundaries in the support of $\mu_{\bar{x}\bar{y}}$ are only possible if $\widetilde{g}(\tilde{x}^*)$ were constant over some interval, which violates Assumption 6. Hence the range of $g(x^*)$ must be $\mathbb{R}$ and we can pick y_0^* sufficiently large so that $g(x^*) = y_0^*$ has a unique root $x^* = x_0^*$. If $E\left[e^{i\xi\Delta\widetilde{x}}\right]$ is a factor of $E\left[e^{i\xi x}|g(x^*) = y_0^*\right] = E\left[e^{i\xi x^* + \Delta x}|g(x^*) = y_0^*\right] = e^{i\xi x_0^*} E\left[e^{i\xi\Delta x}|g(x^*) = y_0^*\right] = e^{i\xi x_0^*} E\left[e^{i\xi\Delta x}\right]$ then it is also a factor of $E\left[e^{i\xi\Delta x}\right]$ since $e^{i\xi x_0^*}$ is the c.f. of a point mass.

In summary, since $E\left[e^{i\xi\Delta\widetilde{x}}\right]$ is a factor of $E\left[e^{i\xi\Delta x}\right]$ and $E\left[e^{i\gamma\Delta y}\right]$ is a factor of $E\left[e^{i\gamma\Delta\widetilde{y}}\right]$, we have just shown that there exist random variables $\Delta\bar{x}$ and $\Delta\bar{y}$ such that $E\left[e^{i\xi\Delta x}\right]/E\left[e^{i\xi\Delta\widetilde{x}}\right] = E\left[e^{i\xi\Delta\bar{x}}\right]$ and $E\left[e^{i\gamma\Delta\widetilde{y}}\right]/E\left[e^{i\gamma\Delta y}\right] = E\left[e^{i\gamma\Delta\bar{y}}\right]$, implying that Equation (A.7) can be written as

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

thus indicating the observational equivalence of two models of the form Model A.3 and A.4. ■

Lemma A.4 *Let Assumptions 4-6 hold. If Models A.3 and A.4 are distinct but observationally equivalent, then the distributions of $\Delta\bar{x}$ and $\Delta\bar{y}$ both admit an absolutely continuous density with respect to the Lebesgue measure on $\mathbb{R}$ that never vanishes. Moreover, $\bar{x}$ and $\bar{y}$ are also supported on $\mathbb{R}$.*

Proof. By the Lebesgue decomposition theorem (see Theorem 8.1.A in Loève (1977)), the probability measure of $(\bar{x}, \bar{y})$ can be decomposed into the sum of a measure λ that is absolutely continuous with respect to the Lebesgue measure and a singular measure σ supported on a set $\mathbb{S}$ of Lebesgue measure zero. For simplicity of discussion we define $f_{\bar{x}\bar{y}}(\bar{x}, \bar{y})$ for $(\bar{x}, \bar{y}) \in \mathbb{R}^2 \backslash \mathbb{S}$ to be equal to the density with respect to the Lebesgue measure associated with λ (as defined via Equation (A.1)).[2] In a slight abuse of notation, we say that "$f_{\bar{x}\bar{y}}(\bar{x}, \bar{y})$ is infinite" when $(\bar{x}, \bar{y}) \in \mathbb{S}$. Clearly, the statement "$f_{\bar{x}\bar{y}}(\bar{x}, \bar{y})$ is infinite" is not detailed enough to fully characterize the distribution of $(\bar{x}, \bar{y})$ over $\mathbb{S}$, but it will be sufficient for our purposes, because our proof only relies

[2] Since we compute the density from λ, not the original probability measure, the singular component of the measure does not affect Equation (A.1).

on the location of the points in $\mathbb{S}$ and not on their exact probability measure. We adopt similar conventions for the distributions of $\Delta\bar{x}$, $\Delta\bar{y}$, $\bar{x}$ and $\bar{y}$.

Observational equivalence between Models A.3 and A.4 can be written as[3]

$$\sum_k f_{\Delta\bar{x}}\left(\bar{x}-h_k\left(\bar{y}\right)\right) f_{\bar{y}}\left(\bar{y}\right)=\widetilde{f}_{\Delta\bar{y}}\left(\bar{y}-\widetilde{g}\left(\bar{x}\right)\right) f_{\bar{x}}\left(\bar{x}\right), \tag{A.8}$$

where the summation is over the branches $h_k\left(\bar{y}\right)$ of the inverse of $g\left(x^*\right)$. The number of branches may vary with $\bar{y}$, although this is not explicit in the notation. There are at most a finite number $m+1$ of such branches because m (the number of points where $g'\left(x^*\right)=0$) is finite and $g'\left(x^*\right)$ is continuous.

Consider any point $\bar{x}_0 \in \mathcal{S}_{\bar{x}}$ such that $f_{\bar{x}}\left(\bar{x}_0\right)>0$ and any point $\bar{y}_0 \in \mathcal{S}_{\bar{y}}$ such that $f_{\bar{y}}\left(\bar{y}_0\right)>0$. Note that such points are dense in $\mathcal{S}_{\bar{x}}$ and $\mathcal{S}_{\bar{y}}$, respectively, by the very definition of the support. Suppose that we have that $\widetilde{f}_{\Delta\bar{y}}\left(\bar{y}_0-\widetilde{g}\left(\bar{x}_0\right)\right)$ becomes zero or infinite at the point $\left(\bar{x}_0,\bar{y}_0\right)$. Since $f_{\bar{x}}\left(\bar{x}_0\right)$ is finite by Assumption 4, it follows that $\widetilde{f}_{\Delta\bar{y}}\left(\bar{y}_0-\widetilde{g}\left(\bar{x}_0\right)\right) f_{\bar{x}}\left(\bar{x}_0\right)$ becomes zero or infinite at the same points as $\widetilde{f}_{\Delta\bar{y}}\left(\bar{y}_0-\widetilde{g}\left(\bar{x}_0\right)\right)$. By Equation (A.8), $\sum_k f_{\Delta\bar{x}}\left(\bar{x}-h_k\left(\bar{y}\right)\right) f_{\bar{y}}\left(\bar{y}\right)$ also becomes zero or infinite at the same point(s).

Now, $\bar{y}_0$ is such that $f_{\bar{y}}\left(\bar{y}_0\right)>0$. If $f_{\bar{y}}\left(\bar{y}_0\right)$ were infinite, then $\sum_k f_{\Delta\bar{x}}\left(\bar{x}-h_k\left(\bar{y}\right)\right) f_{\bar{y}}\left(\bar{y}\right)$ would be infinite along a the line $\bar{y}=\bar{y}_0$. But it would then be impossible for it to be equal to the function $\widetilde{f}_{\Delta\bar{y}}\left(\bar{y}-\widetilde{g}\left(\bar{x}\right)\right) f_{\bar{x}}\left(\bar{x}\right)$, which could only be infinite along a curve of constant $\bar{y}-\widetilde{g}\left(\bar{x}\right)$. The two statements are compatible only if $\widetilde{g}\left(\bar{x}\right)$ is also constant, which is ruled out by Assumption 6. We are left with the possibility that $f_{\bar{y}}\left(\bar{y}_0\right)$ is nonzero and finite. Hence, the fact that $\sum_k f_{\Delta\bar{x}}\left(\bar{x}_0-h_k\left(\bar{y}_0\right)\right) f_{\bar{y}}\left(\bar{y}_0\right)$ is zero or infinite must be due to $\sum_k f_{\Delta\bar{x}}\left(\bar{x}_0-h_k\left(\bar{y}_0\right)\right)$ being zero or infinite. We have just shown that whenever $\widetilde{f}_{\Delta\bar{y}}\left(\bar{y}_0-\widetilde{g}\left(\bar{x}_0\right)\right)$ is zero or infinite, the same behavior occurs for $\sum_k f_{\Delta\bar{x}}\left(\bar{x}_0-h_k\left(\bar{y}_0\right)\right)$.

First consider the case where $\widetilde{f}_{\Delta\bar{y}}\left(\bar{y}_0-\widetilde{g}\left(\bar{x}_0\right)\right)$ is infinite at a point. The same behavior would also occur along the whole curve $\mathcal{V}_{v_0}$ containing points $(\bar{x},\bar{y})$ giving the same value of $v_0 \equiv \bar{y}_0-\widetilde{g}\left(\bar{x}_0\right)=\bar{y}-\widetilde{g}\left(\bar{x}\right)$ and along the set $\mathcal{U}^*$ of values of $(\bar{x},\bar{y})$ giving the same value of $u_0 \equiv \bar{x}_0-h_k\left(\bar{y}_0\right)=\bar{x}-h_{k'}\left(\bar{y}\right)$ for some k,k' (because it suffices that one term in the sum $\sum_k f_{\Delta\bar{x}}\left(\bar{x}_0-h_k\left(\bar{y}_0\right)\right)$ is infinite for this sum of positive quantities to be infinite). These two sets contain, respectively, the curves $\mathcal{V}_{v_0}$ and $\mathcal{U}_{u_0}$, where

$$\mathcal{V}_v=\left\{\left(\widetilde{x}^*,\widetilde{g}\left(\widetilde{x}^*\right)+v\right):\widetilde{x}^* \in \mathcal{S}_{\widetilde{x}^*}\right\} \text{ and } \mathcal{U}_u=\left\{\left(x^*+u,g\left(x^*\right)\right):x^* \in \mathcal{S}_{x^*}\right\}. \tag{A.9}$$

[3] We consider the statement $\infty=\infty$ to indicate that both sides fail to admit a Lebesgue density at the same points. Also, away from these singular points, the equality holds pointwise if the densities are defined as in Equation (A.1).

If $\mathcal{V}_{v_0}$ and $\mathcal{U}_{u_0}$ did not coincide, then it would be possible to recursively construct the following sequence of sets

$$\begin{aligned}
\mathcal{V}^0 &\equiv \mathcal{V}_{v_0} \\
\mathcal{U}^0 &\equiv \mathcal{U}_{u_0} \\
\mathcal{V}^{n+1} &= \bigcup_{v:\mathcal{V}_v \cap \mathcal{U}^n \neq \emptyset} \mathcal{V}_v \\
\mathcal{U}^{n+1} &= \bigcup_{u:\mathcal{U}_u \cap \mathcal{V}^{n+1} \neq \emptyset} \mathcal{U}_u
\end{aligned}$$

that is such that[4] $\mathcal{V}^n \to \mathcal{S}_{\bar{x}\bar{y}}$ and $\mathcal{U}^n \to \mathcal{S}_{\bar{x}\bar{y}}$. This implies that $f_{\Delta\bar{x}}$ and $\widetilde{f}_{\Delta\bar{y}}$ are everywhere infinite, in contradiction to Lebesgue's decomposition theorem, which states that these divergences could only occur on a set of null measure. Therefore, the curves $\mathcal{V}_{v_0}$ and $\mathcal{U}_{u_0}$ have to coincide:

$$\left(\widetilde{x}^*, \widetilde{g}\left(\widetilde{x}^*\right) + v_0\right) = \left(x^* + u_0, g\left(x^*\right)\right),$$

implying that

$$\widetilde{g}\left(x^* + u_0\right) + v_0 = g\left(x^*\right),$$

i.e., $\widetilde{g}(\cdot)$ and $g(\cdot)$ are just horizontally and vertically shifted versions of each other. But any nonzero shift would imply that either one of the models is violating one of the zero mean assumptions on the disturbances.[5] Hence, for any pair of valid models A.3 and A.4, we must have $\widetilde{g}\left(x^*\right) = g\left(x^*\right)$.

Now consider the case where $\widetilde{f}_{\Delta\bar{y}}\left(\bar{y}_0 - \widetilde{g}\left(\bar{x}_0\right)\right)$ is zero at some point. Then, $\widetilde{f}_{\Delta\bar{y}}\left(\bar{y}_0 - \widetilde{g}\left(\bar{x}_0\right)\right)$ would have to be zero along the whole curve $\mathcal{V}_{v_0}$ containing points $(\bar{x}, \bar{y})$ giving the same value of $v_0 \equiv \bar{y}_0 - \widetilde{g}\left(\bar{x}_0\right) = \bar{y} - \widetilde{g}\left(\bar{x}\right)$. Furthermore, the quantity $\sum_k f_{\Delta\bar{x}}\left(\bar{x} - h_k\left(\bar{y}\right)\right)$ must also vanish at those points. This means that $f_{\Delta\bar{x}}\left(u_0\right) = 0$ for $u_0 = \bar{x} - h_k\left(\bar{y}\right)$ for all k and all $(\bar{x}, \bar{y}) \in \mathcal{V}_{v_0}$.

If the $h_k(\cdot)$ do not constitute the inverse of $\widetilde{g}(\cdot)$, it is then possible to recursively construct the following sequence of sets

$$\begin{aligned}
\mathcal{V}^0 &\equiv \mathcal{V}_{v_0} \qquad\qquad (A.10) \\
\mathcal{U}^0 &\equiv \left\{(\bar{x}, \bar{y}) : \left(\bar{x} - h_k\left(\bar{y}\right)\right) = \left(\bar{x}_0 - h_{k'}\left(\bar{y}_0\right)\right) \text{ for some } k, k' \text{ for some } \left(\bar{x}_0, \bar{y}_0\right) \in \mathcal{V}_{v_0}\right\} \\
\mathcal{V}^{n+1} &= \bigcup_{v:\mathcal{V}_v \cap \mathcal{U}^n \neq \emptyset} \mathcal{V}_v \\
\mathcal{U}^{n+1} &= \left\{(\bar{x}, \bar{y}) : \left(\bar{x} - h_k\left(\bar{y}\right)\right) = \left(\bar{x}_0 - h_{k'}\left(\bar{y}_0\right)\right) \text{ for some } k, k' \text{ for some } \left(\bar{x}_0, \bar{y}_0\right) \in \mathcal{V}_v^{n+1}\right\}
\end{aligned}$$

that is such that $\mathcal{V}^n \to \mathcal{S}_{\bar{x}\bar{y}}$ and $\mathcal{U}^n \to \mathcal{S}_{\bar{x}\bar{y}}$. Hence, $\widetilde{f}_{\Delta\bar{y}}(v)$ and $f_{\Delta\bar{x}}(u)$ would have to vanish everywhere, which is impossible. This contradiction leads to the conclusion that, in fact, the $h_k(\cdot)$ do constitute the inverse of $\widetilde{g}(\cdot)$, i.e. $g\left(x^*\right) = \widetilde{g}\left(x^*\right)$.

[4] Convergence of sets is in the Hausdorff metric.

[5] The only exception in the linear specification, where two nonzero shifts along each axes may cancel each other. But in this case, the shifted curve is identical to the original one.

We have just shown that if the densities of $\Delta\bar{y}$ or $\Delta\bar{x}$ either vanish or are infinite at some point, then $g(x^*) = \widetilde{g}(x^*)$. The density of x^* could then be determined (up to a multiplicative constant determined by the normalization of unit total probability) from the density $f_{\bar{x}\bar{y}}(\bar{x}, \bar{y})$ along the line $\bar{y} = g(\bar{x}) + u$ for some $u \in \mathcal{S}_{\Delta\bar{y}}$.

This means that if there are any points where $\widetilde{f}_{\Delta\bar{y}}$ or $f_{\Delta\bar{x}}$ become zero or infinite, then Model A.3 and A.4 are such that $\widetilde{g}(x^*) = g(x^*)$ and $\widetilde{f}_{\tilde{x}^*}(x^*) = f_{x^*}(x^*)$. So any pair of *distinct* but *observationally equivalent* models must be such that $\widetilde{f}_{\Delta\bar{y}}$ and $f_{\Delta\bar{x}}$ are well-defined densities with respect to the Lebesgue measure that are nonzero and finite. Since $\widetilde{f}_{\Delta\bar{y}}$ and $f_{\Delta\bar{x}}$ are supported on $\mathbb{R}$, so must $f_{\bar{x}}$ and $f_{\bar{y}}$, in light of Equation (A.8). ■

Lemma A.5 *Under Assumptions 4-6, Model A.3 with a specification $g(x^*)$ that is not strictly monotone cannot be observationally equivalent to Model A.4 (whether its specification is strictly monotone or not).*

Proof. If $g(x^*)$ is not strictly monotone, we have

$$f_{y^*}(y^*) = \sum_k \frac{f_{x^*}(h_k(y^*))}{|g'(h_k(y^*))|}$$

where $h_k(\cdot)$ denotes one of the branches of the inverse of $g(\cdot)$. There are at most a finite number $m+1$ of such branches because $g'(x^*)$ is continuous and vanishes a finite number m of times (by Assumption 6). Since $f_{x^*}(x^*) \neq 0$ for x^* in a neighborhood of any one of the x_k^* where $g'(x_k^*) = 0$, the density of y^* diverges as $y^* \to g(x_k^*)$. In the model with errors along $\bar{x}$ only, $y^* = \bar{y}$, so the observed density of $\bar{y}$ exhibits divergences whenever $\bar{y} \to g(x_k^*)$ for some k.

In Model A.4, with errors only along $\bar{y}$, the observable joint density of $\bar{x}$ and $\bar{y}$ is

$$\widetilde{f}_{x^*}(\bar{x})\, \widetilde{f}_{\Delta\bar{y}}(\bar{y} - \widetilde{g}(\bar{x}))$$

where none of the quantities can diverge either by Assumption 4 or by Lemma A.4 and hence the marginal

$$f_{\bar{y}}(\bar{y}) = \int \widetilde{f}_{x^*}(\bar{x})\, \widetilde{f}_{\Delta\bar{y}}(\bar{y} - \widetilde{g}(\bar{x}))\, d\bar{x}$$

cannot diverge either since

$$f_{\bar{y}}(\bar{y}) \leq \sup_v \widetilde{f}_{\Delta\bar{y}}(v) \int \widetilde{f}_{x^*}(\bar{x})\, d\bar{x} = \sup_v \widetilde{f}_{\Delta\bar{y}}(v),$$

where $\sup_v \widetilde{f}_{\Delta\bar{y}}(v) < \infty$ by Lemma A.4. Hence Models A.3 and A.4 cannot be observationally equivalent. ■

Lemma A.6 *Under Assumptions 4-5, Model A.3 with a strictly monotone specification $g(x^*)$ cannot be observationally equivalent to Model A.4 with a nonmonotone specification $\widetilde{g}(\widetilde{x}^*)$.*

Proof. If Model A.3 is monotone, the inverse of $g(x^*)$ has a unique branch $h(\bar{y})$ and Equation (A.8) reduces to

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

By Lemma A.4 the existence of two observationally equivalent models implies that all densities are strictly positive and we can take logs on each side:

$$\ln f_{\Delta\bar{x}}(\bar{x} - h(\bar{y})) + \ln f_{\bar{y}}(\bar{y}) = \ln \widetilde{f}_{\Delta\bar{y}}(\bar{y} - \widetilde{g}(\bar{x})) + \ln f_{\bar{x}}(\bar{x}). \tag{A.11}$$

Taking the mixed derivative of each side with respect to $\bar{x}$ and $\bar{y}$ yields:[6]

$$F''_{\Delta\bar{x}}(\bar{x} - h(\bar{y}))\, h'(\bar{y}) = \widetilde{F}''_{\Delta\bar{y}}(\bar{y} - \widetilde{g}(\bar{x}))\, \widetilde{g}'(\bar{x}) \tag{A.12}$$

where

$$F_{\Delta\bar{x}}(u) \equiv \ln f_{\Delta\bar{x}}(u) \tag{A.13}$$

$$\widetilde{F}_{\Delta\bar{y}}(v) \equiv \ln \widetilde{f}_{\Delta\bar{y}}(v). \tag{A.14}$$

Note that $h'(\bar{y})$ and $\widetilde{g}'(\bar{x})$ exist by Assumption 5 and the assumed strict monotonicity of $g(x^*)$. We may assume second differentiability of $F_{\Delta\bar{x}}$ and $\widetilde{F}_{\Delta\bar{y}}$ using an argument similar to Lemma A.4. As soon as $F''_{\Delta\bar{x}}$ fails to exist, it also fails to exist along a whole curve of constant $\bar{x} - h(\bar{y})$ and similarly for $\widetilde{F}''_{\Delta\bar{y}}$ along a curve of constant $\bar{y} - \widetilde{g}(\bar{x})$. If these two curves coincide, then $g(x^*) = \widetilde{g}(x^*)$ and both models would have to be identical and monotone. If these curves do not coincide, then we could recursively show that $F_{\Delta\bar{x}}(u)$ and $\widetilde{F}_{\Delta\bar{y}}(v)$ are nowhere twice differentiable (using the same argument as in the proof of Lemma A.4, Equation (A.10)). This is impossible because if one considers the following differential of the right-hand side of (A.11):

$$\begin{aligned}\frac{\partial}{\partial v}\frac{\partial}{\partial \bar{x}}\left(\left[\ln \widetilde{f}_{\Delta\bar{y}}(\bar{y} - \widetilde{g}(\bar{x})) + \ln f_{\bar{x}}(\bar{x})\right]_{\bar{y}=\widetilde{g}(\bar{x})+v}\right) &= \frac{\partial}{\partial v}\frac{\partial}{\partial \bar{x}}\left(\ln \widetilde{f}_{\Delta\bar{y}}(v) + \ln f_{\bar{x}}(\bar{x})\right) \\ &= \frac{\partial}{\partial \bar{x}}\frac{\partial}{\partial v}\ln f_{\bar{x}}(\bar{x}) = 0,\end{aligned}$$

but evaluating the same differential using the left-hand side of (A.11) yields:

$$\begin{aligned}&\frac{\partial}{\partial v}\frac{\partial}{\partial \bar{x}}\left(\left[\ln f_{\Delta\bar{x}}(\bar{x} - h(\bar{y})) + \ln f_{\bar{y}}(\bar{y})\right]_{\bar{y}=\widetilde{g}(\bar{x})+v}\right) \\ = \;& \frac{\partial}{\partial v}\frac{\partial}{\partial \bar{x}}\left(\ln f_{\Delta\bar{x}}(\bar{x} - h(\widetilde{g}(\bar{x}) + v)) + \ln f_{\bar{y}}(\widetilde{g}(\bar{x}) + v)\right).\end{aligned}$$

[6] Note that the mixed derivative of a function of only one variable (say, $a(x)$) must be zero regardless of its differentiability: $\lim_{\varepsilon\to 0}\lim_{\eta\to 0}\frac{1}{\eta}\left(\frac{a(x+\varepsilon)-a(x)}{\varepsilon} - \frac{a(x+\varepsilon)-a(x)}{\varepsilon}\right) = \lim_{\varepsilon\to 0}\lim_{\eta\to 0} 0 = 0$.

But if $\ln f_{\Delta\bar{x}}$ is indeed everywhere not twice differentiable, the second expression cannot give 0 (the two terms on the right cannot cancel each other everywhere perfectly since they depend differently on the arguments). Hence, we may indeed assume second differentiability of $F_{\Delta\bar{x}}$ (and $\widetilde{F}_{\Delta\bar{y}}$, by a similar argument).

Next, we note that neither $F''_{\Delta\bar{x}}(u)$ nor $\widetilde{F}''_{\Delta\bar{y}}(v)$ can ever change sign, because changes in sign in the expression (A.12) can only occur on curves of constant $\bar{x}-h(\bar{y})$ or constant $\bar{y}$ on the left and curves of constant $\bar{y}-\widetilde{g}(\bar{x})$ or constant $\bar{x}$ on the right. These curves cannot coincide unless h is the inverse of $\widetilde{g}$, making the two models identical (and both monotone). Hence, for distinct models, $\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))$ and $F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))$ never change signs. Rearranging (A.12) yields:

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

Hence if $\widetilde{g}'(\bar{x})$ ever changes sign, so must $h'(\bar{y})$, in contradiction with the assumption that Model A.3 is monotone. ■

Lemma A.7 *Let Assumption 4 hold, $h(\cdot) \equiv g^{-1}(\cdot)$ and let $g(\cdot)$ and $\widetilde{g}(\cdot)$ be as defined in Models A.3 and A.4, respectively. These models are assumed to be distinct and both strictly monotone. If two functions $a(\cdot)$ and $b(\cdot)$ are such that $a(\bar{y}-\widetilde{g}(\bar{x})) = b(\bar{x}-h(\bar{y}))\ \forall(\bar{x},\bar{y}) \in \mathbb{R}^2$, then $a(\cdot)$ and $b(\cdot)$ are constant functions over $\mathbb{R}$. Similarly if $a(\bar{y}-\widetilde{g}(\bar{x})) = 0 \Leftrightarrow b(\bar{x}-h(\bar{y})) = 0\ \forall(\bar{x},\bar{y}) \in \mathbb{R}^2$, then $a(\cdot)$ and $b(\cdot)$ are zero over $\mathbb{R}$ if either one vanishes at a single point.*

Proof. Note that, by Lemma A.4, $\{(\bar{y}-\widetilde{g}(\bar{x}), \bar{x}-h(\bar{y})) : \forall(\bar{x},\bar{y}) \in \mathbb{R}^2\} = \mathbb{R}^2$. It is therefore possible to vary $\bar{x}$ and $\bar{y}$ so that $\Delta\bar{y} = \bar{y}-\widetilde{g}(\bar{x})$ remains constant while $\Delta\bar{x} = \bar{x}-h(\bar{y})$ varies or vice-versa. Hence, it is possible to vary $(\bar{x},\bar{y})$ in such a way such that $\Delta\bar{x}$ varies but $\Delta\bar{y}$ remains constant. Having $a(\Delta\bar{y})$ constant implies that $b(\Delta\bar{x})$ also is, even though its argument is varying. This shows that $b(\Delta\bar{x})$ is constant along a one-dimensional slice of constant $\Delta\bar{y}$. Then, varying $(\bar{x},\bar{y})$ so that the argument of the $b(\Delta\bar{x})$ is constant, we can show that the $a(\Delta\bar{y})$ is constant along a one-dimensional slice of constant $\Delta\bar{x}$. Repeating the process (using the same argument as in the proof of Lemma A.4, Equation (A.10)) we can show that $a(\Delta\bar{y})$ and $b(\Delta\bar{x})$ are constant for all $(\Delta\bar{x},\Delta\bar{y}) \in \mathbb{R}^2$ and therefore for all $(\bar{x},\bar{y}) \in \mathbb{R}^2$. A similar argument demonstrates the second conclusion of the Lemma. ■

A.2 Proof of Theorem 1

We now know from Lemmas A.3-A.6 that proving Theorem 1 is equivalent to finding two distinct but observationally equivalent Models A.3 and A.4 that are both strictly monotone and involve random variables admitting nonvanishing densities with respect to the Lebesgue measure.

Under Model A.3, the joint density of $\bar{x}$ and $\bar{y}$ can be written as:

$$f_{\bar{x}\bar{y}}(\bar{x},\bar{y}) = f_{\Delta\bar{x}}(\bar{x} - h(\bar{y}))\, f_{\bar{y}}(\bar{y}) \tag{A.15}$$

where $h(y) \equiv g^{-1}(y)$ (which exists and has a single branch by Lemmas A.5-A.6), while under Model A.4, we have

$$f_{\bar{x}\bar{y}}(\bar{x},\bar{y}) = \widetilde{f}_{\Delta\bar{y}}(\bar{y} - \widetilde{g}(\bar{x}))\, f_{\bar{x}}(\bar{x}) \tag{A.16}$$

where the $\sim$ on the densities emphasizes the quantities that differ under the alternative model.

Since the two models must be observationally equivalent, we equate (A.15) and (A.16):

$$f_{\Delta\bar{x}}(\bar{x} - h(\bar{y}))\, f_{\bar{y}}(\bar{y}) = \widetilde{f}_{\Delta\bar{y}}(\bar{y} - \widetilde{g}(\bar{x}))\, f_{\bar{x}}(\bar{x}). \tag{A.17}$$

After rearranging (A.17) and taking logs, we obtain:

$$\ln \widetilde{f}_{\Delta\bar{y}}(\bar{y} - \widetilde{g}(\bar{x})) - \ln f_{\Delta\bar{x}}(\bar{x} - h(\bar{y})) = \ln f_{\bar{y}}(\bar{y}) - \ln f_{\bar{x}}(\bar{x}), \tag{A.18}$$

where these densities are always positive (by Lemma A.4), so that the $\ln(\cdot)$ are always well-defined.

We will find necessary conditions for Equation (A.18) to hold, in order to narrow down the search for possible solutions that would provide distinct but observationally equivalent models. Next, we will need to check that these solutions actually lead to proper densities (i.e. with finite area) for all variables in order to obtain necessary and sufficient condition for identifiability.

Note that differentiability of $g(x^*)$ (Assumption 5), combined with $g'(x^*) \neq 0$ (by Lemmas A.5-A.6) implies that $h(\bar{y}) \equiv g^{-1}(\bar{y})$ is differentiable.

Let F denote the logarithms of the corresponding lowercase density and rewrite Equation (A.18) as

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

Differentiating with respect to $\bar{x}$ and $\bar{y}$, this implies that

$$\begin{aligned} \frac{\partial^2}{\partial\bar{x}\partial\bar{y}}\widetilde{F}_{\Delta\bar{y}}(\bar{y} - \widetilde{g}(\bar{x})) - \frac{\partial^2}{\partial\bar{x}\partial\bar{y}}F_{\Delta\bar{x}}(\bar{x} - h(\bar{y})) &= 0 \\ \widetilde{F}''_{\Delta\bar{y}}(\bar{y} - \widetilde{g}(\bar{x}))\,\widetilde{g}'(\bar{x}) - F''_{\Delta\bar{x}}(\bar{x} - h(\bar{y}))\,h'(\bar{y}) &= 0 \end{aligned} \tag{A.19}$$

In the above, we may assume twice differentiability of $\widetilde{F}_{\Delta\bar{y}}$ and $\widetilde{F}_{\Delta\bar{x}}$, by the same argument as in Lemma A.6 (in essence, lack of differentiability would imply that the two model must be identical). We can rearrange Equation (A.19) to yield

$$\widetilde{F}''_{\Delta\bar{y}}(\bar{y} - \widetilde{g}(\bar{x})) = \frac{h'(\bar{y})}{\widetilde{g}'(\bar{x})}F''_{\Delta\bar{x}}(\bar{x} - h(\bar{y})), \tag{A.20}$$

where the ratio $h'(\bar{y})/\widetilde{g}'(\bar{x})$ is nonzero and finite by assumption. Hence if $F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))$ is zero, then so is $\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))$ and vice versa. If either of those two functions vanishes at a point, by Lemma A.7, they must vanish everywhere. It would follow that $\widetilde{F}_{\Delta\bar{y}}(\Delta\bar{y})$ and $F_{\Delta\bar{x}}(\Delta\bar{x})$ would be linear and that the corresponding densities $\widetilde{f}_{\Delta\bar{y}}(\Delta\bar{y})$ and $f_{\Delta\bar{x}}(\Delta\bar{x})$ would be exponential over $\mathbb{R}$, which is an improper density. It follows that our presumption that either $F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))$ or $\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))$ vanish at some point is incorrect.

Hence we may assume that $F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))$ and $\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))$ do not vanish. Since these functions are continuous, this means they never change sign. Also note that, by assumption, $h'(\bar{y})$ and $\widetilde{g}'(\bar{x})$ never change sign or vanish either. We can thus, without loss of generality, rewrite Equation (A.19) as:

$$\frac{\left|\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))\right|}{\left|F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))\right|}=\frac{\left|h'(\bar{y})\right|}{\left|\widetilde{g}'(\bar{x})\right|} \tag{A.21}$$

or

$$\ln\left|\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))\right|-\ln\left|F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))\right|=\ln\left|h'(\bar{y})\right|-\ln\left|\widetilde{g}'(\bar{x})\right|$$

Again, since the right-hand side is a difference of functions of $\bar{y}$ and of $\bar{x}$, we must have[7]

$$\begin{aligned}\frac{\partial^2}{\partial x\partial y}\ln\left|\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))\right|-\frac{\partial^2}{\partial x\partial y}\ln\left|F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))\right| &= 0\\ \left(\ln\left|\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))\right|\right)''\widetilde{g}'(\bar{x})-\left(\ln\left|F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))\right|\right)''h'(\bar{y}) &= 0\end{aligned}$$

By the same argument as before, if $\left(\ln\left|\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))\right|\right)''=0$ or $\left(\ln\left|F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))\right|\right)''=0$ at a point then they must vanish everywhere.

Hence there are two mutually exclusive situations, either $\left(\ln\left|\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))\right|\right)''$ and $\left(\ln\left|F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))\right|\right)''$ never vanish (see Case A below) or they vanish everywhere (see Case B below). In both cases, we can also re-use the argument that lack of sufficient continuous differentiability implies identification, as in Lemma A.6, hence, for the purpose of finding models that are not identified, we can assume sufficient continuous differentiability.

Case A: If $\left(\ln\left|\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))\right|\right)''$ and $\left(\ln\left|F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))\right|\right)''$ do not vanish, we may write

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

[7]The notation $\left(\ln\left|\widetilde{F}''_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))\right|\right)''$ stands for $\left(\ln\left|\widetilde{F}''_{\Delta\bar{y}}(u)\right|\right)''|_{u=\bar{y}-\widetilde{g}(\bar{x})}$.

combined with Equation (A.21) this implies:

$$\begin{aligned}
\frac{\left|\widetilde{F}''_{\Delta\bar{y}}\left(\bar{y}-\widetilde{g}\left(\bar{x}\right)\right)\right|}{\left|F''_{\Delta\bar{x}}\left(\bar{x}-h\left(\bar{y}\right)\right)\right|} &= \frac{\left|\left(\ln\left|\widetilde{F}''_{\Delta\bar{y}}\left(\bar{y}-\widetilde{g}\left(\bar{x}\right)\right)\right|\right)''\right|}{\left|\left(\ln\left|F''_{\Delta\bar{x}}\left(\bar{x}-h\left(\bar{y}\right)\right)\right|\right)''\right|} \\
\frac{\left|\left(\ln\left|\widetilde{F}''_{\Delta\bar{y}}\left(\bar{y}-\widetilde{g}\left(\bar{x}\right)\right)\right|\right)''\right|}{\left|\widetilde{F}''_{\Delta\bar{y}}\left(\bar{y}-\widetilde{g}\left(\bar{x}\right)\right)\right|} &= \frac{\left|\left(\ln\left|F''_{\Delta\bar{x}}\left(\bar{x}-h\left(\bar{y}\right)\right)\right|\right)''\right|}{\left|F''_{\Delta\bar{x}}\left(\bar{x}-h\left(\bar{y}\right)\right)\right|}
\end{aligned} \tag{A.22}$$

By Lemma A.7, each side of this equality must equal a constant, say A. Note that this equality is only a necessary condition for lack of identifiability. For instance, it does not ensure that $\left|\widetilde{F}''_{\Delta\bar{y}}\left(\bar{y}-\widetilde{g}\left(\bar{x}\right)\right)\right| / \left|F''_{\Delta\bar{x}}\left(\bar{x}-h\left(\bar{y}\right)\right)\right|$ can actually be written as a ratio of a function of $\bar{y}$ and a function of $\bar{x}$, as required by Equation (A.21). This will need to be subsequently checked.

We now find densities such that the left-hand (or right-hand) side of Equation (A.22) is constant. Letting $u=\bar{y}-\widetilde{g}\left(\bar{x}\right)$ and $F\left(\cdot\right)\equiv\widetilde{F}_{\Delta\bar{y}}\left(\cdot\right)$ (or similarly, $u=\bar{x}-h\left(\bar{y}\right)$ and $F\left(\cdot\right)\equiv F_{\Delta\bar{x}}\left(\cdot\right)$), we must have that

$$\begin{aligned}
\frac{\left(\ln\left|F''\left(u\right)\right|\right)''}{F''\left(u\right)} &= \pm A \\
\left(\ln\left|F''\left(u\right)\right|\right)'' &= \pm AF''\left(u\right) \\
\left(\ln\left|F''\left(u\right)\right|\right)' &= \pm AF'\left(u\right)+B \\
\ln\left|F''\left(u\right)\right| &= \pm AF\left(u\right)+Bu+C \\
F''\left(u\right) &= \pm\exp\left(\pm AF\left(u\right)+Bu+C\right) \\
F''\left(u\right) &= -\exp\left(AF\left(u\right)+Bu+C\right)
\end{aligned} \tag{A.23}$$

where A, B, C are some constants and where one of the "$\pm$" has been incorporated into the constant A and the other has been set to "$-$", because the "$+$" solution does not lead to a proper density.

The solution $F\left(u\right)$ to Equation (A.23) is:

$$F\left(u\right)=-\frac{B}{A}u-\frac{C}{A}+\frac{1}{A}\ln\left(\frac{2D^{2}}{A}\rho\left(D\left(u-u_{0}\right)\right)\right) \tag{A.24}$$

where

$$\rho\left(v\right)=1-\tanh^{2}\left(v\right)=4\left(\exp\left(v\right)+\exp\left(-v\right)\right)^{-2}$$

and where A, B, C, D, u_0 are constants. This solution can be verified by substitution into the differential equation and by noting that any initial conditions in $F\left(0\right)$ and $F'\left(0\right)$ can be accommodated by adjusting the constants D, u_0. Note that this solution is unique (up to the constants D and u_0).

The density corresponding to $F(u)$ is

$$f(u) = C_1 \exp\left(-\frac{B}{A}u\right)\left(\rho\left(D\left(u-u_0\right)\right)\right)^{1/A}$$

where C_1 is such that the density integrates to 1. To check that this is a valid solution, we first calculate what the implied forms of $\widetilde{g}(\bar{x})$ and $h(\bar{y})$ are. From Equation (A.20), we know that

$$\frac{\widetilde{F}''_{\Delta\bar{y}}\left(\bar{y}-\widetilde{g}\left(\bar{x}\right)\right)}{F''_{\Delta\bar{x}}\left(\bar{x}-h\left(\bar{y}\right)\right)} = \frac{h'\left(\bar{y}\right)}{\widetilde{g}'\left(\bar{x}\right)} \tag{A.25}$$

where we can find an expression for $F''_{\Delta\bar{x}}(\cdot)$ and $\widetilde{F}''_{\Delta\bar{y}}(\cdot)$, generically denoted $F''(\cdot)$ using Equations (A.23) and (A.24):

$$\begin{aligned} F''(u) &= -\exp\left(A\left(-\frac{B}{A}u-\frac{C}{A}+\frac{1}{A}\ln\left(\frac{2D^2}{A}\rho\left(D\left(u-u_0\right)\right)\right)\right)+Bu+C\right) \\ &= -\frac{2D^2}{A}\rho\left(D\left(u-u_0\right)\right). \end{aligned}$$

The constants D and u_0 may differ for $F''_{\Delta\bar{x}}(\cdot)$ and $\widetilde{F}''_{\Delta\bar{y}}(\cdot)$ and we distinguish them by subscripts $\Delta\bar{x}$ or $\Delta\bar{y}$. The constant A is the same, however. Next, we calculate the ratio:

$$\begin{aligned} \frac{\widetilde{F}''_{\Delta\bar{y}}\left(\bar{y}-\widetilde{g}\left(\bar{x}\right)\right)}{F''_{\Delta\bar{x}}\left(\bar{x}-h\left(\bar{y}\right)\right)} &= \frac{-\frac{2D^2_{\Delta\bar{y}}}{A}\rho\left(D_{\Delta\bar{y}}\left(\bar{y}-\widetilde{g}\left(\bar{x}\right)-u_{0\Delta\bar{y}}\right)\right)}{-\frac{2D^2_{\Delta\bar{x}}}{A}\rho\left(D_{\Delta\bar{x}}\left(\bar{x}-h\left(\bar{y}\right)-u_{0\Delta\bar{x}}\right)\right)} \\ &= \frac{D^2_{\Delta\bar{y}}\left(\exp\left(D_{\Delta\bar{y}}\left(\bar{y}-\widetilde{g}\left(\bar{x}\right)-u_{0\Delta\bar{y}}\right)\right)+\exp\left(-D_{\Delta\bar{y}}\left(\bar{y}-\widetilde{g}\left(\bar{x}\right)-u_{0\Delta\bar{y}}\right)\right)\right)^{-2}}{D^2_{\Delta\bar{x}}\left(\exp\left(D_{\Delta\bar{x}}\left(\bar{x}-h\left(\bar{y}\right)-u_{0\Delta\bar{x}}\right)\right)+\exp\left(-D_{\Delta\bar{x}}\left(\bar{x}-h\left(\bar{y}\right)-u_{0\Delta\bar{x}}\right)\right)\right)^{-2}} \\ &= \frac{D^{-2}_{\Delta\bar{x}}\left(2+\exp\left(2D_{\Delta\bar{x}}\left(\bar{x}-h\left(\bar{y}\right)-u_{0\Delta\bar{x}}\right)\right)+\exp\left(-2D_{\Delta\bar{x}}\left(\bar{x}-h\left(\bar{y}\right)-u_{0\Delta\bar{x}}\right)\right)\right)}{D^{-2}_{\Delta\bar{y}}\left(2+\exp\left(2D_{\Delta\bar{y}}\left(\bar{y}-\widetilde{g}\left(\bar{x}\right)-u_{0\Delta\bar{y}}\right)\right)+\exp\left(-2D_{\Delta\bar{y}}\left(\bar{y}-\widetilde{g}\left(\bar{x}\right)-u_{0\Delta\bar{y}}\right)\right)\right)} \end{aligned}$$

and note that it cannot be written as a ratio of a function of $\bar{y}$ and a function of $\bar{x}$ (unless $\widetilde{g}(\bar{x})$ or $h(\bar{y})$ are constant, a situation ruled out by Assumptions 5 and 6). Hence Equation (A.21) cannot possibly hold and this solution is not valid. Hence, except possibly when $\left(\ln F''(u)\right)'' = 0$, there exists no pair of observationally equivalent models of the forms of Model A.3 and A.4.

Case B: We now consider the (so far excluded) case where $\left(\ln\left|F''(u)\right|\right)'' = 0$ for $F = F_{\Delta\bar{x}}$ and $\widetilde{F}_{\Delta\bar{y}}$. We have

$$\begin{aligned} \left(\ln\left|F''(u)\right|\right)'' &= 0 \\ \left|F''(u)\right| &= \exp\left(Au+B\right) \qquad (A.26) \\ F''(u) &= \pm\exp\left(Au+B\right) \\ F'(u) &= \pm A^{-1}\exp\left(Au+B\right)+C \\ F(u) &= -A^{-2}\exp\left(Au+B\right)+Cu+D \qquad (A.27) \end{aligned}$$

for some adjustable constants A, B, C, D with $A \neq 0$ (the case $A = 0$ is covered in Case C below). We have selected the negative branch of the "$\pm$" of since it is the only one yielding a proper density. The density corresponding to (A.27) is of the form

$$f(u) = \exp\left(-A^{-2}\exp(Au + B) + Cu + D\right) \tag{A.28}$$

where the constants A, B, C, D are selected so as to satisfy the normalization constraint and the zero mean assumption. In the sequel, we will distinguish the constants A, B, C, D by subscripts $\Delta\bar{x}, \Delta\bar{y}$ corresponding to the densities of $\Delta\bar{x}$ and $\Delta\bar{y}$, respectively. We first determine $h(\bar{y})$ and $g(\bar{x})$ through relationship (A.21):

$$\begin{aligned}\frac{|h'(\bar{y})|}{|\widetilde{g}'(\bar{x})|} &= \frac{\left|\widetilde{F}''_{\Delta\bar{y}}(\bar{y} - \widetilde{g}(\bar{x}))\right|}{|F''_{\Delta\bar{x}}(\bar{x} - h(\bar{y}))|} = \frac{\exp(A_{\Delta\bar{y}}(\bar{y} - \widetilde{g}(\bar{x})) + B_{\Delta\bar{y}})}{\exp(A_{\Delta\bar{x}}(\bar{x} - h(\bar{y})) + B_{\Delta\bar{x}})} \\ &= \frac{\exp(A_{\Delta\bar{x}}h(\bar{y}) + A_{\Delta\bar{y}}\bar{y} + B_{\Delta\bar{y}})}{\exp(A_{\Delta\bar{y}}\widetilde{g}(\bar{x}) + A_{\Delta\bar{x}}\bar{x} + B_{\Delta\bar{x}})}\end{aligned}$$

Rearranging, we must have

$$\frac{|h'(\bar{y})|}{\exp(A_{\Delta\bar{x}}h(\bar{y}) + A_{\Delta\bar{y}}\bar{y} + B_{\Delta\bar{y}})} = \frac{|\widetilde{g}'(\bar{x})|}{\exp(A_{\Delta\bar{y}}\widetilde{g}(\bar{x}) + A_{\Delta\bar{x}}\bar{x} + B_{\Delta\bar{x}})}$$

and each side must be equal to the same constant (say, $-A_{hg}$) since they depend on different variables. The solution to the differential equation

$$h'(\bar{y}) = \pm A_{hg}\exp(A_{\Delta\bar{x}}h(\bar{y}) + A_{\Delta\bar{y}}\bar{y} + B_{\Delta\bar{y}}) \tag{A.29}$$

is

$$h(y) = -\frac{B_{\Delta\bar{y}}}{A_{\Delta\bar{x}}} - \frac{1}{A_{\Delta\bar{x}}}\ln\left(\pm\frac{A_{\Delta\bar{x}}A_{hg}}{A_{\Delta\bar{y}}}\left(e^{A_{\Delta\bar{y}}\bar{y}} + C_{1\Delta\bar{y}}A_{\Delta\bar{y}}\right)\right), \tag{A.30}$$

where $C_{1\Delta\bar{y}}$ is a constant. (This can be shown by substitution of (A.30) into (A.29) and by noting that any initial condition $h(0)$ can be accommodated by adjusting $C_{1\Delta\bar{y}}$.) Similarly,

$$\widetilde{g}'(\bar{x}) = \pm A_{hg}\exp(A_{\Delta\bar{y}}\widetilde{g}(\bar{x}) + A_{\Delta\bar{x}}\bar{x} + B_{\Delta\bar{x}})$$

and

$$\widetilde{g}(\bar{x}) = -\frac{B_{\Delta\bar{x}}}{A_{\Delta\bar{y}}} - \frac{1}{A_{\Delta\bar{y}}}\ln\left(\pm\frac{A_{\Delta\bar{y}}A_{hg}}{A_{\Delta\bar{x}}}\left(e^{A_{\Delta\bar{x}}\bar{x}} + C_{1\Delta\bar{x}}A_{\Delta\bar{x}}\right)\right) \tag{A.31}$$

where $C_{1\Delta\bar{x}}$ is a constant. From Equations (A.17), (A.28) (A.30) and (A.31), we have

$$\begin{aligned}\frac{f_{\bar{y}}(\bar{y})}{f_{\bar{x}}(\bar{x})} &= \frac{\widetilde{f}_{\Delta\bar{y}}(\bar{y} - \widetilde{g}(\bar{x}))}{f_{\Delta\bar{x}}(\bar{x} - h(\bar{y}))} = \frac{\exp\left(-A_{\Delta\bar{y}}^{-2}\exp(A_{\Delta\bar{y}}(\bar{y} - \widetilde{g}(\bar{x})) + B_{\Delta\bar{y}}) + C_{\Delta\bar{y}}(\bar{y} - \widetilde{g}(\bar{x})) + D_{\Delta\bar{y}}\right)}{\exp\left(-A_{\Delta\bar{x}}^{-2}\exp(A_{\Delta\bar{x}}(\bar{x} - h(\bar{y})) + B_{\Delta\bar{x}}) + C_{\Delta\bar{x}}(\bar{x} - h(\bar{y})) + D_{\Delta\bar{x}}\right)} \\ &= \frac{\exp\left(A_{\Delta\bar{y}}^{-2}\exp(A_{\Delta\bar{y}}\bar{y} + B_{\Delta\bar{x}} + B_{\Delta\bar{y}})\left(\frac{\pm A_{\Delta\bar{y}}A_{hg}}{A_{\Delta\bar{x}}}\left(e^{A_{\Delta\bar{x}}\bar{x}} + C_{1\Delta\bar{x}}A_{\Delta\bar{x}}\right)\right) + C_{\Delta\bar{y}}(\bar{y} - \widetilde{g}(\bar{x})) + D_{\Delta\bar{y}}\right)}{\exp\left(A_{\Delta\bar{x}}^{-2}\exp(A_{\Delta\bar{x}}\bar{x} + B_{\Delta\bar{y}} + B_{\Delta\bar{x}})\left(\frac{\pm A_{\Delta\bar{x}}A_{hg}}{A_{\Delta\bar{y}}}\left(e^{A_{\Delta\bar{y}}\bar{y}} + C_{1\Delta\bar{y}}A_{\Delta\bar{y}}\right)\right) + C_{\Delta\bar{x}}(\bar{x} - h(\bar{y})) + D_{\Delta\bar{x}}\right)}\end{aligned}$$

$$
\begin{aligned}
&= \frac{\exp\left(\exp\left(B_{\Delta\bar{x}}+B_{\Delta\bar{y}}\right)\frac{\pm A_{hg}}{A_{\Delta\bar{y}}A_{\Delta\bar{x}}}\exp\left(A_{\Delta\bar{y}}\bar{y}\right)\exp\left(A_{\Delta\bar{x}}\bar{x}\right)\right)}{\exp\left(\exp\left(B_{\Delta\bar{x}}+B_{\Delta\bar{y}}\right)\frac{\pm A_{hg}}{A_{\Delta\bar{x}}A_{\Delta\bar{y}}}\exp\left(A_{\Delta\bar{y}}\bar{y}\right)\exp\left(A_{\Delta\bar{x}}\bar{x}\right)\right)}\times \\
&\times\frac{\exp\left(A_{\Delta\bar{y}}^{-2}\exp\left(B_{\Delta\bar{x}}+B_{\Delta\bar{y}}\right)\frac{\pm A_{\Delta\bar{y}}A_{hg}}{A_{\Delta\bar{x}}}\exp\left(A_{\Delta\bar{y}}\bar{y}\right)\left(C_{1\Delta\bar{x}}A_{\Delta\bar{x}}\right)+C_{\Delta\bar{y}}\left(\bar{y}-\widetilde{g}\left(\bar{x}\right)\right)+D_{\Delta\bar{y}}\right)}{\exp\left(A_{\Delta\bar{x}}^{-2}\exp\left(B_{\Delta\bar{x}}+B_{\Delta\bar{y}}\right)\frac{\pm A_{\Delta\bar{x}}A_{hg}}{A_{\Delta\bar{y}}}\exp\left(A_{\Delta\bar{x}}\bar{x}\right)\left(C_{1\Delta\bar{y}}A_{\Delta\bar{y}}\right)+C_{\Delta\bar{x}}\left(\bar{x}-h\left(\bar{y}\right)\right)+D_{\Delta\bar{x}}\right)} \\
&= \frac{\exp\left(\exp\left(B_{\Delta\bar{x}}+B_{\Delta\bar{y}}\right)\frac{\pm A_{hg}C_{1\Delta\bar{x}}}{A_{\Delta\bar{y}}}\exp\left(A_{\Delta\bar{y}}\bar{y}\right)+C_{\Delta\bar{y}}\bar{y}+D_{\Delta\bar{y}}\right)\exp\left(C_{\Delta\bar{x}}h\left(\bar{y}\right)\right)}{\exp\left(\exp\left(B_{\Delta\bar{x}}+B_{\Delta\bar{y}}\right)\frac{\pm A_{hg}C_{1\Delta\bar{y}}}{A_{\Delta\bar{x}}}\exp\left(A_{\Delta\bar{x}}\bar{x}\right)+C_{\Delta\bar{x}}\bar{x}+D_{\Delta\bar{x}}\right)\exp\left(C_{\Delta\bar{y}}\widetilde{g}\left(\bar{x}\right)\right)} \\
&= \frac{\exp\left(\exp\left(B_{\Delta\bar{x}}+B_{\Delta\bar{y}}\right)\frac{\pm A_{hg}C_{1\Delta\bar{x}}}{A_{\Delta\bar{y}}}\exp\left(A_{\Delta\bar{y}}\bar{y}\right)+C_{\Delta\bar{y}}\bar{y}+D_{\Delta\bar{y}}\right)}{\exp\left(\exp\left(B_{\Delta\bar{x}}+B_{\Delta\bar{y}}\right)\frac{\pm A_{hg}C_{1\Delta\bar{y}}}{A_{\Delta\bar{x}}}\exp\left(A_{\Delta\bar{x}}\bar{x}\right)+C_{\Delta\bar{x}}\bar{x}+D_{\Delta\bar{x}}\right)}\times \\
&\times\frac{\exp\left(-\frac{C_{\Delta\bar{x}}B_{\Delta\bar{y}}}{A_{\Delta\bar{x}}}\right)\left(\frac{\pm A_{\Delta\bar{x}}A_{hg}}{A_{\Delta\bar{y}}}\right)^{-\frac{C_{\Delta\bar{x}}}{A_{\Delta\bar{x}}}}\left(e^{A_{\Delta\bar{y}}\bar{y}}+C_{1\Delta\bar{y}}A_{\Delta\bar{y}}\right)^{-\frac{C_{\Delta\bar{x}}}{A_{\Delta\bar{x}}}}}{\exp\left(-\frac{C_{\Delta\bar{y}}B_{\Delta\bar{x}}}{A_{\Delta\bar{y}}}\right)\left(\frac{\pm A_{\Delta\bar{y}}A_{hg}}{A_{\Delta\bar{x}}}\right)^{-\frac{C_{\Delta\bar{y}}}{A_{\Delta\bar{y}}}}\left(e^{A_{\Delta\bar{x}}\bar{x}}+C_{1\Delta\bar{x}}A_{\Delta\bar{x}}\right)^{-\frac{C_{\Delta\bar{y}}}{A_{\Delta\bar{y}}}}},
\end{aligned}
$$

implying that

$$
\begin{aligned}
f_{\bar{y}}\left(\bar{y}\right) &= A_{n\Delta\bar{y}}\exp\left(\exp\left(B_{\Delta\bar{x}}+B_{\Delta\bar{y}}\right)\frac{\pm A_{hg}C_{1\Delta\bar{x}}}{A_{\Delta\bar{y}}}\exp\left(A_{\Delta\bar{y}}\bar{y}\right)+C_{\Delta\bar{y}}\bar{y}\right)\left(e^{A_{\Delta\bar{y}}\bar{y}}+C_{1\Delta\bar{y}}A_{\Delta\bar{y}}\right)^{-\frac{C_{\Delta\bar{x}}}{A_{\Delta\bar{x}}}} \\
f_{\bar{x}}\left(\bar{x}\right) &= A_{n\Delta\bar{x}}\exp\left(\exp\left(B_{\Delta\bar{x}}+B_{\Delta\bar{y}}\right)\frac{\pm A_{hg}C_{1\Delta\bar{y}}}{A_{\Delta\bar{x}}}\exp\left(A_{\Delta\bar{x}}\bar{x}\right)+C_{\Delta\bar{x}}\bar{x}\right)\left(e^{A_{\Delta\bar{x}}\bar{x}}+C_{1\Delta\bar{x}}A_{\Delta\bar{x}}\right)^{-\frac{C_{\Delta\bar{y}}}{A_{\Delta\bar{y}}}}.
\end{aligned}
$$

where the constants $A_{n\Delta\bar{y}}$ and $A_{n\Delta\bar{x}}$ incorporate any prefactor that would have cancelled in the ratio $f_{\bar{y}}\left(\bar{y}\right)/f_{\bar{x}}\left(\bar{x}\right)$ as well as the constants $\exp\left(D_{\Delta\bar{y}}\right)\exp\left(-C_{\Delta\bar{x}}B_{\Delta\bar{y}}/A_{\Delta\bar{x}}\right)\left(\pm A_{\Delta\bar{x}}A_{hg}/A_{\Delta\bar{y}}\right)^{-\frac{C_{\Delta\bar{x}}}{A_{\Delta\bar{x}}}}$ and $\exp\left(D_{\Delta\bar{x}}\right)\exp\left(-C_{\Delta\bar{y}}B_{\Delta\bar{x}}/A_{\Delta\bar{y}}\right)\left(\pm A_{\Delta\bar{y}}A_{hg}/A_{\Delta\bar{x}}\right)^{-\frac{C_{\Delta\bar{y}}}{A_{\Delta\bar{y}}}}$, respectively. The constants $A_{n\Delta\bar{y}}$ and $A_{n\Delta\bar{x}}$ are determined by the fact that these densities must integrate to 1. It can be readily, albeit tediously, verified that it is possible to set the signs of all constants so as to obtain valid densities for all variables. Hence, we have found one special case where Model 1 is not identified. This is Case 2 in the statement of Theorem 1.

Case C: In the special case where $A=0$ in Equation (A.26) (not included in Case B above), we let $B_2=\exp\left(B\right)$ and write, for $F=F_{\Delta x},\widetilde{F}_{\Delta y}$:

$$
\begin{aligned}
F''\left(u\right) &= B_2 \\
F\left(u\right) &= B_2u^2+Cu+D
\end{aligned}
$$

for some constants B_2, C, D (that differ for $F_{\Delta x}$ and $\widetilde{F}_{\Delta y}$) to conclude that $f\left(u\right)$ is a normal and therefore that $\Delta\bar{x}$ and $\Delta\bar{y}$ are normally distributed. Since under Model

A.3 $\Delta\bar{x}$ is a factor of Δx and, under model A.4, $\Delta\bar{y}$ is a factor of Δy, we conclude that either Δx must have a normal factor or Δy must have a normal factor. Next,

$$\frac{|h'(\bar{y})|}{|\widetilde{g}'(\bar{x})|} = \frac{\left|\widetilde{F}''_{\Delta\bar{y}}(\bar{y} - \widetilde{g}(\bar{x}))\right|}{|F''_{\Delta\bar{x}}(\bar{x} - h(\bar{y}))|} = B_3$$

where B_3 is the ratio of the constants B_2 obtained for $F_{\Delta x}$ and $\widetilde{F}_{\Delta y}$. Rearranging, we obtain

$$|h'(\bar{y})| = B_3 |\widetilde{g}'(\bar{x})|$$

and it follows that $h'(\bar{y})$ and $\widetilde{g}'(\bar{x})$ must be constant, i.e., that $h(\bar{y})$ and $\widetilde{g}(\bar{x})$ are linear. From $\frac{f_{\bar{y}}(\bar{y})}{f_{\bar{x}}(\bar{x})} = \frac{\widetilde{f}_{\Delta\bar{y}}(\bar{y}-\widetilde{g}(\bar{x}))}{f_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))}$, we can show that $f_{\bar{y}}(\bar{y})$ and $f_{\bar{x}}(\bar{x})$ must also be normal. Either Model A.3 or A.4 then implies that x^* must be normal. So we recover the more familiar unidentified Case 3 in the statement of Theorem 1.

Cases A, B and C are mutually exclusive and have explored every possible solutions to our differential equations that could have possibly lead to two distinct observationally equivalent models. We can thus conclude that when the unidentified Cases B and C do not apply, the model is identified, which corresponds to Case 1 in the statement of Theorem 1. (Case 1 in the statement of Theorem 1 includes not only the identified Case A above, but also all "dead-ends" where the construction of two distinct observationally equivalent models failed for a reason or another.)

When $f_{x^*}(x^*)$ and $g(x^*)$ are identified the distributions of Δx and Δy are also identified. Indeed, $E\left[e^{i\xi\Delta x}\right] = E\left[e^{i\xi x}\right]/E\left[e^{i\xi x^*}\right]$ where $E\left[e^{i\xi x^*}\right]$ is nonzero for all ξ in a dense subset of $\mathbb{R}$ by Assumption 3. Since $E\left[e^{i\xi\Delta x}\right]$ is continuous in ξ, $E\left[e^{i\xi\Delta x}\right]$ can be recovered for all ξ by a limiting process. The distribution of Δx is then uniquely determined by its c.f. $E\left[e^{i\xi\Delta x}\right]$. Similarly, identification of the distribution of Δy follows from the identity $E\left[e^{i\gamma\Delta y}\right] = E\left[e^{i\gamma y}\right]/E\left[e^{i\gamma g(x^*)}\right]$.

B Specific example

This section provides independent verification of Case 2 of Theorem 1 by showing that a specific density of the stated form can be generated by two different models, one with only errors along y and one with only errors along x. Examples with errors along both directions can be constructed by simply adding the same error terms in the two equivalent models.

Consider the joint density:

$$f_{xy}(x,y) = C\exp(x)\exp(-y)\exp(-e^x)\exp\left(-e^{-y}\right)\exp\left(-e^x e^{-y}\right)$$

where C in a normalization constant equal to $(e\,\mathrm{Ei}(1))^{-1}$ with $\mathrm{Ei}(t) = \int_t^\infty \frac{e^{-u}}{u}du$ (known as the "exponential integral"). In particular, $\int_1^\infty \frac{e^{-u}}{u}du \approx 0.21938$. By direct

substitution, one can verify that $f_{xy}(x,y)$ can be written in two ways:

$$\begin{aligned} f_{xy}(x,y) &= \widetilde{f}_{\Delta y}\left(y-\widetilde{g}(x)\right)\widetilde{f}_x(x) \\ &= f_{\Delta x}\left(x-h(y)\right)f_y(y) \end{aligned}$$

where

$$\begin{aligned} \widetilde{f}_{\Delta y}(v) &= \exp\left(-e^{-v+\gamma}-v+\gamma\right) \\ \widetilde{g}(x) &= \gamma+\ln\left(e^x+1\right) \\ \widetilde{f}_x(x) &= C\frac{\exp\left(-e^x+x\right)}{1+e^x} \end{aligned}$$

and

$$\begin{aligned} f_{\Delta x}(u) &= \exp\left(-e^{u-\gamma}+u-\gamma\right) \\ h(y) &= -\gamma-\ln\left(e^{-y}+1\right) \\ &\Longrightarrow \quad g(x)=h^{-1}(x)=-\gamma-\ln\left(e^{-x}-e^{\gamma}\right) \\ f_y(y) &= C\frac{\exp\left(-e^{-y}-y\right)}{1+e^{-y}} \end{aligned}$$

where $\gamma=-\int_0^\infty \ln(u)\,e^{-u}du\approx 0.57722$ (known as the Euler-Mascheroni constant).

Without performing detailed simulations it is clear how an estimator would behave in this case. The likelihood function has, in the limit, two disjoint maxima of equal value. However, for any distribution "close" to this unidentified case, the two maxima would adopt slightly different values. Hence, the maximum likelihood estimator based on a finite sample would find either one of these maxima at random and would never converge to a single value of the parameter asymptotically.

C Sieve Maximum Likelihood: Estimation and Asymptotics

In this section, we consider the estimation of a parametric regression model: $y=m(x^*;\theta_0)+\Delta y$, where the function $m()$ is known and Δy is independent of x^*. We assume $\theta_0\in\Theta$, which is a compact subset of $\mathbb{R}^{d_\theta}$. Let $\{D_i\equiv(y_i,x_i)\}_{i=1}^n$ denote a random sample of $D\equiv(y,x)$, where $x=x^*+\Delta x$ and Δx is independent of x^*. We have shown that m, $f_{\Delta y}$, f_v, and f_{x^*} are identified from $f_{y,x}$. Let $\alpha_0\equiv(\theta_0,f_{01},f_{02},f_{03})^T\equiv(\theta_0,f_{\Delta y},f_v,f_{x^*})^T$ be the true values of the nuisance parameters. Our sieve ML estimator $\widehat{\alpha}$ for α_0 is based on the following equation

$$f(y,x;\alpha_0)=\int f_{\Delta y}\left(y-m(x^*;\theta_0)\right)f_v\left(x-x^*\right)f_{x^*}\left(x^*\right)dx^*.$$

We start with imposing some smoothness restrictions on the unknown functions α_0 in the Hölder space. Let $\xi = (\xi_1, ..., \xi_d)^T \in \mathbb{R}^d$, $a = (a_1, ..., a_d)^T$ be a vector of non-negative integers, and

$$\nabla^a g(\xi) \equiv \frac{\partial^{a_1+\cdots+a_d} g(\xi_1, ..., \xi_d)}{\partial \xi_1^{a_1} \cdots \partial \xi_d^{a_d}}$$

denote the $(a_1 + \cdots + a_d)$ -th derivative. Let $\|\cdot\|_E$ denote the Euclidean norm. Let $\underline{\gamma}$ be the largest integer satisfying $\gamma > \underline{\gamma}$. The Hölder space $\Lambda^\gamma(\mathcal{V})$ of order $\gamma > 0$ with $\mathcal{V} \subseteq \mathbb{R}^d$ is a space of functions $g : \mathcal{V} \mapsto \mathbb{R}$ such that the first $\underline{\gamma}$ derivatives are continuous and bounded, and the $\underline{\gamma}$-th derivative are Hölder continuous with the exponent $\gamma - \underline{\gamma} \in (0, 1]$. The Hölder norm is defined as:

$$\|g\|_{\Lambda^\gamma} = \sup_{\xi \in \mathcal{V}} |g(\xi)| + \max_{a_1+\cdots+a_d=\underline{\gamma}} \sup_{\xi \neq \xi' \in \mathcal{V}} \frac{|\nabla^a g(\xi) - \nabla^a g(\xi')|}{\left(\|\xi - \xi'\|_E\right)^{\gamma-\underline{\gamma}}},$$

while a Hölder ball is defined as $\Lambda_c^\gamma(\mathcal{V}) \equiv \{g \in \Lambda^\gamma(\mathcal{V}) : \|g\|_{\Lambda^\gamma} \leq c < \infty\}$. We also define

$$\mathcal{F}_1^* = \left\{ \sqrt{f_1(\cdot)} \in \Lambda_c^{\gamma_1}(\mathbb{R}) : f_1(\cdot) > 0, \ \int_{\mathbb{R}} f_1(\Delta y) d\Delta y = 1 \right\},$$

$$\mathcal{F}_2^* = \left\{ \sqrt{f_2(\cdot)} \in \Lambda_c^{\gamma_2}(\mathbb{R}) : f_2(\cdot) > 0, \ \int_{\mathbb{R}} f_2(\Delta x) d\Delta x = 1 \right\},$$

$$\mathcal{F}_3^* = \left\{ \sqrt{f_3(\cdot)} \in \Lambda_c^{\gamma_3}(\mathbb{R}) : f_3(\cdot) > 0, \ \int_{\mathbb{R}} f_3(x^*) dx^* = 1 \right\}.$$

It is useful to introduce restricted subsets $\mathcal{F}_i \subseteq \mathcal{F}_i^*$, $i = 1, 2, 3$ to obtain an asymptotic treatment that allows for additional user-specified restrictions on the spaces $\mathcal{F}_i^*$. This extension serves two purposes. First, some applications may suggest plausible constraints, such as monotonicity, which may help reduce the estimation errors. Second, it is possible to find settings in which the semiparametric efficiency bound for this estimation problem may be degenerate (i.e. root n consistent estimation of a certain semiparametric functional is not possible) without additional constraints on the spaces $\mathcal{F}_i^*$. Introducing constrained spaces $\mathcal{F}_i$ may restore root n consistency in such cases if suitable constraints can be found. For instance, in the related context of measurement error estimation in the presence of repeated measurements (Schennach (2004)), it is known that root n consistent estimation is only possible under a suitable balance between smoothness of certain quantities and *lack-of-smoothness* of others.

It should be noted that our assumptions (in Section C.2 below) securing asymptotic normality and root n-consistency of a semiparametric estimator are not vacuous: There must exist cases where the semiparametric efficiency bound for our model is not infinite. For instance, in a linear model with non-normal regressors, the slope coefficient is given by

$$\frac{E\left[y^2 x\right] - E\left[y^2\right] E\left[x\right]}{E\left[y x^2\right] - E\left[y\right] E\left[x^2\right]} \tag{C.1}$$

provided $E[yx^2] - E[y]E[x^2] \neq 0$. Equation (C.1) suggests a natural root n-consistent estimator, upon substitution of the expectations by their sample counterparts, under standard regularity conditions. This method-of-moment estimator is obviously regular and by Theorem 2.1 in Newey (1990), it follows that its asymptotic variance places an upper bound on the semiparametric efficiency bound.

We assume that the functions $f_{\Delta y}$, $f_{\Delta x}$, and f_{x^*} satisfy the following smoothness restrictions:

Assumption C.1 *(i) Let Assumptions 1-6 hold; (ii) $f_{\Delta y}(\cdot) \in \mathcal{F}_1$ with $\gamma_1 > 1/2$; (iii) $f_{\Delta x}(\cdot) \in \mathcal{F}_2$ with $\gamma_2 > 1/2$; (iv) $f_{x^*}(\cdot) \in \mathcal{F}_3$ with $\gamma_3 > 1/2$;.*

Denote $\mathcal{A} = \Theta \times \mathcal{F}_1 \times \mathcal{F}_2 \times \mathcal{F}_3$ and $\alpha = (\theta, f_1, f_2, f_3)^T$. Let $E[\cdot]$ denote the expectation with respect to the data generating process for D_i. Then,

$$\alpha_0 = \arg\max_{\alpha \in \mathcal{A}} E\left[\ln \int f_1(y - m(x^*; \theta_0)) f_2(x - x^*) f_3(x^*) dx^*\right]. \tag{C.2}$$

Let $\mathcal{A}_n = \Theta \times \mathcal{F}_1^n \times \mathcal{F}_2^n \times \mathcal{F}_3^n$ be a sieve space for $\mathcal{A}$, which is a sequence of approximating spaces that are dense in $\mathcal{A}$ under some pseudo-metric. The sieve MLE $\widehat{\alpha}_n = \left(\widehat{f}_1, \widehat{f}_2, \widehat{f}_3\right)^T \in \mathcal{A}_n$ for $\alpha_0 \in \mathcal{A}$ is:

$$\widehat{\alpha}_n = \arg\max_{\alpha \in \mathcal{A}_n} \sum_{i=1}^{n} \left[\ln \int f_1(y_i - m(x^*; \theta_0)) f_2(x_i - x^*) f_3(x^*) dx^*\right]. \tag{C.3}$$

$\widehat{\alpha}_n = \arg\max_{\alpha \in \mathcal{A}_n} \sum_{t=1}^{n} \ell(Z_t; \alpha)$. Here we present a finite dimensional sieve $\mathcal{F}_1^n \times \mathcal{F}_2^n \times \mathcal{F}_3^n$. For $j = 1, 2, 3$, let $p^{k_{j,n}}(\cdot)$ be a $k_{j,n} \times 1-$vector of known basis functions, such as Fourier series, power series, splines, etc. Then we denote the sieve space for $\mathcal{F}_j, j = 1, 2, 3$ as follows:

$$\mathcal{F}_1^n = \left\{\sqrt{f_1(\cdot)} = p^{k_{1,n}}(\cdot)^T \beta_1 \in \mathcal{F}_1\right\},$$

$$\mathcal{F}_2^n = \left\{\sqrt{f_2(\cdot)} = p^{k_{2,n}}(\cdot)^T \beta_2 \in \mathcal{F}_2\right\},$$

$$\mathcal{F}_3^n = \left\{\sqrt{f_3(\cdot)} = p^{k_{3,n}}(\cdot)^T \beta_3 \in \mathcal{F}_3\right\}.$$

Let $k_n = \min\{k_{1,n}, k_{2,n}, k_{3,n}\}$. We also define the projection of the true value α_0 onto the space $\mathcal{A}_n$ associated with k_n:

$$\Pi_n \alpha \equiv \alpha_n \equiv \underset{\alpha_n = (\theta, f_1, f_2, f_3)^T \in \mathcal{A}_n}{\arg\max} E\left(\ln\left[\int f_1(y_i - m(x^*; \theta_0)) f_2(x_i - x^*) f_3(x^*) dx^*\right]\right).$$

C.1 Consistency

First we define a norm on $\mathcal{A}$ as follows:

$$\|\alpha\|_s = \|\theta\|_E + \sup_{\Delta y} |f_1(\Delta y)\omega(\Delta y)| + \sup_{\Delta x} |f_2(\Delta x)\omega(\Delta x)| + \sup_{x^*} |f_3(x^*)\omega(x^*)|$$

with $\omega(\xi) = \left(1+\xi^2\right)^{-\zeta/2}$ for some $\zeta > 0$. Define

$$\ell(D_i;\alpha) = \int f_1(y_i - m(x^*;\theta)) f_2(x_i - x^*) f_3(x^*) dx^*.$$

We make the following assumptions:

Assumption C.2 *i) the data $\{(y_i, x_i)_{i=1}^n\}$ are i.i.d.; ii) the density of $D \equiv (y, x)$, f_D, satisfies $\int \omega(D)^{-2} f_D(D) dD < \infty$.*

Assumption C.3 *i) $\theta_0 \in \Theta$, a compact subset of $\mathbb{R}^{d_\theta}$; ii) Assumption C.1 holds for (θ, f_1, f_2, f_3) in a neighborhood of α_0 (in the norm $\|\cdot\|_s$).*

Assumption C.4 *(i) $E[\ell(D_i;\alpha_0)^2]$ is bounded; (ii) there are a finite $\kappa > 0$ and a random variable $U(D_i)$ with $E[\{U(D_i)\}^2] < \infty$ such that $\sup_{\alpha\in\mathcal{A}_n:\|\alpha-\alpha_0\|_s\le\delta} |\ell(D_i;\alpha) - \ell(D_i;\alpha_0)| \le \delta^\kappa U(D_i)$.*

Assumption C.5 *$\|\Pi_n\alpha_0 - \alpha_0\|_s = o(1)$ (as $k_n \to \infty$) and $k_n/n \to 0$.*

As shown in Hu and Schennach (2008), we summarize the consistency result with its proof omitted.

Lemma C.1 *Under assumptions C.1, C.2, C.3, C.4, and C.5, we have $\|\widehat{\alpha}_n - \alpha_0\|_s = o_p(1)$.*

Given the consistency, we then need to establish convergence at the rate $o_p\left(n^{-1/4}\right)$ in a suitable norm in order to proceed towards our main semiparametric asymptotic normality and root n consistency result. In order to achieve this convergence rate under relatively weak assumptions, we employ a device introduced by Ai and Chen (2003) and employ a weaker norm $\|\cdot\|$, under which $o_p\left(n^{-1/4}\right)$ convergence is easier to establish.

Consider $\alpha_1, \alpha_2 \in \mathcal{A}$, and assume the existence of a continuous path $\{\alpha(\tau) : \tau \in [0,1]\}$ in $\mathcal{A}$ such that $\alpha(0) = \alpha_1$ and $\alpha(1) = \alpha_2$. If $\ln f_{y,x}(D, (1-\tau)\alpha_0 + \tau\alpha)$ is continuously differentiable at $\tau = 0$ for almost all D and any $\alpha \in \mathcal{A}$, the pathwise derivative of $\ln f_{y,x}(D,\alpha_0)$ at α_0 evaluated at $\alpha - \alpha_0$ can be defined as

$$\frac{d\ln\ell(D,\alpha_0)}{d\alpha}[\alpha - \alpha_0] \equiv \left.\frac{d\ln\ell(D,(1-\tau)\alpha_0 + \tau\alpha)}{d\tau}\right|_{\tau=0} \tag{C.4}$$

almost everywhere (under the probability measure of D). In our setting, the pathwise derivative at α_0 is as follows:

$$\begin{aligned} & \frac{d\ln \ell(D,\alpha_0)}{d\alpha}\left[\alpha-\alpha_0\right] \qquad (C.5)\\ = \; & \frac{1}{f_{y,x}(D,\alpha_0)}\left\{\int f'_{\Delta y}(y-m(x^*;\theta_0))\frac{dm(x^*;\theta_0)}{d\theta}\left[\theta-\theta_0\right]f_{\Delta x}(x-x^*)f_{x^*}(x^*)dx^*+\right.\\ & +\int\left[f_1(y-m(x^*;\theta_0))-f_{\Delta y}(y-m(x^*;\theta_0))\right]f_{\Delta x}(x-x^*)f_{x^*}(x^*)dx^*+\\ & +\int f_{\Delta y}(y-m(x^*;\theta_0))\left[f_2(x-x^*)-f_{\Delta x}(x-x^*)\right]f_{x^*}(x^*)dx^*+\\ & \left.+\int f_{\Delta y}(y-m(x^*;\theta_0))f_{\Delta x}(x-x^*)\left[f_3(x^*)-f_{x^*}(x^*)\right]dx^*\right\}. \end{aligned}$$

We use the Fisher norm $\|\cdot\|$ defined as

$$\|\alpha_1-\alpha_2\|\equiv\sqrt{E\left\{\left(\frac{d\ln \ell(D,\alpha_0)}{d\alpha}\left[\alpha_1-\alpha_2\right]\right)^2\right\}} \qquad (C.6)$$

for any α_1, $\alpha_2\in\mathcal{A}$. In order to establish the asymptotic normality of $\widehat{\theta}_n$, one typically needs that $\widehat{\alpha}_n$ converges to α_0 at a rate faster than $n^{-1/4}$. We need the following assumptions to obtain this rate of convergence:

Assumption C.6 $\|\Pi_n\alpha_0-\alpha_0\|=O\left(k_n^{-\gamma/2}\right)=o\left(n^{-1/4}\right)$ *with* $\gamma\equiv\min\{\gamma_1,\ \gamma_2,\ \gamma_3\}>1/2$.

Assumption C.7 *i) there exists a measurable function* $c(D)$ *with* $E\left\{c(D)^4\right\}<\infty$ *such that* $|\ln \ell(D;\alpha)|\leq c(D)$ *for all* D *and* $\alpha\in\mathcal{A}_n$*; ii)* $\ln \ell(D;\alpha)\in\Lambda_c^{\gamma,\omega}(\mathcal{Y}\times\mathcal{X})$ *for some constant* $c>0$ *with* $\gamma>d_D/2$*, for all* $\alpha\in\mathcal{A}_n$*, where* d_D *is the dimension of* D.

Assumption C.8 $\mathcal{A}$ *is convex in* α_0*, and* $m(x^*;\theta)$ *is pathwise differentiable at* θ_0.

Assumption C.9 *For some* $c_1,c_2>0$*,*

$$c_1E\left(\ln\frac{\ell(D;\alpha_0)}{\ell(D;\alpha)}\right)\leq\|\alpha-\alpha_0\|^2\leq c_2E\left(\ln\frac{\ell(D;\alpha_0)}{\ell(D;\alpha)}\right). \qquad (C.7)$$

holds for all $\alpha\in\mathcal{A}_n$ *with* $\|\alpha-\alpha_0\|_s=o(1)$.

Assumption C.10 $\left(k_nn^{-1/2}\ln n\right)\sup_{\xi_1\in\mathbb{R}}\left\|p^{k_n}\left(\xi_1\right)\right\|_E^2=o\left(1\right)$.

Assumption C.11 $\ln N\left(\varepsilon,\mathcal{A}_n\right)=O\left(k_n\ln\left(k_n/\varepsilon\right)\right)$ *where* $N\left(\varepsilon,\mathcal{A}_n\right)$ *is the minimum number of balls (in the* $\|\cdot\|_s$ *norm) needed to cover the set* $\mathcal{A}_n$.

The discussion of these assumptions can be found in Hu and Schennach (2008). The following convergence rate theorem is a direct application of Theorem 3.1 in Ai and Chen (2003).

Theorem 1 *Under assumptions C.1-C.11, we have* $\|\widehat{\alpha}_n-\alpha_0\|=o_p(n^{-1/4})$.

C.2 Asymptotic normality

We follow the semiparametric MLE framework of Shen (1997) to show the asymptotic normality of the estimator $\widehat{\theta}_n$. We define the inner product

$$\langle v_1, v_2 \rangle = E\left\{ \left(\frac{d \ln \ell(D, \alpha_0)}{d\alpha} [v_1] \right) \left(\frac{d \ln \ell(D, \alpha_0)}{d\alpha} [v_2] \right) \right\}. \tag{C.8}$$

Let $\overline{\mathbf{V}}$ denote the closure of the linear span of $\mathcal{A} - \{\alpha_0\}$ under the norm $\|\cdot\|$ (i.e., $\overline{\mathbf{V}} = \mathbb{R}^{d_\theta} \times \overline{\mathcal{W}}$ with $\overline{\mathcal{W}} \equiv \overline{\mathcal{F}_1 \times \mathcal{F}_2 \times \mathcal{F}_3} - \left\{ \left(f_{\Delta y}, f_{\Delta x}, f_{x^*} \right)^T \right\}$) and define the Hilbert space $\left(\overline{\mathbf{V}}, \langle \cdot, \cdot \rangle \right)$ with its inner product defined in Equation (C.8).

As shown above, we have

$$\begin{aligned} \frac{d \ln \ell(D, \alpha_0)}{d\alpha} [\alpha - \alpha_0] &= \frac{d \ln \ell(D, \alpha_0)}{d\theta} [\theta - \theta_0] + \frac{d \ln \ell(D, \alpha_0)}{df_1} [f_1 - f_{\Delta y}] + \\ &+ \frac{d \ln \ell(D, \alpha_0)}{df_2} [f_2 - f_{\Delta x}] + \frac{d \ln \ell(D, \alpha_0)}{df_3} [f_3 - f_{x^*}]. \end{aligned} \tag{C.9}$$

For each component θ_j of θ, $j = 1, 2, ..., d_\theta$, we define $w_j^* \in \overline{\mathcal{W}}$ as follows:

$$\begin{aligned} w_j^* &\equiv \left(f_{1j}^*, f_{2j}^*, f_{3j}^* \right)^T \\ &= \underset{(f_1, f_2, f_3)^T \in \overline{\mathcal{W}}}{\arg\min} E\left\{ \left(\frac{d \ln \ell(D, \alpha_0)}{d\theta_j} - \frac{d \ln \ell(D, \alpha_0)}{df_1} [f_1] + \right. \right. \\ &\left. \left. - \frac{d \ln \ell(D, \alpha_0)}{df_2} [f_2] - \frac{d \ln \ell(D, \alpha_0)}{df_3} [f_3] \right)^2 \right\}. \end{aligned} \tag{C.10}$$

Define $w^* = \left(w_1^*, w_2^*, ..., w_{d_b}^* \right)$,

$$\begin{aligned} \frac{d \ln \ell(D, \alpha_0)}{df} \left[w_j^* \right] &= \frac{d \ln \ell(D, \alpha_0)}{df_1} \left[f_{1j}^* \right] + \frac{d \ln \ell(D, \alpha_0)}{df_2} \left[f_{2j}^* \right] + \\ &+ \frac{d \ln \ell(D, \alpha_0)}{df_3} \left[f_3^* \right], \\ \frac{d \ln \ell(D, \alpha_0)}{df} [w^*] &= \left(\frac{d \ln \ell(D, \alpha_0)}{df} [w_1^*], ..., \frac{d \ln \ell(D, \alpha_0)}{df} \left[w_{d_b}^* \right] \right), \end{aligned} \tag{C.11}$$

and the row vector

$$G_{w^*}(D) = \frac{d \ln \ell(D, \alpha_0)}{d\theta^T} - \frac{d \ln \ell(D, \alpha_0)}{df} [w^*]. \tag{C.12}$$

We want to show that $\widehat{\theta}_n$ has a multivariate normal distribution asymptotically. It is well known that if $\lambda^T \theta$ has a normal distribution for all λ, then θ has a multivariate

normal distribution. Therefore, we consider $\lambda^T\theta$ as a functional of α. Define $s(\alpha) \equiv \lambda^T\theta$ for $\lambda \in \mathbb{R}^{d_b}$ and $\lambda \neq 0$. If $E\left[G_{w^*}(D)^T G_{w^*}(D)\right]$ is finite positive definite, then the function $s(\alpha)$ is bounded, and the Riesz representation theorem implies that there exists a representor v^* such that

$$s(\alpha) - s(\alpha_0) \equiv \lambda^T(\theta - \theta_0) = \langle v^*, \alpha - \alpha_0 \rangle \tag{C.13}$$

for all $\alpha \in \mathcal{A}$. Here, $v^* \equiv \binom{v_\theta^*}{v_f^*}$, $v_\theta^* = J^{-1}\lambda$, $v_f^* = -w^* v_\theta^*$, with $J = E\left[G_{w^*}(D)^T G_{w^*}(D)\right]$. Under suitable assumptions made below, the Riesz representor v^* exists and is bounded.

We denote

$$\frac{d \ln \ell(D,\alpha)}{d\alpha}[v] \equiv \left.\frac{d \ln \ell(D, \alpha + \tau v)}{d\tau}\right|_{\tau=0} \text{ a.s. } D \text{ for any } v \in \overline{\mathbf{V}}. \tag{C.14}$$

For a sieve MLE, we have that

$$\langle v^*, \widehat{\alpha}_n - \alpha_0 \rangle = \frac{1}{n}\sum_{i=1}^n \frac{d \ln \ell(D_i, \alpha_0)}{d\alpha}[v^*] + o_p\left(n^{-1/2}\right) \tag{C.15}$$

Note that $\left(\frac{d \ln f_{y,x}(D,\alpha)}{d\alpha}[v^*]\right) = G_{w^*}(D) J^{-1}\lambda$. Thus, by the classical central limit theorem, the asymptotic distribution of $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)$ is $N\left(0, J^{-1}\right)$. In fact, the matrix J is the efficient information matrix in this semiparametric estimation, under suitable regularity conditions given in Shen (1997).

We now present the sufficient conditions for the $\sqrt{n}-$normality of $\widehat{\theta}_n$. Define

$$\mathcal{N}_{0n} \equiv \left\{\alpha \in \mathcal{A}_n : \|\alpha - \alpha_0\|_s \leq \upsilon_n, \ \ \|\alpha - \alpha_0\| \leq \upsilon_n n^{-1/4})\right\} \tag{C.16}$$

with $\upsilon_n = o(1)$ and $\mathcal{N}_0$ the same way with $\mathcal{A}_n$ replaced by $\mathcal{A}$. Note that $\mathcal{N}_0$ still depends on n. For $\alpha \in \mathcal{N}_{0n}$ we define a local alternative $\alpha^*(\alpha, \varepsilon_n) = (1 - \varepsilon_n)\alpha + \varepsilon_n(v^* + \alpha_0)$ with $\varepsilon_n = o\left(n^{-1/2}\right)$. Let $\Pi_n \alpha^*(\alpha, \varepsilon_n)$ be the projection of $\alpha^*(\alpha, \varepsilon_n)$ onto $\mathcal{A}_n$.

Assumption C.12 *i)* $E\left[G_{w^*}(D)^T G_{w^*}(D)\right]$ *exists, is bounded and is positive-definite; ii)* $\theta_0 \in int(\times)$.

Assumption C.13 *there is a* $U(D)$ *with* $E\{[U(D)]^2\} < \infty$ *and a non-negative measurable function* η *with* $\lim_{\delta \to 0}\eta(\delta) = 0$ *such that for all* $\alpha \in \mathcal{N}_{0n}$,

$$\sup_{\overline{\alpha} \in \mathcal{N}_0}\left|\frac{d^2 \ln \ell(D;\overline{\alpha})}{d\alpha d\alpha^T}[\alpha - \alpha_0, \upsilon_n^*]\right| \leq U(D) \times \eta(||\alpha - \alpha_0||_s).$$

Assumption C.14 *Uniformly over* $\overline{\alpha} \in \mathcal{N}_0$ *and* $\alpha \in \mathcal{N}_{0n}$,

$$E\left(\frac{d^2 \ln \ell(D;\overline{\alpha})}{d\alpha d\alpha^T}[\alpha - \alpha_0, \upsilon_n^*] - \frac{d^2 \ln \ell(D;\alpha_0)}{d\alpha d\alpha^T}[\alpha - \alpha_0, \upsilon_n^*]\right) = o\left(n^{-1/2}\right).$$

Assumption C.15 *There is a* $v_n^* = \begin{pmatrix} v_\theta^* \\ -(\Pi_n w^*) v_\theta^* \end{pmatrix} \in \mathcal{A}_n - \{\Pi_n \alpha_0\}$ *such that* $\|v_n^* - v^*\| = o(n^{-1/4})$.

A detailed discussion of these assumptions can be found in Shen (1997). By theorem 1 in Shen (1997), we show that the estimator for the parametric component θ_0 is $\sqrt{n}$ consistent and asymptotically normally distributed.

Theorem 2 *Under assumptions C.1-C.15,* $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, J^{-1}\right)$ *where* $J = E\left[G_{w^*}(D)^T G_{w^*}(D)\right]$ *for* $G_{w^*}(D)$ *given in Equation (C.12).*

D Additional Simulation Results

D.1 Smoothing parameter selection

This section illustrates that our choice smoothing parameters ($k_{\Delta y} = 5, k_{\Delta x} = 5, k_x = 6$) is such that the results are not too sensitive to small changes in the smoothing parameters. Specifically, in Table D.1, we verify that the changes in smoothing parameter only affect the mean of the estimator by an amount that is small relative to its standard deviation (std. dev.). Note that we resisted the temptation to merely pick the choice that minimizes the RMSE, as the true RMSE is not empirically accessible in applications. We could have obtained even better results by fine-tuning the smoothing parameters for each model, but did not do so to better emulate real-life settings, where the smoothing parameters may not always be perfectly optimal.

D.2 Small Sample results

Table D.2 illustrates that, even for small samples, the proposed sieve estimation still achieve considerable bias reduction, relative to the naive estimate ignoring the presence of measurement error. As in large samples, the bias reduction is so effective that the RMSE of the sieve estimator is typically much lower than the RMSE of the naive estimator, even though the sieve estimator typically exhibit a larger standard deviation.

Table D.1: Behavior of $\hat{\theta}_{sieve}$ for the same setup as in Table 1, repeated over a range of possible smoothing parameter values.

Case 1

$\theta_1 = 1$			$\theta_2 = 1$			
mean	std. dev.	RMSE	mean	std. dev.	RMSE	Smoothing parameters
1.060	0.189	0.198	0.910	0.131	0.159	$k_{\Delta y} = 5, k_{\Delta x} = 4, k_x = 6$
1.174	0.198	0.263	0.896	0.157	0.188	$k_{\Delta y} = 4, k_{\Delta x} = 4, k_x = 6$
1.058	0.212	0.220	0.924	0.144	0.163	$k_{\Delta y} = 5, k_{\Delta x} = 5, k_x = 6$
0.975	0.204	0.206	0.954	0.137	0.144	$k_{\Delta y} = 5, k_{\Delta x} = 5, k_x = 7$
1.001	0.242	0.242	0.854	0.157	0.213	$k_{\Delta y} = 4, k_{\Delta x} = 5, k_x = 7$
0.904	0.140	0.169	0.925	0.143	0.161	$k_{\Delta y} = 5, k_{\Delta x} = 5, k_x = 5$
1.003	0.293	0.293	0.840	0.127	0.204	$k_{\Delta y} = 6, k_{\Delta x} = 5, k_x = 4$
1.092	0.220	0.238	0.923	0.125	0.146	$k_{\Delta y} = 6, k_{\Delta x} = 5, k_x = 7$
1.047	0.164	0.171	0.917	0.128	0.153	$k_{\Delta y} = 5, k_{\Delta x} = 6, k_x = 7$

Case 2

$\theta_1 = 1$			$\theta_2 = 1$			
mean	std. dev.	RMSE	mean	std. dev.	RMSE	Smoothing parameters
0.938	0.060	0.086	0.917	0.056	0.099	$k_{\Delta y} = 5, k_{\Delta x} = 4, k_x = 6$
0.928	0.069	0.099	0.999	0.064	0.064	$k_{\Delta y} = 4, k_{\Delta x} = 4, k_x = 6$
0.961	0.061	0.073	0.937	0.059	0.086	$k_{\Delta y} = 5, k_{\Delta x} = 5, k_x = 6$
0.920	0.065	0.103	0.960	0.057	0.069	$k_{\Delta y} = 5, k_{\Delta x} = 5, k_x = 7$
0.926	0.052	0.090	1.022	0.057	0.061	$k_{\Delta y} = 4, k_{\Delta x} = 5, k_x = 7$
0.894	0.100	0.145	0.928	0.093	0.117	$k_{\Delta y} = 5, k_{\Delta x} = 5, k_x = 5$
0.892	0.081	0.134	0.969	0.081	0.087	$k_{\Delta y} = 6, k_{\Delta x} = 5, k_x = 4$
0.915	0.067	0.107	0.963	0.061	0.071	$k_{\Delta y} = 6, k_{\Delta x} = 5, k_x = 7$
0.922	0.067	0.102	0.959	0.059	0.071	$k_{\Delta y} = 5, k_{\Delta x} = 6, k_x = 7$

Case 3

$\theta_1 = 1$			$\theta_2 = 1$			
mean	std. dev.	RMSE	mean	std. dev.	RMSE	Smoothing parameters
0.932	0.049	0.083	1.060	0.030	0.068	$k_{\Delta y} = 5, k_{\Delta x} = 4, k_x = 6$
0.924	0.032	0.082	1.062	0.028	0.069	$k_{\Delta y} = 4, k_{\Delta x} = 4, k_x = 6$
0.959	0.079	0.089	1.053	0.037	0.065	$k_{\Delta y} = 5, k_{\Delta x} = 5, k_x = 6$
1.060	0.054	0.081	1.022	0.026	0.034	$k_{\Delta y} = 5, k_{\Delta x} = 5, k_x = 7$
1.041	0.054	0.068	1.027	0.029	0.040	$k_{\Delta y} = 4, k_{\Delta x} = 5, k_x = 7$
0.977	0.080	0.083	1.053	0.040	0.066	$k_{\Delta y} = 5, k_{\Delta x} = 5, k_x = 5$
1.039	0.048	0.062	0.983	0.027	0.032	$k_{\Delta y} = 6, k_{\Delta x} = 5, k_x = 4$
1.040	0.054	0.067	1.026	0.028	0.038	$k_{\Delta y} = 6, k_{\Delta x} = 5, k_x = 7$
1.072	0.056	0.092	1.016	0.029	0.033	$k_{\Delta y} = 5, k_{\Delta x} = 6, k_x = 7$

Case 4

$\theta_1 = 1$			$\theta_2 = 1$			
mean	std. dev.	RMSE	mean	std. dev.	RMSE	Smoothing parameters
1.103	0.164	0.194	1.094	0.154	0.180	$k_{\Delta y} = 5, k_{\Delta x} = 4, k_x = 6$
1.118	0.192	0.226	1.062	0.165	0.177	$k_{\Delta y} = 4, k_{\Delta x} = 4, k_x = 6$
1.079	0.145	0.165	1.088	0.149	0.173	$k_{\Delta y} = 5, k_{\Delta x} = 5, k_x = 6$
1.095	0.137	0.166	1.026	0.161	0.163	$k_{\Delta y} = 5, k_{\Delta x} = 5, k_x = 7$
1.400	0.357	0.536	0.853	0.206	0.253	$k_{\Delta y} = 4, k_{\Delta x} = 5, k_x = 7$
1.196	0.180	0.266	1.067	0.143	0.158	$k_{\Delta y} = 5, k_{\Delta x} = 5, k_x = 5$
1.039	0.178	0.182	1.061	0.218	0.226	$k_{\Delta y} = 6, k_{\Delta x} = 5, k_x = 4$
0.998	0.129	0.129	1.008	0.165	0.165	$k_{\Delta y} = 6, k_{\Delta x} = 5, k_x = 7$
1.066	0.147	0.161	0.975	0.156	0.157	$k_{\Delta y} = 5, k_{\Delta x} = 6, k_x = 7$

Case 5

$\theta_1 = 1$			$\theta_2 = 1$			
mean	std. dev.	RMSE	mean	std. dev.	RMSE	Smoothing parameters
0.824	0.130	0.219	0.942	0.083	0.101	$k_{\Delta y} = 5, k_{\Delta x} = 4, k_x = 6$
0.960	0.239	0.242	0.895	0.210	0.234	$k_{\Delta y} = 4, k_{\Delta x} = 4, k_x = 6$
0.844	0.120	0.196	0.965	0.066	0.074	$k_{\Delta y} = 5, k_{\Delta x} = 5, k_x = 6$
0.866	0.113	0.175	0.968	0.069	0.076	$k_{\Delta y} = 5, k_{\Delta x} = 5, k_x = 7$
1.008	0.208	0.208	0.837	0.227	0.279	$k_{\Delta y} = 4, k_{\Delta x} = 5, k_x = 7$
0.884	0.184	0.217	0.878	0.103	0.159	$k_{\Delta y} = 5, k_{\Delta x} = 5, k_x = 5$
1.110	0.315	0.334	1.019	0.333	0.333	$k_{\Delta y} = 6, k_{\Delta x} = 5, k_x = 4$
0.926	0.144	0.161	0.992	0.081	0.081	$k_{\Delta y} = 6, k_{\Delta x} = 5, k_x = 7$
0.855	0.080	0.165	0.966	0.068	0.076	$k_{\Delta y} = 5, k_{\Delta x} = 6, k_x = 7$

Case 6

$\theta_1 = 1$			$\theta_2 = 1$			
mean	std. dev.	RMSE	mean	std. dev.	RMSE	Smoothing parameters
0.886	0.047	0.122	0.958	0.047	0.062	$k_{\Delta y} = 5, k_{\Delta x} = 4, k_x = 6$
1.052	0.140	0.149	1.011	0.097	0.098	$k_{\Delta y} = 4, k_{\Delta x} = 4, k_x = 6$
0.915	0.059	0.103	0.979	0.054	0.057	$k_{\Delta y} = 5, k_{\Delta x} = 5, k_x = 6$
1.039	0.075	0.084	1.067	0.062	0.091	$k_{\Delta y} = 5, k_{\Delta x} = 5, k_x = 7$
1.130	0.105	0.167	1.047	0.084	0.097	$k_{\Delta y} = 4, k_{\Delta x} = 5, k_x = 7$
0.983	0.076	0.078	1.025	0.067	0.072	$k_{\Delta y} = 5, k_{\Delta x} = 5, k_x = 5$
0.929	0.048	0.085	0.943	0.042	0.070	$k_{\Delta y} = 6, k_{\Delta x} = 5, k_x = 4$

Table D.2: Simulation results as in Table 1, repeated for a sample size of 500.

Case 1: $m(x;\theta) = \theta_1 x + \theta_2 e^x$

Parameter (=true value)	θ_1= 1			θ_2= 1		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
Ignoring meas.error	0.404	0.135	0.610	0.744	0.093	0.271
Accurate data	1.002	0.065	0.065	1.000	0.025	0.025
Sieve MLE	1.040	0.259	0.262	0.964	0.188	0.191

Case 2: $m(x;\theta) = \theta_1 x + \theta_2 x^2$

Parameter (=true value)	θ_1= 1			θ_2= 1		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
Ignoring meas.error	0.748	0.077	0.262	0.541	0.059	0.462
Accurate data	1.000	0.049	0.049	0.998	0.031	0.031
Sieve MLE	0.873	0.325	0.348	0.919	0.162	0.180

Case 3: $m(x;\theta) = \theta_1 x + \theta_2/(1+x^2)$

Parameter (=true value)	θ_1= 1			θ_2= 1		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
Ignoring meas.error	0.634	0.051	0.368	1.037	0.065	0.075
Accurate data	1.000	0.047	0.047	0.998	0.051	0.051
Sieve MLE	1.118	0.110	0.161	0.858	0.072	0.158

Case 4: $m(x;\theta) = (x^2 + \theta_1)(x + \theta_2)$

Parameter (=true value)	θ_1= 1			θ_2= 1		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
Ignoring meas.error	-0.268	0.163	1.279	1.542	0.513	0.746
Accurate data	0.999	0.039	0.039	1.001	0.028	0.028
Sieve MLE	1.072	0.177	0.191	1.059	0.177	0.186

Case 5: $m(x;\theta) = \ln(1 + \theta_1 x + \theta_2 x^2)$

Parameter (=true value)	θ_1= 1			θ_2= 1		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
Ignoring meas.error	0.514	0.083	0.493	0.459	0.069	0.544
Accurate data	0.995	0.115	0.115	1.000	0.111	0.110
Sieve MLE	0.935	0.262	0.270	0.980	0.202	0.203

Case 6: $m(x;\theta) = \theta_1 x + \theta_2 \ln(1 + x^2)$

Parameter (=true value)	θ_1= 1			θ_2= 1		
	mean	std. dev.	RMSE	mean	std. dev.	RMSE
Ignoring meas.error	0.660	0.047	0.342	0.725	0.065	0.282
Accurate data	0.999	0.048	0.048	1.005	0.073	0.073
Sieve MLE	0.922	0.111	0.135	0.984	0.107	0.108

References

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

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

Loève, M. (1977): *Probability Theory I*. New York: Springer.

Lukacs, E. (1970): *Characteristic Functions*. Griffin, London, second edn.

Newey, W. K. (1990): "Semiparametric Efficient Bounds," *Journal of Applied Econometrics*, 5, 99–135.

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

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