Returns to Lying When the Truth is Unobserved

and

Identifying Measurement Error Models Without Instruments

Arthur Lewbel

Boston College

This packet contains three papers that will be discussed in the seminar:

"Returns to Lying When the Truth is Unobserved," by Yingyao Hu, and Arthur Lewbel

"Nonparametric Identification and Estimation of Nonclassical Errors-in-Variables Models Without Additional Information," by Xiaohong Chen, Yingyao Hu, and Arthur Lewbel

and

"Nonparametric Identification of the Classical Errors-in-Variables Model Without Side Information," by Susanne M. Schennach, Yingyao Hu, and Arthur Lewbel.

Identifying the Returns to Lying When the Truth is Unobserved*

Yingyao Hu
Johns Hopkins University

Arthur Lewbel
Boston College

original April, 2007, revised Nov. 2007

Abstract

Consider an observed binary regressor D and an unobserved binary variable D^*, both of which affect some other variable Y. This paper considers nonparametric identification and estimation of the effect of D on Y, conditioning on $D^* = 0$. For example, suppose Y is a person's wage, the unobserved D^* indicates if the person has been to college, and the observed D indicates whether the individual claims to have been to college. This paper then identifies and estimates the difference in average wages between those who falsely claim college experience versus those who tell the truth about not having college. We estimate this average returns to lying to be about 7% to 20%. Nonparametric identification without observing D^* is obtained either by observing a variable V that is roughly analogous to an instrument for ordinary measurement error, or by imposing restrictions on model error moments.

JEL Codes: C14, C13, C20, I2.

Keywords: Binary regressor, misclassification, measurement error, unobserved factor, discrete factor, program evaluation, treatment effects, returns to schooling, wage model.

*We would like to thank Xiaohong Chen for her help on this paper. We also thank participants of UCL, IFS, CEMMAP, BC, Montreal, and Brown seminars for helpful comments, and Douglas Staiger for providing data. All errors are our own.

Department of Economics, Johns Hopkins University, 440 Mergenthaler Hall, 3400 N. Charles Street, Baltimore, MD 21218, USA Tel: 410-516-7610. Email: yhu@jhu.edu, http://www.econ.jhu.edu/people/hu/

Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467 USA. Tel: 617-552-3678. email: lewbel@bc.edu http://www2.bc.edu/~lewbel

1 Introduction

Consider an observed binary regressor D and an unobserved binary variable D^*, both of which affect some other variable Y. This paper considers nonparametric identification and estimation of the effect of D on Y, conditioning on a value of the unobserved D^* (and possibly on a set of other observed covariates X). Formally, what is identified is the function $R(D, X)$ defined by

$$R(D, X) = E(Y \mid D^* = 0, D, X).$$

This can then be used to evaluate

$$r(X) = R(1, X) - R(0, X)$$

and $r = E[r(X)]$, which are respectively, the conditional and unconditional effects of D on Y, holding D^* fixed. When D^* is observed, identification and estimation of R is trivial. Here we obtain identification and provide estimators when D^* is unobserved.

Assuming $E(Y \mid D^*, D, X)$ exists, define a model H and an error η by

$$Y = E(Y \mid D^*, D, X) + \eta = H(D^*, D, X) + \eta \tag{1}$$

where the function H is unknown and the error η is mean zero and uncorrelated with D, D^*, and X. Then, since D and D^* are binary, we may without loss of generality rewrite this model in terms of the unknown R, r, and an unknown function s as

$$Y = R(D, X) + s(D, X)D^* + \eta \tag{2}$$

or equivalently

$$Y = R(0, X) + r(X)D + s(D, X)D^* + \eta. \tag{3}$$

This paper provides conditions that are sufficient to identify the unknown functions R and r, even though D^* is unobserved.

For a specific example, suppose for a sample of individuals the observed D is one if an individual claims or is reported to have some college education (and zero otherwise), and the unobserved D^* is one if the individual actually has some college experience. Let Y be the individual's wage rate. Then r is the

difference in average wages Y between those who claim to have a degree when they actually do not, versus those who honestly report not having a college degree. This paper provides nonparametric identification and associated estimators of the function r. We empirically apply these methods to estimate this average difference in outcomes between truth tellers and liars, when the truth D^* is not observed.

Only responses and not intent can be observed, so we cannot distinguish between intentional lying and false beliefs about D^*. For example, suppose D^* as an actual treatment and D is a perceived treatment (i.e., D is the treatment an individual thinks he received, and so is a false belief rather than an intentional lie). Then r is the average placebo effect, that is, the average difference in outcomes between those who were untreated but believe they received treatment versus those who correctly perceive that they were untreated. This paper then provides identification and an estimator for this placebo effect when the econometrican does not observe who actually received treatment.

Given a Rubin (1974) type unconfoundedness assumption, r will equal the average placebo effect, or the average returns to lying (which could be positive or negative). Unconfoundedness may be a reasonable assumption in the placebo example, but is less likely to hold when lying is intentional. Without unconfoundedness, the difference r in outcomes Y that this paper identifies could be due in part to unobserved differences between truth tellers and liars. For example, r could be positive even if lying itself has no direct effect on wages, if those willing to lie about their education level are on average more aggressive in pursuing their goals than others, or if some of them have spent enough time and effort studying (more on average than other nongraduates) to rationalize claiming that they have college experience. Alternatively r could be negative even if the returns to lying itself is zero, if the liars are more likely to arouse suspicion, or if there exist other negative character flaws that correlate with lying.

The interpretation of r as a placebo effect or returns to lying also assumes that D^* and D are respectively the true and reported values of the same variable. This paper's identification and associated estimator does not require D and D^* to be related in this way (they can be completely different binary variables), and does not require unconfoundedness, however, for the purposes of interpreting the required assumptions and associated results, we will throughout this paper refer to D as the reported value of a true D^* and refer to r as the returns to lying.

Discreteness of D and D^* is also not essential for this paper's identification method, but will simplify the associated estimators. In particular, if we more generally have a reported Z and an unobserved Z^*, we could apply this paper's identification method for any particular values z and z^* of interest by letting $D^* = I(Z^* \neq z)$ and $D = I(Z \neq z)$, where I is the indicator function. Then $D = 1$ when $D^* = 0$ means

lying by claiming a value z when the truth is not z.

When D is a possibly mismeasured or misclassified observation of D^*, then $D - D^*$ is the measurement or misclassification error. Most of the literature on mismeasured binary regressors attempts to estimate the effect of D^* on Y (a treatment effect) and assumes $r(X) = 0$, or equivalently, that the measurement error has no effect on the outcome Y after conditioning on the true D^*. Recent examples include Hu (2006), Mahajan (2006), Lewbel (2007a), and Chen, Hu, and Lewbel (2007). The same is true for general endogenous binary regressor estimators when they are interpreted as arising from mismeasurement. See, e.g., Das (2004), Blundell and Powell (2004), Newey and Powell (2003), and Florens and Malavolti (2003). The assumption that $r(X) = 0$ will be reasonable if the reporting errors $D - D^*$ are due to data collection errors such as accidently checking the wrong box on a survey form. Having $r(X) = 0$ would also hold if the outcome Y could not be affected by the individual's beliefs or reports regarding D, e.g., if D^* were an indicator of whether the individual owns stock and Y is the return on his investment, then that return will only depend on the assets he actually owns and not on his beliefs or self reports about what he owns. Still, there are many applications where it is not reasonable to assume a priori that $r(X)$ is zero, so even when $r(X)$ is not of direct interest, it may be useful to apply this paper's methods to test if it is zero, which would then permit the application of many of the existing treatment or mismeasured or misclassified regressor estimators which all require that $r(X) = 0$.

We propose two different methods to obtain nonparametric identification without observing D^*. One is by observing a variable V that has some special properties, analogous to an instrument. The second way we obtain identification is through restrictions on the first three moments of the model error η. Identification using an instrument V requires V to have some of the properties of a repeated measurement. In particular, Kane and Rouse (1995) and Kane, Rouse, and Staiger (1999) obtain data on both self reports of educational attainment D, and on transcript reports. They provide evidence that this transcript data (like the self reports D) may contain considerable reporting errors on questions like, "Do you have some years of college?" These transcript reports therefore cannot be taken to equal D^*, but we show these transcripts may satisfy the conditions we require for use as an instrument V.

The alternative method we propose for identification does not require an instrument V, but is instead based primarily on assuming that the first three moments of the model error η be independent of the covariates. For example, if η is normal, as might hold by Gibrat's (1931) law for Y being log wages, and homoskedastic, then η will satisfy this assumption.

The next two sections describe identification with and without an instrument. We then propose esti-

mators based on the identification, and provide an empirical application estimating the effects on wages of lying about educational attainment.

2 Identification Using an Instrument

ASSUMPTION A1: *The variable Y, the binary variable D, and a (possibly empty) vector of other covariates X are all observable. The binary variable D^* is unobserved. $E(Y \mid D^*, D, X)$ exists. The functions H, R, r, s and the variable η are defined by equations (1), (2) and (3).*

ASSUMPTION A2: *A variable V is observed with*

$$E(\eta V \mid D, X) = 0, \tag{4}$$

$$E\left(V \mid D, D^* = 1, X\right) = E\left(V \mid D^* = 1, X\right), \tag{5}$$

$$E(V \mid D = 1, X) \neq E(V \mid X). \tag{6}$$

The following Lemmas are useful for interpreting and applying Assumption A2:

LEMMA 1: *Assume $E(D \mid D^* = 1, X) \neq 0$. Equation (5) holds if and only if*

$$Cov\left(D, V \mid D^* = 1, X\right) = 0 \tag{7}$$

LEMMA 2: *Assume $E(D \mid X) \neq 0$. Equation (6) holds if and only if*

$$Cov(D, V \mid X) \neq 0. \tag{8}$$

Proofs of Lemmas and Theorems are in the Appendix. Equation (4) says that the instrument V is uncorrelated with the model error η for any value of the observable regressors D and X. A sufficient condition for equation (4) to hold is if $E(Y \mid D^*, D, X, V) = E(Y \mid D^*, D, X)$. This is a standard property for an instrument.

As shown by Lemmas 1 and 2, equations (5) and (6) say that D and V are correlated, but at least for $D^* = 1$, this relationship only occurs through D^*. Equation (5) means that when $D^* = 1$, the variable D has no additional power to explain V given X. If V is a second mismeasurement of D^*, then (5) or its equivalent (7) is implied by a standard assumption of repeated measurements, namely, that the error in the measurement D be unrelated to the error in the measurement V, while equation (6) can be expected to hold because both measurements are correlated with the true D^*. Equation (6) is close to a standard instrument assumption, if we are thinking of V as an instrument for D (since we are trying to identify the effect of D on Y). Note that equation (6) or Lemma 2 can be easily tested, since they only depend on observables.

To facilitate interpretation of the identifying assumptions, we discuss them in the context of the example in which Y is a wage, D^* is the true indicator of whether an individual has some college experience, D is the individual's self report of college experience, and V is transcript reports of educational attainment, which are an alternative mismeasure of D^*. Let X denote a vector of other observable covariates we may be interested in that can affect either wages, schooling, and/or lying, so X could include observed attributes of the individual and of her job.

In the college and wages example, equation (4) will hold if wages depend on both actual and self reported education, i.e., D^* and D, but not on the transcript reports V. This should hold if employers rely on resumes and worker's actual knowledge and abilities, but don't see college transcripts. Equation (5) or equivalently (7) makes sense, in that errors in college transcripts depend on the actual D^*, but not on what individuals later self report. However, this assumption could be violated if individuals see their own transcripts and base their decision to lie in part on what the transcripts say. Finally, (6) is likely to hold assuming transcripts and self reports are accurate enough on average to both be positively correlated with the truth.

THEOREM 1: *If Assumptions A1 and A2 hold then* $R(D, X)$ *satisfies*

$$R(D, X) = \frac{E\left(YV \mid D, X\right) - E\left(Y \mid D, X\right) E\left(V \mid D^* = 1, X\right)}{E\left(V \mid D, X\right) - E\left(V \mid D^* = 1, X\right)}. \tag{9}$$

It follows immediately from Theorem 1 that $R(D, X)$ is identified if $E\left(V \mid D, X\right) \neq E\left(V \mid D^* = 1, X\right)$ to avoid division by zero and if $E\left(V \mid D^* = 1, X\right)$ can be identified, because the other terms in equation (9) are expectations of observables, conditioned on other observables, and hence are themselves identified.

We now consider two alternative methods of satisfying these conditions needed to identify $R(D, X)$.

ASSUMPTION A3: *Assume*

$$E\left(V \mid D^* = 1, X\right) = 1, \tag{10}$$

and

$$E\left(V \mid D, X\right) \neq 1 \tag{11}$$

Note that if $V \in \{0, 1\}$ (as is the case when V is a mismeasure of D^*, like when V is the transcript report) then equation (10) is equivalent to $\Pr\left(V = 1 \mid D^* = 1, X\right) = 1$. This equation (10) rules out transcript errors of the form $V = 0$ when $D^* = 1$, and therefore requires that only one type of transcript error be possible, namely, $V = 1$ when $D^* = 0$. For example, if D and D^* refer to graduating from college then equation (10) says that anyone who has a diploma will have an accurate transcript, but people who did not graduate may have transcript errors.

Equation (11) requires that there not exist a value of D, X that always yields $V = 1$, or more precisely, that if such a D, X exists, then we cannot identify $R\left(D, X\right)$ for that D, X, since for those people we will not observe any variation in the instrument V. Equation (11) is empirically testable since it depends only on observables.

COROLLARY 1: *If Assumptions A1, A2, and A3 hold then* $R(D, X)$ *is identified by*

$$R(D, X) = \frac{E\left(YV \mid D, X\right) - E\left(Y \mid D, X\right)}{E\left(V \mid D, X\right) - 1} = \frac{E\left(Y\left(V - 1\right) \mid D, X\right)}{E\left(\left(V - 1\right) \mid D, X\right)} \tag{12}$$

Corollary 1 follows from Theorem 1 by substituting equation (10) into equation (9). Equation (11) then ensures that the denominator of equation (12) is nonzero.

Equation (10), which in the wage application rules out one kind of transcript error, may be overly strong. We now consider an alternative assumption and associated identification that does not require this restriction.

ASSUMPTION A3': *There exists an observed binary* $U \in \{0, 1\}$ *(having* U *be an element or subset of* X *is permitted but not required) such that*

$$E\left(V \mid D^* = 1, X\right) = E\left(V \mid U = 1\right) \tag{13}$$

and

$$E\left(V \mid D, X\right) \neq E\left(V \mid U = 1\right) \tag{14}$$

Equation (13) assumes that V has the same mean for people who have $U = 1$ as for people that have $D^* = 1$ and any value of X. One set of sufficient conditions for equation (13) is if $E\left(V \mid D^* = 1, X, U\right) = E\left(V \mid D^* = 1\right)$, so for people having college ($D^* = 1$), the probability of a transcript error is unrelated to one's personal attribute information X and U, and if

$$\Pr\left(D^* = 1 \mid U = 1\right) = 1, \tag{15}$$

so people who have $U = 1$ are an observable subpopulation (e.g., medical doctors or PhD's) that definitely have some college. If equation (15) holds then equation (13) would only be violated if colleges systematically made more errors when producing transcripts for individuals with some value of attributes X, U than for students with other attribute values.

Equation (14) is a technicality that, like equation (11) in Corollary 1, will avoid division by zero in Corollary 2 below. It is difficult to see why it should not hold in general, and it is empirically testable since it depends only on observables. However, if both equations (13) and (15) hold then equation (14) will not hold for values x_0 such that $\Pr\left(U = 1 \mid X = x_0\right) = 1$. This means that $R(D, X)$ cannot be identified for $X = x_0$, which is logical because all members of subgroup x_0 have $U = 1$ which then means they have $D^* = 1$ by equation (15), and therefore none of them can be lying when reporting $D = 1$.

COROLLARY 2: *If Assumptions A1, A2, and A3' hold then* $R(D, X)$ *is identified by*

$$R(D, X) = \frac{E\left(YV \mid D, X\right) - E\left(Y \mid D, X\right) E\left(V \mid U = 1\right)}{E\left(V \mid D, X\right) - E\left(V \mid U = 1\right)}. \tag{16}$$

Corollary 2 follows Theorem 1, by substituting equation (13) into equation (9) to obtain equation (16), and equation (14) makes the denominator in equation (16) be nonzero.

Given identification of $R(D, X)$ by Corollary 1 or 2, the returns to lying $r(X)$ is also identified by $r(X) = R(1, X) - R(0, X)$.

Although rather more difficult to interpret and satisfy than the assumptions in Corollaries 1 and 2, yet another alternative set of identifying assumptions is equations (4), (6) and $Cov\,(D^*, V \mid D, X) = 0$, which by equation (3) implies $Cov\,(Y, V \mid X) = r(X)Cov\,(D, V \mid D, X)$ which can then be solved for, and hence identifies, $r(X)$.

3 Identification Without an Instrument

We now consider identification based on restrictions on moments of η rather than on the presence of an instrument. The method of identification here is similar to that of Chen, Hu, and Lewbel (2007), though that paper imposes the usual measurement error assumption that the outcome Y is conditionally independent of the mismeasure D, conditioning on the true D^*, or equivalently, it assumes that $r(X) = 0$.

ASSUMPTION B2:

$$E\left(\eta \mid D^*, D, X\right) = 0, \tag{17}$$

$$E\left(\eta^k \mid D^*, D, X\right) = E\left(\eta^k\right) \quad \text{for } k = 2, 3, \tag{18}$$

there exists an x_0 *such that*

$$\Pr\left(D = 0 \mid D^* = 1, X = x_0\right) = 0 \quad \text{and} \quad \Pr\left(D = 0 \mid X = x_0\right) > 0, \tag{19}$$

and

$$E\left(Y \mid D^* = 1, D, X\right) \geq E\left(Y \mid D^* = 0, D, X\right) \tag{20}$$

Equation (17) can be assumed to hold without loss of generality by definition of the model error η. Equation (18) says that the second and third moments of the model error η do not depend on D^*, D, X, and so would hold under the common modeling assumption that the error η in a wage equation is independent of the regressors,

Equation (19) implies that people, or at least those in some subpopulation $\{X = x_0\}$, will not underreport and claim to not have been to college if they in fact have been to college. At least in terms of wages, this is plausible in that it is hard to see why someone would lie to an employer by claiming to have less education or training than he or she really possesses.

Finally, equation (20) implies that the impact of D^* on Y conditional on D and X is known to be positive. This makes sense when Y is wages and D^* is the true education level, since ceteris paribus, higher education on average should result in higher wages on average.

Define

$$\begin{aligned}\sigma^2_{Y|D,X}(D,X) &= E\left(Y^2|D,X\right) - \left[E\left(Y|D,X\right)\right]^2, \\ v^3_{Y|D,X}(D,X) &= E\left(\left[Y - E\left(Y|D,X\right)\right]^3|D,X\right),\end{aligned}$$

$$\begin{aligned}\alpha(D,X) &= \sigma^2_{Y|D,X}(D,X) - \sigma^2_{Y|D,X}(0,x_0), \\ \beta(D,X) &= v^3_{Y|D,X}(D,X) - v^3_{Y|D,X}(0,x_0) + 2E\left(Y|D,X\right)\alpha(D,X), \\ \gamma\left(D,X\right) &= \alpha(D,X)^2 + \left[E\left(Y|D,X\right)\right]^2\alpha(D,X) - E\left(Y|D,X\right)\beta(D,X).\end{aligned}$$

THEOREM 2: *Suppose that Assumptions A1 and B2 hold and that* $\alpha(D,X) \neq 0$ *for* $(D,X) \neq (0,x_0)$. *Then,* $R(D,X)$ *and* $s(D,X)$ *are identified as follows:*

i) if $(D,X) = (0,x_0)$*, then* $R(D,X) = E\left(Y|D,X\right)$*;*

ii) if $(D,X) \neq (0,x_0)$*, then*

$$R(D,X) = \frac{\beta(D,X) - \sqrt{\beta(D,X)^2 + 4\alpha(D,X)\gamma\left(D,X\right)}}{2\alpha(D,X)},$$

and

$$s(D,X) = \frac{\alpha(D,X)}{E\left(Y|D,X\right) - R(D,X)} + E\left(Y|D,X\right) - R(D,X).$$

As before given $R(D,X)$ we may identify the returns to lying $r(X)$ using $r(x) = R(1,X) - R(0,X)$. Identification of $s(D,X)$ in Theorem 2 means that the entire conditional mean function H in equation 1 is identified.

The proof of Theorem 2 shows that $R(D,X)$ satisfies a quadratic equation, and equation (20) is only needed to identify which of the two roots is correct.

4 Unconfoundedness

By construction the function $r(X)$ is the difference in the conditional mean of Y (conditioning on D, X, and on $D^* = 0$) when D changes from zero to one. Assuming D is the reported response and D^* is the truth,

here we formally provide the unconfoundedness condition required to have this $r(X)$ equal the returns to lying. Consider the weak version of the Rubin (1974) or Rosenbaum and Rubin (1984) unconfoundedness assumption given by equation (21), interpreting D as a treatment. Letting $Y(d)$ denote what Y equals given the response $D = d$, if

$$E[Y(d) \mid D, D^* = 0, X] = E[Y(d) \mid D^* = 0, X] \tag{21}$$

then it follows immediately from applying, e.g., Heckman, Ichimura, and Todd (1998), that $E[Y(1) - Y(0) \mid D^* = 0, X] = r(X)$ is the conditional average effect of D, and so is the conditional on X average returns to lying.

5 Estimation Using an Instrument

We now provide estimators of $R(D, X)$ and hence of $r(X)$ based on Corollaries 1 and 2 of Theorem 1. We first describe nonparametric estimation based on ordinary sample averages which can be used if X is discrete. We then discuss kernel based nonparametric estimation, and finally we provide a simple least squares based semiparametric estimator that does not require any kernels, bandwidths, or other smoothers regardless of whether X contains continuous or discrete elements.

5.1 Nonparametric, Discrete X Estimation

When X is discrete, replacing the expectations in equation (16) with sample averages gives the estimators

$$\widehat{R}(d, x) = \frac{\widehat{\mu}_{Y,V,X,d} - \widehat{\mu}_{Y,X,d}\widehat{\mu}}{\widehat{\mu}_{V,X,d} - \widehat{\mu}_{X,d}\widehat{\mu}}, \quad \widehat{r}(x) = \widehat{R}(1, x) - \widehat{R}(0, x). \tag{22}$$

with

$$\begin{aligned}
\widehat{\mu}_{Y,V,X,d} &= \frac{1}{n}\sum_{i=1}^{n} Y_i V_i I(X_i = x, D_i = d), \quad \widehat{\mu}_{Y,X,d} = \frac{1}{n}\sum_{i=1}^{n} Y_i I(X_i = x, D_i = d), \\
\widehat{\mu}_{V,X,d} &= \frac{1}{n}\sum_{i=1}^{n} V_i I(X_i = x, D_i = d), \quad \widehat{\mu}_{X,d} = \frac{1}{n}\sum_{i=1}^{n} I(X_i = x, D_i = d), \\
\widehat{\mu}_{V,U} &= \frac{1}{n}\sum_{i=1}^{n} V_i U_i, \quad \widehat{\mu}_U = \frac{1}{n}\sum_{i=1}^{n} U_i, \quad \widehat{\mu} = \widehat{\mu}_{V,U}/\widehat{\mu}_U
\end{aligned}$$

Estimation based on equation (12) is the same replacing $\widehat{\mu}$ with the number one in equation (22)

We also consider the unconditional mean returns $R_d = E\left[R\left(d, X\right)\right]$ and unconditional average returns to lying $r = E\left[r\left(X\right)\right]$, which may be estimated by

$$\widehat{R}_d = \frac{1}{n}\sum_{i=1}^{n}\widehat{R}(d, X_i), \quad \widehat{r} = \frac{1}{n}\sum_{i=1}^{n}\widehat{r}(X_i). \tag{23}$$

Assuming independent, identically distributed draws of $\{Y_i, V_i, X_i, D_i, U_i\}$, and existence of relevant variances, it follows immediately from the Lindeberg-Levy central limit theorem and the delta method that $\widehat{R}(d, x)$, $\widehat{r}(x)$, $\widehat{R}_d$, and $\widehat{r}$ are root n consistent and asymptotically normal, with variance formulas as provided in the appendix, or that can be obtained by an ordinary bootstrap. Analogous limiting distribution results will hold with heteroskedastic or nonindependent data generating processes, as long as a central limit theorem still applies.

5.2 General Nonparametric Estimation

Letting $\mu = E\left(V \mid U = 1\right)$, equation (16) can be rewritten as

$$R(D, X) = \frac{E\left[Y\left(V - \mu\right) \mid D, X\right]}{E\left[\left(V - \mu\right) \mid D, X\right]}. \tag{24}$$

Equation (12) can also be written in the form of equation (24) by letting $\mu = 1$.

Assume n independent, identically distributed draws of $\{Y_i, V_i, X_i, D_i, U_i\}$. Let $\widehat{\mu} = \widehat{\mu}_{V,U}/\widehat{\mu}_U$ if estimation is based on equation (16), otherwise let $\widehat{\mu} = 1$ if estimation is based on equation (12). Let $X_i = (Z_i, C_i)$ where Z and C are, respectively, the vectors of discretely and continuously distributed elements of X. Similarly let $x = (z, c)$. Based on equation (24), a kernel based estimator for $R(D, X)$ is

$$\widehat{R}(d, x) = \frac{\Sigma_{i=1}^{n} Y_i\left(V_i - \widehat{\mu}\right) K[(C_i = c)/b] I(Z_i = z) I(D_i = d)}{\Sigma_{i=1}^{n}\left(V_i - \widehat{\mu}\right) K[(C_i = c)/b] I(Z_i = z) I(D_i = d)} \tag{25}$$

where K is a kernel function and b is a bandwidth that goes to zero as n goes to infinity. Equation (25) is numerically identical to the ratio of two ordinary nonparametric Nadaraya-Watson kernel regressions of $Y\left(V - \widehat{\mu}\right)$ and $V - \widehat{\mu}$ on X, D, which under standard conditions are consistent and asymptotically normal. These will have the same slower than root n rate of convergence as regressions that used the constant μ in place of the estimator $\widehat{\mu}$, because $\widehat{\mu}$ either equals the constant one, or it converges at the rate root n by the law of large numbers. Alternatively, equation (24) can be rewritten as the conditional moment condition

$$E\left[\left(Y - R(D, X)\right)\left(V - \mu\right) \mid D, X\right] = 0 \tag{26}$$

which may be estimated using, e.g., the functional GMM estimator of Ai and Chen (2003), or by Lewbel's (2007b) local GMM estimator , with limiting distributions as provided by those references.

Given $\widehat{R}(d,x)$ from equation (25) we may as before construct $\widehat{r}(x) = \widehat{R}(1,x) - \widehat{R}(0,x)$, and unconditional returns $\widehat{R}_d$ and $\widehat{r}$ by equation (23). We also construct trimmed unconditional returns $\widehat{r}_t = \frac{1}{n}\sum_{i=1}^{n}\widehat{r}(X_i)I_{ti}$ and similarly for $\widehat{R}_{dt}$, where I_{ti} is a trimming parameter that equals one for most observations i, but equals zero for tail observations. Assuming regularity conditions such as Newey (1994) these trimmed unconditional returns are root n consistent and asymptotically normal of trimmed means r_t and R_{dt}.

5.3 Simple Semiparametric Estimation

Assume we have a parameterization $R(D, X, \theta)$ for the function $R(D, X)$ with a vector of parameters θ. The function $s(D, X)$ and the distribution of the model error η are not parameterized. Then based on the definition of μ and equation (26), θ and μ could be jointly estimated based on Corollary 2 by applying GMM to the moments

$$E\left[(V-\mu)\,U\right] = 0 \tag{27}$$

$$E\left[\psi\left(D,X\right)\left(Y - R(D,X,\theta)\right)\left(V-\mu\right)\right] = 0 \tag{28}$$

for a chosen vector of functions $\psi\left(D,X\right)$. For estimation based on Corollary 1, the estimator would just use the moments given by equation (28) with $\mu = 1$.

Let $W = \left(1, D, X'\right)'$. If R has the linear specification $R(D,X,\theta) = W'\theta$ then let $\psi\left(D,X\right) = W$ to yield moments $E\left[W\left(Y - W'\theta\right)\left(V-\mu\right)\right] = 0$, so $\theta = E\left[\left(V-\mu\right)WW'\right]^{-1}E\left[\left(V-\mu\right)WY\right]$. This then yields a weighted linear least squares regression based estimator

$$\widehat{\theta} = \left[\sum_{i=1}^{n}\left(V_i - \widehat{\mu}\right)W_i W_i'\right]^{-1}\left[\sum_{i=1}^{n}\left(V_i - \widehat{\mu}\right)W_i Y_i\right] \tag{29}$$

based on Corollary 2, or the same expression with $\widehat{\mu} = 1$ based on Corollary 1. Given $\widehat{\theta}$ we then have $\widehat{R}(D,X) = W'\widehat{\theta}$. In this semiparametric specification $r(x)$ is a constant with $\widehat{r}(x) = \widehat{r} = \widehat{\theta}_1$, the first element of $\widehat{\theta}$. Note that both GMM based on equation (28) and the special case of weighted linear regression based on equation (29) do not require any kernels, bandwidths, or other smoothers for their implementation.

6 Estimation Without an Instrument

We now consider estimation based on Theorem 2. As in the previous section, let K be a kernel function, b be a bandwidth, and $X_i = (Z_i, C_i)$ where Z and C are, respectively, the vectors of discretely and continuously distributed elements of X. Also let $x = (z, c)$. For $k = 1, 2, 3$, define

$$\widehat{E}\left(Y^k | D = d, X = x\right) = \frac{\Sigma_{i=1}^n Y_i^k K[(C_i = c)/b] I(Z_i = z) I(D_i = d)}{\Sigma_{i=1}^n K[(C_i = c)/b] I(Z_i = z) I(D_i = d)} \tag{30}$$

This is a standard Nadayara-Watson Kernel regression combining discrete and continuous data, which provides a uniformly consistent estimator of $E\left(Y^k | D = d, X = x\right)$ under standard conditions. Define

$$\begin{aligned}
\widehat{\sigma}^2_{Y|D,X}(d, x) &= \widehat{E}\left(Y^2 | D = d, X = x\right) - \left[\widehat{E}\left(Y | D = d, X = x\right)\right]^2, \\
\widehat{v}^3_{Y|D,X}(d, x) &= \widehat{E}\left(\left[Y - \widehat{E}\left(Y | D = d, X = x\right)\right]^3 | D = d, X = x\right),
\end{aligned}$$

$$\begin{aligned}
\widehat{\alpha}(d, x) &= \widehat{\sigma}^2_{Y|D,X}(d, x) - \widehat{\sigma}^2_{Y|D,X}(0, x_0), \\
\widehat{\beta}(d, x) &= \widehat{v}^3_{Y|D,X}(d, x) - v^3_{Y|D,X}(0, x_0) + 2\widehat{E}\left(Y | D = d, X = x\right)\widehat{\alpha}(d, x), \\
\widehat{\gamma}(d, x) &= \widehat{\alpha}(d, x)^2 + \left[\widehat{E}\left(Y | D = d, X = x\right)\right]^2 \widehat{\alpha}(d, x) - \widehat{E}\left(Y | D = d, X = x\right)\widehat{\beta}(d, x).
\end{aligned}$$

Based on Theorem 2 and uniform consistency of the kernel regressions, a consistent estimator of $R(d, x)$ is then

$$\widehat{R}(0, x_0) = \widehat{E}\left(Y | D = 0, X = x_0\right),$$

$$\widehat{R}(d, x) = \frac{\widehat{\beta}(d, x) - \sqrt{\widehat{\beta}(d, x)^2 + 4\widehat{\alpha}(d, x)\widehat{\gamma}(d, x)}}{2\widehat{\alpha}(d, x)} \text{ for } (d, x) \neq (0, x_0).$$

If X does not contain any continuously distributed elements, then these estimators are smooth functions of cell means, and so are root n consistent and asymptotically normal by the Lindeberg Levy central limit theorem and the delta method. Given $\widehat{R}(d, x)$ from equation (25) we may as before construct $\widehat{r}(x) = \widehat{R}(1, x) - \widehat{R}(0, x)$, and unconditional returns $\widehat{R}_d$ and $\widehat{r}$ by equation (23). Also as before, Root n consistent, asymptotically normal convergence of trimmed means of $\widehat{R}_d$ and $\widehat{r}$ is possible using regularity conditions as in Newey (1994) for two step plug in estimators.

7 Returns to Lying about College

Here we report results of empirically implementing our estimators of $r(x)$ where D is self reports of schooling and Y is log wages. We will for convenience refer to these results as returns to lying, but strong caveats are required for that interpretation. First, we are only estimating conditional means, so our results fail to control for the selection effects that lie at the heart of the modern literature on wages and schooling going back at least to Heckman (1979). In addition, unconfoundedness with respect to lying based on equation (21) may not hold for the reasons listed in the introduction. Finally, our sample is likely to not be representative of the general population. It is therefore safest to interpret the estimates here as simply difference in means between truth tellers and liars for a limited sample, rather than as formal returns to lying.

7.1 Preliminary Data Analysis

Kane, Rouse, and Staiger (1999) estimate a model of wages as a function of having either some college, an associate degree or higher, or a bachelors degree or higher. Their model also includes other covariates, and they use data on both self reports and transcript reports of education level. Their data is from the National Longitudinal Study of High School Class of 1972 (NLS-72) and a Post-secondary Education Transcript Survey (PETS). We use their data set of $n = 5912$ observations to estimate the returns to lying, defining Y to be log wage in 1986, D to be one if an individual self reports having "some college" and zero otherwise, while V is one for a transcript report of having "some college" and zero otherwise (both before 1979). We also provide estimates where D and V are self and transcript reports of having an associate degree or more, and reports of having a bachelor's degree or more. We take X to be the same set of other regressors Kane, Rouse, and Staiger (1999) used, which are a 1972 standardized test score and zero-one dummy variables for female, black nonhispanic, hispanic, and other nonhispanic.

The means of D and V (which equal the fractions of our sample that report having that level of college or higher) are 0.6739 and 0.6539 for "some college," 0.4322 and 0.3884 respectively for "Associate degree," and 0.3557 and 0.3383 for "Bachelors degree." The mean of U is 0.03468, so about 3.5% of our sample have transcripts indicating graduate degrees, and the average log wage Y is 2.228.

If D^* were observed along with Y and D, then the functions $r(x)$ and $s(d,x)$ could be immediately estimated from equation equation (3). Table 1 provides preliminary estimates of r and s based on this equation, under the assumption that transcripts have no errors. The row "r if V=D*" in Table 1 is the sample estimates of $E(Y|V=0, D=1) - E(Y|V=0, D=0)$, which would equal an estimate of $r = E[r(X)]$

if $V = D^*$, that is, if the transcripts V were always correct. The row, "r if V=D*, linear" is the coefficient of D in a linear regression of Y on D, V, $D * V$, and X, and so is another estimate of r that would be valid if if $V = D^*$ and given a linear model for log wages.

The third row of Table 1 is the sample analog of $E(Y|V = 1) - E(Y|V = 0)$, which if $V = D^*$ would be an estimate of the returns to schooling $s = E[s(D, X)]$ (or more precisely, the difference in conditional means of log wages between those with $D^* = 1$, versus those with $D^* = 0$, which is returns to schooling if the effects of schooling satisfy an unconfoundedness condition). In this and all other tables, standard errors are obtained by 400 bootstrap replications, and are given in parentheses.

Table 1 also shows the fraction of truth tellers and liars, if the transcripts V were always correct. The rows labeled E(DV) and E[(1-D)(1-V)] gives the fraction of observations where self and transcript reports agree that the individual respectively either has or does not have the given level of college. The row labeled E[D(1-V)] gives the fraction of relevant liars if the transcripts are correct, that is, it is the fraction who claim to have the given level of college, $D = 1$, while their transcripts say they do not, $V = 0$. This fraction is a little over 5% of the sample for some college or Associate degree, but only about half that amount appear to lie about having Bachelor's degree.

If V has no errors, then Table 1 indicates a small amount of lying in the opposite direction, given by the row labeled E[(1-D)V]. These are people who self report having less education than is indicated by their transcripts, ranging from a little over half a percent of the sample regarding college degrees to almost 3% for "some college." It is difficult to see a motive for lying in this direction, which suggests ordinary reporting errors in self reports, transcript reports, or both.

Table 1: Returns to Lying and Schooling Treating Transcripts as True

	Some college	Associate degree	Bachelor's degree
r if V=D*	0.1266 (0.03129)	0.2322 (0.02748)	0.1948 (0.04451)
r if V=D*, linear	0.07868 (0.02864)	0.1681 (0.02777)	0.1269 (0.04082)
s if V=D*	0.2831 (0.01366)	0.2958 (0.01288)	0.3181 (0.01280)
E(DV)	0.6204	0.3794	0.3325
E[D(1-V)]	0.05345	0.05277	0.02317
E[(1-D)V]	0.03349	0.008965	0.005751
E[(1-D)(1-V)]	0.2926	0.5589	0.6385

Standard Errors are in Parentheses

Prior to estimating $r(x)$, we examined equation (6) of Assumption A2, which is testable. A sufficient condition for equation (6) to hold is that $E(V|D=1) - E(V) \neq 0$. In our data the t-statistic for the null hypothesis $E(V|D=1) = E(V)$ is over 40 for each of the three levels of schooling considered, which strongly supports this assumption.

7.2 Estimates Based on Corollary 1

Table 2 summarizes estimates of $r(x)$ based on Corollary 1. Nonparametric estimates of $\widehat{r}(x) = \widehat{R}(1,x) - \widehat{R}(0,x)$ are obtained with $\widehat{R}(d,x)$ given by equation (25) with $\widehat{\mu}_{V|U} = 1$, where the variable C in X is the test score, and Z is the vector of other elements of X. The first row of Table 2 contains r, the sample average of $\widehat{r}(X)$, while the second row has the estimated trimmed mean r_t, which is the sample average of $\widehat{r}(X)$ after removing the highest 5% and lowest 5% of $\widehat{r}(X)$ in the sample. Next are the lower quartile, middle quartile (median) and upper quartile r_{q1}, r_{med}, and r_{q3}, of $\widehat{r}(X)$ in the sample. The final row, "r semi, linear" is a semiparametric estimate of r using equation (29). As before, standard errors are based on 400 bootstrap replications. One set of sufficient regularity conditions for bootstrapping here is Theorem B in Chen, Linton, and Van Keilegom (2003).

For the nonparametric estimates, the kernel function K is a standard normal density function, with bandwidth $b = 0.1836$ given by Silverman's rule. Doubling or halving this bandwidth changed most estimates by less than 10%, indicating that the results were generally not sensitive to bandwidth choice. An exception is that mean and trimmed mean estimates for the Bachelor's degree, which are small in Table 2, become larger (closer to the median r estimate) when the bandwidth is doubled. The results for bachelor's degree are also much less precisely estimated than for some college or associate degree, with generally twice as large standard errors. Based on Table 1, we might expect that far fewer individuals lie about having a bachelor's degree, so the resulting imprecision in the Bachelor's degree estimates could be due to a much smaller fraction of data points that are informative about lying.

The nonparametric mean and median estimates of r are generally significant in Table 2, except for the mean estimates for the Bachelor's degree. Overall, these results indicate that those who lie by claiming to have have some college have about 7% higher wages than those who tell the truth about not having any college on average, and those who lie by claiming to have an associate or bachelor's degree have about 18% higher wages. However, the variability in these returns is large, ranging from zero or negative returns at the first quartile to returns of 14% for some college to 31% for a degree at the third quartile. The semiparametric estimates of r are similar to the mean of the nonparametric estimates, though the variation in the quantiles

of the nonparametric estimates suggests that the semiparametric specification, which assumes r is constant, is not likely to hold.

Recall that estimation based on Corollary 1 assumes no observations with $D^* = 1$ and $D = 0$. If transcripts V are very accurate, then V should be close to D^*, so $E[(1-D)V]$ in Table 1 should be close to zero, and the estimates of r in Table 1 should be close to those in Table 2. The evidence on this is mixed. $E[(1-D)V]$ is close to zero for the two types of degrees, but less so for "some college." The linear model estimates in Table 1 are close to the semiparametric linear model estimates in Table 1, however, the nonparametric estimates of r in Table 1 are rather larger than the mean and median nonparametric estimates in Table 2. In linear models measurement error generally causes attenuation bias, but in contrast here the potentially mismeasured data estimates appear too large rather than too small. This could be due to nonlinearity, or because the potentially mismeasured variable V is highly correlated with another regressor, D.

We should expect that the returns to lying would be smaller than the returns of actually having some college or a degree. These returns to actual schooling are not identified from the assumptions in Corollary 1 or 2. Table 1 gives estimates of returns to schooling s of 28% for some college to 32% for a bachelor's degree, though these estimates are only reliable if transcripts V are accurate. These are indeed higher than the returns to lying, as one would expect. Also, while we would expect the returns to schooling to increase monotonically with the level of schooling, we do not necessarily expect the returns to lying to increase in the same way, because those returns depend on other factors like the plausibility of the lie.

Table 2: Returns to Lying, Nonparametric and Semiparametric Corollary 1 IV Estimates

	Some college	Associate degree	Bachelor's degree
r nonparametric	0.07051 (0.03420)	0.1759 (0.02998)	0.04140 (0.07665)
r_t nonparametric	0.07355 (0.03166)	0.1896 (0.02894)	0.09440 (0.07203)
r_{q1} nonparametric	-0.05768 (0.04930)	0.09691 (0.04219)	-0.04684 (0.09336)
r_{med} nonparametric	0.06447 (0.03663)	0.1992 (0.03995)	0.1748 (0.05683)
r_{q3} nonparametric	0.1421 (0.03903)	0.3111 (0.03920)	0.2478 (0.06094)
r semi, linear	0.08008 (0.02940)	0.1702 (0.02668)	0.1281 (0.04353)

Kane, Rouse, and Staiger (1999) report some substantial error rates in transcripts, however, those findings are based on model estimates that could be faulty, rather than any type of direct verification. Based on

our empirical results comparing Tables 1 and 2, it is possible that transcripts are generally accurate, and in that case the ability of our estimator to produce reasonable estimates of r would not be impressive, since one could then just as easily generate good estimates of r using regressions or cell means as in Table 1. Therefore, to check the robustness of our methodology, we reestimated the model after randomly changing 20% of the observations of V to $1 - V$, thereby artificially making V a much weaker instrument. The resulting estimates of the mean and trimmed mean of r were generally higher than those reported in Tables 1 and 2 (consistent with our earlier result that, in our application, measurement error in V seems to raise rather than lower estimates of the returns to lying), but the estimates of the median of r with this noisy V data are very close to the median estimates in table 2 (though of course with much larger standard errors). Specifically, the r_{med} estimates with substantial measurement error added to V were 0.070, 0.190, and 0.170, compared to the r_{med} estimates in Table 2 of 0.064, 0.199, and 0.175.

To summarize how $\widehat{r}(x)$ varies with regressors x, Table 3 reports the estimated coefficients from linearly regressing the nonparametric estimates $\widehat{r}(x)$ on x and on a constant. The results show a few interesting patterns, including that women appear to have a higher return to lying than men, and that for individuals with above average high school test scores also have above average returns to lying about a higher degree of education. These results are consistent with the notion that returns to lying should be highest for those can lie most plausibly (e.g., those with high ability) or for those who may be perceived as less likely to lie (such as women). However, these results should not be over interpreted, since they are not particularly stable and many are not statistically insignificant.

Table 3: Nonparametric Corollary 1 IV Returns to Lying Linearized Coefficient Estimates

X	Some college	Associate degree	Bachelor's degree
blacknh	-0.09208 (0.1246)	-0.1521 (0.1114)	0.04640 (0.2464)
hispanic	0.01220 (0.1289)	-0.1492 (0.1968)	0.05146 (0.5439)
othernh	0.2176 (0.1304)	0.03830 (0.1844)	-0.005045 (0.4755)
female	0.09291 (0.06570)	0.1791 (0.05585)	0.01029 (0.1449)
mscore	-0.009755 (0.03807)	0.05248 (0.03832)	0.1608 (0.09928)
constant	0.02449 (0.04635)	0.09574 (0.04338)	-0.001377 (0.1018)

8 Alternative Estimates

To check the robustness of our results to alternative identifying assumptions, we now provide estimates based on Corollary 2 and Theorem 2. First consider estimation based on Corollary 2, which replaces Assumption A3 with Assumption A3', and so requires an additional variable U. We define U to equal one for individual's that both self report having a masters degree or a PhD and are in the top decile of the standardized test scores. We could have based U on transcript reports of a graduate degree instead, but then by construction we would have $\widehat{\mu}_{V|U} = 1$, which would then yield numerically identical estimates to those previously reported based on Corollary 1. In our data $\widehat{\mu}_{V|U}$ is .971 for a Bachelor's degree, .981 for an Associate degree, and 1.000 for some college (so the estimates for "some college" in Table 4 are the same as in Table 2). The estimates of returns to lying are somewhat lower in Table 4 than in Table 2. In a few cases they are much lower (e.g., the median returns to lying about a bachelor's degree are only 7% in Table 4 versus 17% in Table 2) but the standard errors are also larger in Table 4, so the differences between the tables are not statistically significantly.

Table 4: Returns to Lying, Nonparametric and Semiparametric Corollary 2 IV Estimates

	Some college	Associate degree	Bachelor's degree
r nonparametric	0.07052 (0.03420)	0.1696 (0.3335)	0.1250 (1.918)
r_t nonparametric	0.07355 (0.03166)	0.1796 (0.04158)	0.07109 (0.1217)
r_{q1} nonparametric	-0.05768 (0.04930)	0.09099 (0.06185)	-0.1654 (0.1841)
r_{med} nonparametric	0.06447 (0.03663)	0.1287 (0.04903)	0.06696 (0.1003)
r_{q3} nonparametric	0.1421 (0.03903)	0.3214 (0.05156)	0.3002 (0.1596)
r semi, linear	0.08008 (0.02940)	0.1610 (0.03362)	0.05613 (1.138)

In Table 5 we report the returns to lying using the estimator based on Theorem 2, which does not use data on either the instrument V or U. These estimates are based only on self reports, and so do not use the transcript data in any way. For these estimates we let $X_0 = X$, which (as with the estimates based on Corollary 1) implies the assumption that that no one reports $D = 0$ when $D^* = 1$ (and hence that transcripts are wrong for the observations in the data that have $D = 0$ and $V = 1$).

As should be expected, the estimates in Table 5 are less precise than those in Table 2, in part because they do not exploit any transcript information, and they assume no heteroskedasticity in the model error η,

which may not hold in this application. They are also more variable in part because they depend on higher moments of the data, and so will be more sensitive to outliers in the first stage nonparametric estimates. Still, the estimates in Table 5 are generally consistent with those in Table 2, in particular, as with Table 4, almost all of the differences between Tables 2 and 5 are not statistically significantly

Table 5: Returns to Lying, Nonparametric and Semiparametric Theorem 2 Estimates Without IV

	Some college	Associate degree	Bachelor's degree
r nonparametric	-0.4127 (28.66)	0.1917 (2.915)	0.1247 (18.27)
r_t nonparametric	0.05064 (0.1402)	0.1684 (0.1738)	0.09186 (0.2489)
r_{q1} nonparametric	-0.05096 (0.1446)	-0.1065 (0.2406)	-0.5425 (0.3659)
r_{med} nonparametric	0.1179 (0.06115)	0.1495 (0.06191)	0.1958 (0.05549)
r_{q3} nonparametric	0.2570 (0.1019)	0.2813 (0.1428)	0.3308 (0.2038)

Given the substantial differences in estimators and identifying assumptions between Corollary 1 and Theorem 2, it is reassuring that the resulting estimates are robust across the two methodologies.

In the Appendix we report the estimates of $E\left[R(d,X)\right]$ corresponding to Tables 2, 4, and 5. As one would expect, these are generally more stable than the estimates of $E\left[r(X)\right]$ reported in Tables 2, 4, and 5, since $r(X)$ is a difference $R(1,X) - R(0,X)$ rather than a level $R(d,X)$.

9 Conclusions

We provide identification and associated estimators for the conditional mean of an outcome Y, conditioned upon an observed discrete variable D and an unobserved discrete variable D^*.

In our empirical application, Y is log wages, while D and D^* are self reports and actual levels of educational attainment. We find that wages are on average about 7% higher for those who lie about having some college, and from 7% to 20% higher on average for those who lie about having a college degree, relative to those who tell the truth about not having college or those degrees. Estimates at the median appear to be more reliable and robust than estimates of the mean returns. Our median results are about the same based on either semiparametric or nonparametric estimation, and are roughly comparable whether identification and associated estimation is based on using transcript reports as an instrument, or is based on higher moment error independence assumptions without exploiting transcript data.

In this application D and D^* refer to the same binary event (educational attainment), with D a self report of D^*, so what is identified is the mean effects of having D either agree with or contradict D^*, which given unconfoundedness identifies either the returns to lying (if misreports of D are intentional) or a placebo effect.

To apply our methodology, it is not necessary for D and D^* to refer to the same binary event. More generally, one could estimate the average effect of any binary treatment or choice D (e.g., exposure to a law, a tax, or an advertisement) on any outcome Y (e.g., compliance with a law, income, expenditures on a product) where the effect is averaged only over some subpopulation of interest indexed by D^* (e.g., potential criminals, the poor, or a target audience of potential buyers), and where we do not observe exactly who is in the subpopulation of interest. Our identification strategy may thereby be relevant to a wide variety of applications, not just returns to lying.

10 Appendix

Proof of Lemmas 1 and 2: Consider Lemma 2 first:

$$\begin{aligned} Cov\left(D, V \mid X\right) &= E\left(DV \mid X\right) - E\left(D \mid X\right) E\left(V \mid X\right) \\ &= E\left[DE\left(V \mid D, X\right) \mid X\right] - E\left(D \mid X\right) E\left(V \mid X\right) \\ &= \Pr\left(D = 1 \mid X\right) E\left(V \mid D = 1, X\right) - E\left(D \mid X\right) E\left(V \mid X\right) \\ &= E\left(D \mid X\right)\left[E\left(V \mid D = 1, X\right) - E\left(V \mid X\right)\right] \end{aligned}$$

so $Cov\left(D, V \mid X\right) \neq 0$ if and only if the right side of the above expression is nonzero. The proof of Lemma 1 works exactly the same way.

Proof of Theorem 1:

First observe that

$$\begin{aligned} E\left(D^*V \mid D, X\right) &= \sum_{d^*=0}^{1} \Pr\left(D^* = d^* \mid D, X\right) E\left(D^*V \mid D^* = d^*, D, X\right) \\ &= \Pr\left(D^* = 1 \mid D, X\right) E\left(V \mid D^* = 1, D, X\right) \\ &= E\left(D^* \mid D, X\right) E\left(V \mid D^* = 1, X\right) \end{aligned}$$

and using this result we have

$$\begin{aligned} E\left(YV \mid D, X\right) &= R(D, X)E\left(V \mid D, X\right) + s(D, X)E\left[D^*V \mid D, X\right] + E\left(\eta V \mid D, X\right) \\ &= R(D, X)E\left(V \mid D, X\right) + s(D, X)E\left(D^* \mid D, X\right)E\left(V \mid D^* = 1, X\right). \end{aligned}$$

Also

$$E\left(Y \mid D, X\right) = R(D, X) + s(D, X)E\left[D^*|D, X\right]$$

Use the latter equation to substitute $s(D, X)E\left[D^*|D, X\right]$ out of the former equation, and solve what remains for $R(D, X)$ to obtain equation (9).

Proof of Theorem 2: Begin with equation (2), $Y = R(D, X) + s(D, X)D^* + \eta$ with $R(D, X) = R(X) + r(X)D$. Assumption B2.1-2 implies that

$$\begin{aligned} \mu_{Y|D,X} &\equiv E\left(Y|D, X\right) \\ &= E\left(\left(R(D, X) + s(D, X)D^*\right)|D, X\right) \\ &= R(D, X) + s(D, X)E\left(D^*|D, X\right), \end{aligned} \tag{31}$$

$$\begin{aligned} \mu_{Y^2|D,X} &\equiv E\left(Y^2|D, X\right) \\ &= E\left(\left(R(D, X) + s(D, X)D^* + \eta\right)^2|D, X\right) \\ &= E\left(\left(R(D, X) + s(D, X)D^*\right)^2|D, X\right) + E\eta^2 \\ &= R(D, X)^2 + 2R(D, X)s(D, X)E\left(D^*|D, X\right) + s(D, X)^2E\left(D^*|D, X\right) + E\eta^2 \\ &= R^2 + 2R\left(\mu_{Y|D,X} - R\right) + s\left(\mu_{Y|D,X} - R\right) + E\eta^2 \\ &= \mu_{Y|D,X}R + (R + s)\left(\mu_{Y|D,X} - R\right) + E\eta^2, \end{aligned} \tag{32}$$

and

$$\begin{aligned} \mu_{Y^3|D,X} &\equiv E\left(Y^3|D, X\right) \\ &= E\left(\left(R(D, X) + s(D, X)D^* + \eta\right)^3|D, X\right) \\ &= E\left[\left(R(D, X) + s(D, X)D^*\right)^3|D, X\right] + 3E\left[\left(R(D, X) + s(D, X)D^*\right)|D, X\right]E\eta^2 + E\eta^3 \\ &= R(D, X)^3 + 3R(D, X)^2s(D, X)E\left(D^*|D, X\right) \\ &\quad +3R(D, X)s(D, X)^2E\left(D^*|D, X\right) + s(D, X)^3E\left(D^*|D, X\right) \\ &\quad +3\mu_{Y|D,X}E\eta^2 + E\eta^3. \end{aligned} \tag{33}$$

We then show that assumption B2.3 implies the identification of $E\left(\eta^k\right)$ for $k = 2, 3$. This assumption implies that

$$\begin{aligned} & E\left(D^*|D=0, X=x_0\right) \\ = & \Pr\left(D^*=1|D=0, X=x_0\right) \\ = & \Pr\left(D=0|D^*=1, X=x_0\right) \frac{\Pr\left(D^*=1|X=x_0\right)}{\Pr\left(D=0|X=x_0\right)} \\ = & 0, \end{aligned}$$

and therefore,

$$\begin{aligned} \mu_{Y|0,x_0} &\equiv E\left(Y|D=0, X=x_0\right) \\ &= R(0,x_0) + s(0,x_0) E\left(D^*|D=0, X=x_0\right) \\ &= R(0,x_0), \end{aligned}$$

$$\begin{aligned} \mu_{Y^2|0,x_0} &\equiv E\left(Y^2|D=0, X=x_0\right) \\ &= R(0,x_0)^2 + 2R(D,X)s(D,X)E\left(D^*|D=0, X=x_0\right) \\ &\quad + s(D,X)^2 E\left(D^*|D=0, X=x_0\right) + E\eta^2 \\ &= R(0,x_0)^2 + E\eta^2 \\ &= \mu_{Y|0,x_0}^2 + E\eta^2, \end{aligned}$$

and

$$\begin{aligned} \mu_{Y^3|0,x_0} &= E\left(Y^3|D=0, X=x_0\right) \\ &= R(0,x_0)^3 + 3\mu_{Y|0,x_0} E\eta^2 + E\eta^3 \\ &= \mu_{Y|0,x_0}^3 + 3\mu_{Y|0,x_0}\left(\mu_{Y^2|0,x_0} - \mu_{Y|0,x_0}^2\right) + E\eta^3. \end{aligned}$$

Therefore, we have

$$\begin{aligned} E\eta^2 &= \mu_{Y^2|0,x_0} - \mu_{Y|0,x_0}^2 \\ &\equiv \sigma_{Y|0,x_0}^2, \end{aligned}$$

and

$$\begin{aligned} E\eta^3 &= \mu_{Y^3|0,x_0} + 2\mu^3_{Y|0,x_0} - 3\mu_{Y|0,x_0}\mu_{Y^2|0,x_0} \\ &= E\left(\left(Y - \mu_{Y|0,x_0}\right)^3 | D = 0, X = x_0\right) \\ &\equiv v^3_{Y|0,x_0}. \end{aligned}$$

In the next step, we eliminate $s(D, X)$ and $E\left(D^*|D, X\right)$ in equations 31-33 to obtain a restriction only containing $R(D, X)$ and known variables. We will use the following two equations repeatedly.

$$(R + s)\left(\mu_{Y|D,X} - R\right) = \mu_{Y^2|D,X} - E\eta^2 - \mu_{Y|D,X}R \tag{34}$$

$$sE\left(D^*|D, X\right) = \mu_{Y|D,X} - R \tag{35}$$

Notice that

$$s = \frac{\mu_{Y^2|D,X} - \mu^2_{Y|D,X} - \sigma^2_{Y|0,x_0}}{\mu_{Y|D,X} - R} + \mu_{Y|D,X} - R$$

which also implies that we can't identify $s(0, x_0)$ because $\mu_{Y|D=0,x_0} = R(0, x_0)$. Consider

$$
\begin{aligned}
\mu_{Y^3|D,X} &\equiv E\left(Y^3|D, X\right) \\
&= E\left(\left(R(D, X) + s(D, X)D^* + \eta\right)^3 |D, X\right) \\
&= E\left(\left(R + sD^*\right)^3 |D, X\right) + 3E\left(\left(R + sD^*\right)|D, X\right) E\eta^2 + E\left(\eta^3\right) \\
&= R(D, X)^3 + 3R(D, X)^2 s(D, X) E\left(D^*|D, X\right) \\
&\quad +3R(D, X)s(D, X)^2 E\left(D^*|D, X\right) + s(D, X)^3 E\left(D^*|D, X\right) \\
&\quad +3\left[R(D, X) + s(D, X)E\left(D^*|D, X\right)\right] E\eta^2 + E\eta^3 \\
&= R^3 + 3R^2\left(\mu_{Y|D,X} - R\right) + 3Rs\left(\mu_{Y|D,X} - R\right) + s^2\left(\mu_{Y|D,X} - R\right) \\
&\quad +3\mu_{Y|D,X} E\eta^2 + E\eta^3 \\
&= R^3 + 3R^2\left(\mu_{Y|D,X} - R\right) + 2Rs\left(\mu_{Y|D,X} - R\right) + s\left(R + s\right)\left(\mu_{Y|D,X} - R\right) \\
&\quad +3\mu_{Y|D,X} E\eta^2 + E\eta^3 \\
&= R^3 + 3R^2\left(\mu_{Y|D,X} - R\right) + 2Rs\left(\mu_{Y|D,X} - R\right) + s\left(\mu_{Y^2|D,X} - E\eta^2 - \mu_{Y|D,X}R\right) \\
&\quad +3\mu_{Y|D,X} E\eta^2 + E\eta^3 \\
&= R^3 + R^2\left(\mu_{Y|D,X} - R\right) + 2R\left(R + s\right)\left(\mu_{Y|D,X} - R\right) + s\left(\mu_{Y^2|D,X} - E\eta^2 - \mu_{Y|D,X}R\right) \\
&\quad +3\mu_{Y|D,X} E\eta^2 + E\eta^3 \\
&= R^3 + R^2\left(\mu_{Y|D,X} - R\right) + 2R\left(\mu_{Y^2|D,X} - E\eta^2 - \mu_{Y|D,X}R\right) + s\left(\mu_{Y^2|D,X} - E\eta^2 - \mu_{Y|D,X}R\right) \\
&\quad +3\mu_{Y|D,X} E\eta^2 + E\eta^3 \\
&= R^3 + R^2\left(\mu_{Y|D,X} - R\right) + R\left(\mu_{Y^2|D,X} - E\eta^2 - \mu_{Y|D,X}R\right) + \left(R + s\right)\left(\mu_{Y^2|D,X} - E\eta^2 - \mu_{Y|D,X}R\right) \\
&\quad +3\mu_{Y|D,X} E\eta^2 + E\eta^3 \\
&= R\left(\mu_{Y^2|D,X} - E\eta^2\right) + \left(R + s\right)\left(\mu_{Y^2|D,X} - E\eta^2 - \mu_{Y|D,X}R\right) + 3\mu_{Y|D,X} E\eta^2 + E\eta^3 \\
&= R\left(\mu_{Y^2|D,X} - E\eta^2\right) + \frac{\mu_{Y^2|D,X} - E\eta^2 - \mu_{Y|D,X}R}{\left(\mu_{Y|D,X} - R\right)}\left(\mu_{Y^2|D,X} - E\eta^2 - \mu_{Y|D,X}R\right) \\
&\quad +3\mu_{Y|D,X} E\eta^2 + E\eta^3.
\end{aligned}
$$

That is

$$\begin{aligned} 0 &= \left(\mu_{Y^2|D,X} - E\eta^2 - \mu_{Y|D,X}R\right)^2 + \left(\mu_{Y^2|D,X} - E\eta^2\right)\left(\mu_{Y|D,X} - R\right)R \\ &\quad - \left(\mu_{Y^3|D,X} - \left(3\mu_{Y|D,X}E\eta^2 + E\eta^3\right)\right)\left(\mu_{Y|D,X} - R\right). \end{aligned}$$

The restrictions on R simplify to the quadratic equation

$$-\alpha R^2 + \beta R + \gamma = 0,$$

where

$$\begin{aligned} \alpha &= -\left(\mu^2_{Y|D,X} - \left(\mu_{Y^2|D,X} - E\eta^2\right)\right), \\ \beta &= \left(-\left(\mu_{Y^2|D,X} - E\eta^2\right)\mu_{Y|D,X} + \mu_{Y^3|D,X} - \left(3\mu_{Y|D,X}E\eta^2 + E\eta^3\right)\right), \\ \gamma &= \left(\mu_{Y^2|D,X} - E\eta^2\right)^2 - \left(\mu_{Y^3|D,X} - \left(3\mu_{Y|D,X}E\eta^2 + E\eta^3\right)\right)\mu_{Y|D,X}. \end{aligned}$$

Notice that

$$\sigma^2_{Y|D,X} = \mu_{Y^2|D,X} - \mu^2_{Y|D,X},$$

$$\begin{aligned} v^3_{Y|D,X} &\equiv E\left(\left(Y - \mu_{Y|D,X}\right)^3 |D, X\right) \\ &= \mu_{Y^3|D,X} + 2\mu^3_{Y|D,X} - 3\mu_{Y|D,X}\mu_{Y^2|D,X}. \end{aligned}$$

We then simplify the expressions of α, β, and γ as follows:

$$\begin{aligned} \alpha &= -\left(\mu^2_{Y|D,X} - \left(\mu_{Y^2|D,X} - E\eta^2\right)\right) \\ &= \left(\sigma^2_{Y|D,X} - \sigma^2_{Y|0,x_0}\right), \end{aligned}$$

$$
\begin{aligned}
\beta &= \left(-\left(\mu_{Y^2|D,X} - E\eta^2\right)\mu_{Y|D,X} + \mu_{Y^3|D,X} - \left(3\mu_{Y|D,X}E\eta^2 + E\eta^3\right)\right) \\
&= \left(\mu_{Y^3|D,X} - 2\mu_{Y|D,X}E\eta^2 - E\eta^3 - \mu_{Y|D,X}\mu_{Y^2|D,X}\right) \\
&= v^3_{Y|D,X} - 2\mu^3_{Y|D,X} + 3\mu_{Y|D,X}\mu_{Y^2|D,X} - 2\mu_{Y|D,X}E\eta^2 - E\eta^3 - \mu_{Y|D,X}\mu_{Y^2|D,X} \\
&= v^3_{Y|D,X} - E\eta^3 - 2\mu^3_{Y|D,X} - 2\mu_{Y|D,X}E\eta^2 + 2\mu_{Y|D,X}\mu_{Y^2|D,X} \\
&= v^3_{Y|D,X} - E\eta^3 - 2\mu^3_{Y|D,X} - 2\mu_{Y|D,X}E\eta^2 + 2\mu_{Y|D,X}\left(\sigma^2_{Y|D,X} + \mu^2_{Y|D,X}\right) \\
&= v^3_{Y|D,X} - E\eta^3 + 2\mu_{Y|D,X}\left(\sigma^2_{Y|D,X} - E\eta^2\right) \\
&= v^3_{Y|D,X} - v^3_{Y|0,x_0} + 2\mu_{Y|D,X}\left(\sigma^2_{Y|D,X} - \sigma^2_{Y|0,x_0}\right) \\
&= v^3_{Y|D,X} - v^3_{Y|0,x_0} + 2\mu_{Y|D,X}\alpha,
\end{aligned}
$$

$$
\begin{aligned}
\gamma &= \left(\mu_{Y^2|D,X} - E\eta^2\right)^2 - \left(\mu_{Y^3|D,X} - \left(3\mu_{Y|D,X}E\eta^2 + E\eta^3\right)\right)\mu_{Y|D,X} \\
&= \left(\sigma^2_{Y|D,X} + \mu^2_{Y|D,X} - E\eta^2\right)^2 - \left(\mu_{Y^3|D,X} - \left(3\mu_{Y|D,X}E\eta^2 + E\eta^3\right)\right)\mu_{Y|D,X} \\
&= \mu^4_{Y|D,X} + 2\mu^2_{Y|D,X}\left(\sigma^2_{Y|D,X} - E\eta^2\right) + \left(\sigma^2_{Y|D,X} - E\eta^2\right)^2 \\
&\quad -\mu_{Y^3|D,X}\mu_{Y|D,X} + 3\mu^2_{Y|D,X}E\eta^2 + \mu_{Y|D,X}E\eta^3 \\
&= \mu^4_{Y|D,X} + 2\mu^2_{Y|D,X}\sigma^2_{Y|D,X} + \left(\sigma^2_{Y|D,X} - E\eta^2\right)^2 - \mu_{Y^3|D,X}\mu_{Y|D,X} + \mu^2_{Y|D,X}E\eta^2 + \mu_{Y|D,X}E\eta^3 \\
&= \mu^4_{Y|D,X} + 2\mu^2_{Y|D,X}\sigma^2_{Y|D,X} + \left(\sigma^2_{Y|D,X} - E\eta^2\right)^2 \\
&\quad - \left(v^3_{Y|D,X} - 2\mu^3_{Y|D,X} + 3\mu_{Y|D,X}\mu_{Y^2|D,X}\right)\mu_{Y|D,X} + \mu^2_{Y|D,X}E\eta^2 + \mu_{Y|D,X}E\eta^3 \\
&= \mu^4_{Y|D,X} + 2\mu^2_{Y|D,X}\sigma^2_{Y|D,X} + \left(\sigma^2_{Y|D,X} - E\eta^2\right)^2 \\
&\quad +2\mu^4_{Y|D,X} - 3\mu^2_{Y|D,X}\mu_{Y^2|D,X} + \mu^2_{Y|D,X}E\eta^2 + \mu_{Y|D,X}\left(E\eta^3 - v^3_{Y|D,X}\right) \\
&= \mu^4_{Y|D,X} + 2\mu^2_{Y|D,X}\sigma^2_{Y|D,X} + \left(\sigma^2_{Y|D,X} - E\eta^2\right)^2 \\
&\quad +2\mu^4_{Y|D,X} - 3\mu^2_{Y|D,X}\left(\sigma^2_{Y|D,X} + \mu^2_{Y|D,X}\right) + \mu^2_{Y|D,X}E\eta^2 + \mu_{Y|D,X}\left(E\eta^3 - v^3_{Y|D,X}\right) \\
&= \left(\sigma^2_{Y|D,X} - E\eta^2\right)^2 - \mu^2_{Y|D,X}\left(\sigma^2_{Y|D,X} - E\eta^2\right) - \mu_{Y|D,X}\left(v^3_{Y|D,X} - E\eta^3\right) \\
&= \left(\sigma^2_{Y|D,X} - \sigma^2_{Y|0,x_0}\right)^2 - \mu^2_{Y|D,X}\left(\sigma^2_{Y|D,X} - \sigma^2_{Y|0,x_0}\right) - \mu_{Y|D,X}\left(v^3_{Y|D,X} - v^3_{Y|0,x_0}\right) \\
&= \alpha^2 - \mu^2_{Y|D,X}\alpha - \mu_{Y|D,X}\left(\beta - 2\mu_{Y|D,X}\alpha\right) \\
&= \alpha^2 + \mu^2_{Y|D,X}\alpha - \mu_{Y|D,X}\beta.
\end{aligned}
$$

In summary, we have

$$-\alpha R^2 + \beta R + \gamma = 0$$

$$
\begin{aligned}
\alpha &= \sigma^2_{Y|D,X} - \sigma^2_{Y|0,x_0} \\
\beta &= v^3_{Y|D,X} - v^3_{Y|0,x_0} + 2\mu_{Y|D,X}\alpha \\
\gamma &= \alpha^2 + \mu^2_{Y|D,X}\alpha - \mu_{Y|D,X}\beta
\end{aligned}
$$

That means

$$R = \frac{\beta + \sqrt{\beta^2 + 4\alpha\gamma}}{2\alpha} \text{ or } \frac{\beta - \sqrt{\beta^2 + 4\alpha\gamma}}{2\alpha}.$$

In fact, we may show that equations 35 and 34 implies

$$\alpha \geq 0$$

Consider

$$\begin{aligned} s &= \frac{\mu_{Y^2|D,X} - \mu^2_{Y|D,X} - E\eta^2}{\mu_{Y|D,X} - R} + \mu_{Y|D,X} - R \\ &= \frac{\alpha}{\mu_{Y|D,X} - R} + \mu_{Y|D,X} - R \end{aligned}$$

and

$$\begin{aligned} E\left(D^*|D,X\right) &= \frac{\mu_{Y|D,X} - R}{s} \\ &= \frac{\left(\mu_{Y|D,X} - R\right)^2}{\left(\mu_{Y|D,X} - R\right)^2 + \alpha}. \end{aligned}$$

Therefore, $0 \leq E\left(D^*|D,X\right) \leq 1$ implies that $\alpha \geq 0$.

The last step is to eliminate one of the two roots to achieve point identification. Notice that

$$E\left(Y|D^*,D,X\right) = R(D,X) + s(D,X)D^*.$$

Assumption B2.4 implies that

$$s(D,X) \geq 0.$$

Consider

$$\begin{aligned} \mu_{Y|D,X} &= R + sE\left(D^*|D,X\right) \\ &= R\left[1 - E\left(D^*|D,X\right)\right] + (R+s)\,E\left(D^*|D,X\right). \end{aligned}$$

Therefore, $0 \leq E\left(D^*|D,X\right) \leq 1$ and $s(D,X) \geq 0$ imply

$$R \leq \mu_{Y|D,X} \leq s + R,$$

Thus, we may identify R as the smaller root if $\mu_{Y|D,X}$ is between the two roots. , i.e.,

$$-\alpha\mu^2_{Y|D,X} + \beta\mu_{Y|D,X} + \gamma \geq 0,$$

which holds because

$$\begin{aligned} & -\alpha\mu_{Y|D,X}^2 + \beta\mu_{Y|D,X} + \gamma \\ = \; & -\alpha\mu_{Y|D,X}^2 + \beta\mu_{Y|D,X} + \alpha^2 + \mu_{Y|D,X}^2\alpha - \mu_{Y|D,X}\beta \\ = \; & \alpha^2 \geq 0. \end{aligned}$$

Therefore, we have

$$R(D, X) = \frac{\beta - \sqrt{\beta^2 + 4\alpha\gamma}}{2\alpha}.$$

Notice that R equals the larger root if $s(D, X) \leq 0$. The function $s(D, X)$ then follows.

Discrete Limiting Distributions for equation (16). Let

$$\begin{aligned} \widehat{\alpha}(x) &= \left(\widehat{\mu}_{Y,V,X,1}, \widehat{\mu}_{Y,V,X,0}, \widehat{\mu}_{Y,X,1}, \widehat{\mu}_{Y,X,0}, \widehat{\mu}_{V,X,1}, \widehat{\mu}_{V,X,0}, \widehat{\mu}_{X,1}, \widehat{\mu}_{X,0}, \widehat{\mu}_{VU}, \widehat{\mu}_{U}\right)^T \\ \alpha_0 &= E\left[\widehat{\alpha}(x)\right] \\ \widehat{R}(d, \widehat{\alpha}(x)) &\equiv \frac{\left(\widehat{\mu}_{Y,V,X,1}^{d}\widehat{\mu}_{Y,V,X,0}^{1-d}\right)\widehat{\mu}_{U} - \left(\widehat{\mu}_{Y,X,1}^{d}\widehat{\mu}_{Y,X,0}^{1-d}\right)\widehat{\mu}_{VU}}{\left(\widehat{\mu}_{V,X,1}^{d}\widehat{\mu}_{V,X,0}^{1-d}\right)\widehat{\mu}_{U} - \left(\widehat{\mu}_{X,1}^{d}\widehat{\mu}_{X,0}^{1-d}\right)\widehat{\mu}_{VU}} \\ \widehat{r}(x) &= \widehat{R}(1, \widehat{\alpha}(x)) - \widehat{R}(0, \widehat{\alpha}(x)) \end{aligned}$$

$$\begin{aligned} \gamma &= \frac{\partial}{\partial t} R\left(d, \alpha_0 + t\left(\widehat{\alpha} - \alpha_0\right)\right)\Big|_{t=0} \\ &\equiv G\left(d, \alpha_0\right)^T\left(\widehat{\alpha} - \alpha_0\right) \end{aligned}$$

$$V\left(\widehat{\alpha}(x)\right) = n \times E\left[\left(\widehat{\alpha} - \alpha_0\right)\left(\widehat{\alpha} - \alpha_0\right)^T\right]$$

Assuming independent, identically distributed draws and existence of $V\left(\widehat{\alpha}(x)\right)$, by the Lindeberg-Levy central limit theorem and the delta method

$$\begin{aligned} \sqrt{n}\left[\widehat{R}(d, x) - R(d, x)\right] &\rightarrow {}^{d} N\left(0, \Omega_R\right) \\ \Omega_R &= G\left(d, \alpha_0(x)\right)^T V\left(\widehat{\alpha}(x)\right) G\left(d, \alpha_0(x)\right) \end{aligned}$$

and

$$\begin{aligned} \sqrt{n}\left[\widehat{r}(x) - r(x)\right] &\rightarrow {}^{d} N\left(0, \Omega_r\right) \\ \Omega_r &= \left[G\left(1, \alpha_0(x)\right) - G\left(0, \alpha_0(x)\right)\right]^T V\left(\widehat{\alpha}(x)\right)\left[G\left(1, \alpha_0(x)\right) - G\left(0, \alpha_0(x)\right)\right] \end{aligned}$$

Table 6: R(0,X), Nonparametric and Semiparametric Corollary 1 IV Estimates

	Some college	Associate degree	Bachelor's degree
R0 nonparametric	2.072 (0.01514)	2.125 (0.009659)	2.143 (0.007992)
$R0_t$ nonparametric	2.065 (0.01536)	2.125 (0.01016)	2.144 (0.008568)
$R0_{q1}$ nonparametric	1.863 (0.02520)	1.940 (0.01939)	1.975 (0.01834)
$R0_{med}$ nonparametric	2.003 (0.03859)	2.089 (0.03224)	2.143 (0.02788)
$R0_{q3}$ nonparametric	2.309 (0.02681)	2.326 (0.01762)	2.319 (0.01663)
R0 semi, linear	2.025 (0.01174)	2.094 (0.008751)	2.114 (0.007448)

Table 7: R(1,X), Nonparametric and Semiparametric Corollary 1 IV Estimates

	Some college	Associate degree	Bachelor's degree
R1 nonparametric	2.142 (0.03011)	2.301 (0.02844)	2.184 (0.07708)
$R1_t$ nonparametric	2.152 (0.02985)	2.324 (0.02751)	2.240 (0.07308)
$R1_{q1}$ nonparametric	1.997 (0.04430)	2.179 (0.05181)	2.131 (0.09785)
$R1_{med}$ nonparametric	2.173 (0.04633)	2.382 (0.02936)	2.312 (0.05680)
$R1_{q3}$ nonparametric	2.340 (0.04635)	2.455 (0.03637)	2.424 (0.05945)
R1 semi, linear	2.188 (0.02898)	2.351 (0.02604)	2.339 (0.04339)

Table 8: R(0,X), Nonparametric and Semiparametric Corollary 2 IV Estimates

	Some college	Associate degree	Bachelor's degree
R0 nonparametric	2.072 (0.01514)	2.125 (0.009665)	2.143 (0.007997)
$R0_t$ nonparametric	2.065 (0.01536)	2.125 (0.01016)	2.144 (0.008579)
$R0_{q1}$ nonparametric	1.863 (0.02520)	1.939 (0.01940)	1.975 (0.01834)
$R0_{med}$ nonparametric	2.003 (0.03859)	2.089 (0.03225)	2.143 (0.02788)
$R0_{q3}$ nonparametric	2.309 (0.02681)	2.326 (0.01763)	2.319 (0.01665)
R0 semi, linear	2.025 (0.01174)	2.094 (0.008754)	2.114 (0.007451)

Table 9: R(1,X), Nonparametric and Semiparametric Corollary 2 IV Estimates

	Some college	Associate degree	Bachelor's degree
R1 nonparametric	2.142 (0.03011)	2.295 (0.3326)	2.268 (1.918)
$R1_t$ nonparametric	2.152 (0.02986)	2.319 (0.04103)	2.223 (0.1219)
$R1_{q1}$ nonparametric	1.997 (0.04430)	2.181 (0.06094)	2.092 (0.1694)
$R1_{med}$ nonparametric	2.173 (0.04633)	2.380 (0.04084)	2.189 (0.1026)
$R1_{q3}$ nonparametric	2.340 (0.04635)	2.449 (0.04508)	2.397 (0.1731)
R1 semi, linear	2.188 (0.02898)	2.341 (0.03397)	2.267 (1.149)

Table 10: R(0,X), Nonparametric and Semiparametric Theorem 2 Estimates Without IV

	Some college	Associate degree	Bachelor's degree
R0 nonparametric	2.078 (0.01383)	2.126 (0.009571)	2.144 (0.007897)
$R0_t$ nonparametric	2.074 (0.01447)	2.123 (0.01025)	2.148 (0.008459)
$R0_{q1}$ nonparametric	1.891 (0.02270)	1.942 (0.01916)	1.974 (0.01820)
$R0_{med}$ nonparametric	2.022 (0.03761)	2.095 (0.03227)	2.146 (0.02767)
$R0_{q3}$ nonparametric	2.288 (0.02247)	2.321 (0.01708)	2.324 (0.01620)

Table 11: R(1,X), Nonparametric and Semiparametric Theorem 2 Estimates Without IV

	Some college	Associate degree	Bachelor's degree
R1 nonparametric	1.666 (28.66)	2.318 (2.915)	2.269 (18.27)
$R1_t$ nonparametric	2.141 (0.1418)	2.310 (0.1719)	2.227 (0.2483)
$R1_{q1}$ nonparametric	1.832 (0.1459)	2.069 (0.2132)	1.525 (0.3052)
$R1_{med}$ nonparametric	2.223 (0.07273)	2.247 (0.08727)	2.222 (0.07330)
$R1_{q3}$ nonparametric	2.419 (0.09065)	2.501 (0.1239)	2.552 (0.2033)

References

[1] AI, C. AND X. CHEN (2003), "Efficient Estimation of Models With Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795-1844.

[2] BLUNDELL, R. AND J. L. POWELL, (2004), "Endogeneity in Semiparametric Binary Response Models" *Review of Economic Studies*, 71, 655-679.

[3] CHEN, X., Y. HU AND A. LEWBEL, (2007), "Nonparametric Identification of Regression Models Containing a Misclassified Dichotomous Regressor Without Instruments," unpublished manuscript.

[4] CHEN, X., O. LINTON, AND I. VAN KEILEGOM, (2003) "Estimation of Semiparametric Models when the Criterion Function Is Not Smooth," *Econometrica*, 71, 1591-1608,

[5] DAS, M., (2004), "Instrumental Variables Estimators of Nonparametric Models With Discrete Endogenous Regressors," *Journal of Econometrics*, 124, 335-361.

[6] FLORENS, J.-P. AND L. MALAVOLTI, (2003), "Instrumental Regression with Discrete Variables," unpublished manuscript.

[7] GIBRAT, R. (1931), Les Inegalites Economiques, Librairie du Recueil Sirey, Paris

[8] HECKMAN, J. J. (1979), "Sample selection bias as a specification error," *Econometrica*, 47, 153–161.

[9] HECKMAN, J. J., H. ICHIMURA AND P. TODD, (1998), "Matching as an Econometric Evaluations Estimator, Review of Economic Studies, 65, 261-294.

[10] HU, Y. (2006), "Identification and estimation of nonlinear models with misclassification error using instrumental variables," U. Texas at Austin unpublished manuscript.

[11] KANE, T. J., AND C. E. ROUSE, (1995), "Labor market returns to two- and four- year college," *American Economic Review*, 85, 600-614

[12] KANE, T. J., C. E. ROUSE, AND D. STAIGER, (1999), "Estimating Returns to Schooling When Schooling is Misreported," NBER working paper #7235.

[13] LEWBEL, A., (2007a), “Estimation of Average Treatment Effects With Misclassification,” *Econometrica*, 75, 537-551forthcoming.

[14] LEWBEL, A., (2007b), "A Local Generalized Method of Moments Estimator,” *Economics Letters*, 94, 124-128.

[15] MAHAJAN, A. (2006) ”Identification and Estimation of Regression Models with Misclassification," *Econometrica*, 74, 631-665.

[16] NEWEY, W. K., (1994), “Kernel Estimation of Partial Means and a General Variance Estimator,” *Econometric Theory*, 10, 233-253.

[17] NEWEY, W. K. AND J. L. POWELL, (2003), “Instrumental Variable Estimation of Nonparametric Models,” *Econometrica*, 71, 1565-1578.

[18] ROSENBAUM, P. AND D. RUBIN, (1984), “Reducing Bias in Observational Studies Using Subclassification on the Propensity Score,” *Journal of the American Statistical Association*, 79, 516-524.

[19] RUBIN, D. B. (1974), “Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies, ” *Journal of Educational Psychology*, 76, 688-701.

Nonparametric Identification and Estimation of Nonclassical Errors-in-Variables Models Without Additional Information*

Xiaohong Chen† Yale University

Yingyao Hu‡ Johns Hopkins University

Arthur Lewbel§ Boston College

First version: November 2006; Revised October 2007.

Abstract

This paper considers identification and estimation of a nonparametric regression model with an unobserved discrete covariate. The sample consists of a dependent variable and a set of covariates, one of which is discrete and arbitrarily correlates with the unobserved covariate. The observed discrete covariate has the same support as the unobserved covariate, and can be interpreted as a proxy or mismeasure of the unobserved one, but with a nonclassical measurement error that has an unknown distribution. We obtain nonparametric identification of the model given monotonicity of the regression function and a rank condition that is directly testable given the data. Our identification strategy does not require additional sample information, such as instrumental variables or a secondary sample. We then estimate the model via the method of sieve maximum likelihood, and provide root-n asymptotic normality and semiparametric efficiency of smooth functionals of interest. Two small simulations are presented to illustrate the identification and the estimation results.

Keywords: Errors-In-Variables (EIV), Identification; Nonclassical measurement error; Nonparametric regression; Sieve maximum likelihood.

*We thank participants at June 2007 North American Summer Meetings of the Econometric Society at Duke for helpful comments. Chen acknowledges support from NSF.

†Department of Economics, Yale University, Box 208281, New Haven, CT 06520-8281, USA. Tel: 203-432-5852. Email: xiaohong.chen@yale.edu.

‡Department of Economics, Johns Hopkins University, 440 Mergenthaler Hall, 3400 N. Charles Street, Baltimore, MD 21218, USA. Tel: 410-516-7610. Email: yhu@jhu.edu.

§Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467 USA. Tel: 617-522-3678. Email: lewbel@bc.edu.

1 INTRODUCTION

We consider identification and estimation of the nonparametric regression model

$$Y = m(X^*) + \eta, \quad E[\eta|X^*] = 0 \tag{1.1}$$

where Y and X^* are scalars and X^* is not observed. We assume X^* is discretely distributed, so for example X^* could be categorical, qualitative, or count data. We observe a random sample of Y and a scalar X, where X could be arbitrarily correlated with the unobserved X^*, and η is independent of X and X^*. We assume X has the same support as X^*. The extension to

$$Y = m(X^*, W) + \eta, \quad E[\eta|X^*, W] = 0$$

where W is an additional vector of observed error-free covariates is immediate (and is included in the estimation section) because our assumptions and identification results for model (1.1) can be all restated as conditional upon W. Discreteness of X and X^* (with same supports) means that the measurement error $X - X^*$ will be *nonclassical*, in particular, it will not be independent of X^* and generally has nonzero mean. see, e.g., Bound, Brown and Mathiowetz (2001) for a review of nonclassical measurement errors.

This type of discrete measurement error is common in many data sets, in particular, it arises in contexts where X^* indexes or classifies the group that an individual belongs to, which is sometimes misreported, yielding classification errors. For example, Kane and Rouse (1995) find that school transcript reports of years of schooling often contain errors, so X^* could indicate one's actual years of schooling and X is the transcript report. Finney (1964) discusses misclassification in biological assay. Gustman and Steinmeier (2004) report that many individuals that actually have a defined benefit retirement plan claimed to have a defined contribution plan, and vice versa, so here X^* and X would binary indicators of actual versus reported pension type. Hirsch and MacPherson (2003) document misclassification in surveys of union status. More generally X^* and X could be the actual and reported values in any count data or multiple choice survey question, with differences between X^* and X arising from either imperfect knowledge or recording and transcription errors. Balke and Pearl (1997) model imperfect compliance, where X is some assigned experimental treatment that differs from the actual treatment received X^* because of compliance difficulties.

Many estimators and associated empirical analyses have been proposed to deal with misclassified discrete variables. Examples in addition to the above citations include Aigner (1973), Chua and Fuller (1987), Hsiao (1991), Poterba and Summers (1995), Bollinger (1996), Hausman, Abrevaya, and Scott-Morton (1998), Lewbel (2000), (2007), Hu (2006), and Mahajan (2006). However, to the best of our knowledge, there is no published work that allows for nonparametric point identification and estimation of nonparametric regression models with nonclassically mismeasured discrete regressors without parametric restrictions or additional sample information such as instrumental variables, repeated measurements, or validation data, which our paper provides. In short, we nonparametrically recover and hence identify the conditional density $f_{Y|X^*}$ (or equivalently, the regression function m and the distribution of the regression error η) just from the observed joint distribution $f_{Y,X}$, while imposing minimal restrictions on the joint distribution of X^* and X. We will also recover $f_{X|X^*}$ and f_{X^*}which respectively imply identifying the conditional distribution of

the measurement error, and the marginal distribution of the unobserved regressor f_{X^*}, and also implies identification of the joint distributions f_{Y,X^*} and f_{X,X^*}.

Although we interpret X as a measure of X^* that is contaminated by measurement or misclassification error, more generally X^* could represent some latent, unobserved quantifiable discrete variable like a health status or life expectancy quantile, and X could be some observed proxy such as a body mass index quantile or the response to a health related categorical survey question. Equation (1.1) could then be interpreted as a latent factor model $Y = m^* + \eta$, with two unobserved independent factors m^* and η, with identification based on observing the proxy X and on existence of a measurable function $m(\cdot)$ such that $m^* = m(X^*)$.

The relationship between the latent model $f_{Y|X^*}$ and the observed density $f_{Y,X}$ is

$$f_{Y,X}(y,x) = \int f_{Y|X^*}(y|x^*) f_{X,X^*}(x,x^*)\, dx^*. \tag{1.2}$$

Existing papers identifying the latent model $f_{Y|X^*}$ use one of the three methods: i) assuming the measurement error structure $f_{X|X^*}$ belongs to a parametric family (see, e.g., Fuller (1987), Bickel and Ritov (1987), Hsiao (1991), Murphy and Van der Vaart (1996), Wang, Lin, Gutierrez and Carroll (1998), Liang, Hardle and Carroll (1999), Taupin (2001), Hong and Tamer (2003), Liang and Wang (2005)); ii) assuming there exists an additional exogenous variable Z in the sample (such as an instrument or a repeated measure) that does not enter the latent model $f_{Y|X^*}$, and exploiting assumed restrictions on $f_{Y|X^*,Z}$ and $f_{X,X^*,Z}$ to identify $f_{Y|X^*}$ given the joint distribution of $\{y,x,z\}$ (see, e.g., Hausman, Ichimura, Newey and Powell (1991), Li and Vuong (1998), Li (2002), Wang (2004), Schennach (2004), Carroll, Ruppert, Crainiceanu, Tosteson and Karagas (2004), Mahajan (2006), Lewbel (2007), Hu (2006) and Hu and Schennach (2006)); or iii) assuming a secondary sample to provide information on f_{X,X^*} to permit recovery of $f_{Y|X^*}$ from the observed $f_{Y,X}$ in the primary sample (see, e.g., Carroll and Stefanski (1990), Lee and Sepanski (1995), Chen, Hong, and Tamer (2005), Chen, Hong, and Tarozzi (2007), Hu and Ridder (2006)). Detailed reviews on existing approaches and results can be found in several recent books and surveys on measurement error models; see, e.g., Carroll, Ruppert, Stefanski and Crainiceanu (2006), Chen, Hong and Nekipelov (2007), Bound, Brown and Mathiowetz (2001), Wansbeek and Meijer (2000), and Cheng and Van Ness (1999).

In this paper, we obtain identification by exploiting nonparametric features of the latent model $f_{Y|X^*}$, such as independence of the regression error term η and discreteness of X^*. Our results are useful because many applications specify the latent model of interest $f_{Y|X^*}$, while often little is known about f_{X,X^*}, that is, about the nature of the measurement error or the exact relationship between the unobserved latent X^* and a proxy X. In addition, our key "rank" condition for identification is directly testable from the data.

Our identification method is based on characteristic functions. Suppose X and X^* have support $\mathcal{X} = \{1,2,...,J\}$. Then by equation (1.1), $\exp(itY) = \exp(it\eta)\sum_{j=1}^{J} 1(X^* = j)\exp[im(j)t]$ for any given constant t, where $1()$ is the indicator function that equals one if

its argument is true and zero otherwise. This equation and independence of η yield moments

$$E\left[\exp\left(itY\right)f_X(x)\mid X=x\right]=E\left[\exp\left(it\eta\right)\right]\sum_{x^*=1}^{J}f_{X,X^*}\left(x,x^*\right)\exp\left[im\left(x^*\right)t\right] \tag{1.3}$$

Evaluating equation (1.3) for $t\in\{t_1,...,t_K\}$ and $x\in\{1,2,...,J\}$ provides KJ equations in J^2+J+K unknown constants. These unknown constants are the values of $f_{X,X^*}\left(x,x^*\right)$, $m\left(x^*\right)$, and $E\left[\exp\left(it\eta\right)\right]$ for $t\in\{t_1,...,t_K\}$, $x\in\{1,2,...,J\}$, and $x^*\in\{1,2,...,J\}$. Given a large enough value of K, these moments provide more equations than unknowns. We provide sufficient regularity assumptions to ensure existence of some set of constants $\{t_1,...,t_K\}$ such that these equations do not have multiple solutions, and the resulting unique solution to these equations provides identification of $m(\cdot)$, f_η and f_{X,X^*}, and hence of $f_{Y|X^*}$.

As equation (1.3) shows, our identification results depend on many moments of Y or equivalently of η, rather than just on the conditional mean restriction $E[\eta|X^*]=0$ that would suffice for identification if X^* were observed. Previous results, for example, Reiersol (1950), Kendall and Stuart (1979), and Lewbel (1997), exist that obtain identification based on higher moments as we do (without instruments, repeated measures, or validation data), but all these previous results have assumed either classical measurement error or/and parametric restrictions.

Equation (1.3) implies that independence of the regression error η is actually stronger than necessary for identification, since e.g. we would obtain the same equations used for identification if we only had $E\left[\exp\left(it\eta\right)|X^*,X\right]=E\left[\exp\left(it\eta\right)\right]$ for $t\in\{t_1,...,t_K\}$. To illustrate this point further, we provide an alternative identification result for the dichotomous X^* case without the independence assumption, and in this case we can identify (solve) for all the unknowns in closed-form.

Estimation could be based directly on equation (1.3) using, for example, Hansen's (1982) Generalized Method of Moments (GMM). However, this would require knowing or choosing constants $t_1,...,t_K$. Moreover, under the independence assumption of η and X^*, we have potentially infinitely many constants t that solves equation (1.3); hence GMM estimation using finitely many such t's will not be efficient in general. One could apply the infinite-dimensional Z-estimation as described in Van der Vaart and Wellner (1996), here we instead apply the method of sieve Maximum Likelihood (ML) of Grenander (1981) and Geman and Hwang (1982), which does not require knowing or choosing constants $t_1,...,t_K$, and easily allows for an additional vector of error-free covariates W. The sieve ML estimator essentially replaces the unknown functions f_η, m, and $f_{X^*|X,W}$ with polynomials, Fourier series, splines, wavelets or other sieve approximators, and estimates the parameters of these approximations by maximum likelihood. By simple applications of the general theory on sieve MLE developed in Wong and Shen (1995), Shen (1997), Van de Geer (1993, 2000) and others, we provide consistency and convergence rate of the sieve MLE, and root-n asymptotic normality and semiparametric efficiency of smooth functionals of interest, such as the weighted averaged derivatives of the latent nonparametric regression function $m(X^*,W)$, or the finite-dimensional parameters (β) in a semiparametric specification of $m(X^*,W;\beta)$.

The rest of this paper is organized as follows. Section 2 provides the identification results. Section 3 describes the sieve ML estimator and presents its large sample properties. Section

4 provides two small simulation studies. Section 5 briefly concludes and all the proofs are in the appendix.

2 NONPARAMETRIC IDENTIFICATION

Our basic nonparametric regression model is equation (1.1) with scalar Y and $X^* \in \mathcal{X} = \{1, 2, ..., J\}$. We observe a random sample of $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, where X is a proxy of X^*. The goal is to consider restrictions on the latent model $f_{Y|X^*}$ that suffice to nonparametrically identify $f_{Y|X^*}$ and $f_{X|X^*}$ from $f_{Y|X}$.

Assumption 2.1 $X \perp \eta | X^*$.

This assumption implies that the measurement error $X - X^*$ is independent of the dependent variable Y conditional on the true value X^*. In other words, we have $f_{Y|X^*,X}(y|x^*, x) = f_{Y|X^*}(y|x^*)$ for all $(x, x^*, y) \in \mathcal{X} \times \mathcal{X} \times \mathcal{Y}$. This is equivalent to the classical measurement error property that the outcome Y depends only on the true X^* and not on the mismeasured version X.

Assumption 2.2 $X^* \perp \eta$.

This assumption implies that the regression error η is independent of the regressor X^* so $f_{Y|X^*}(y|x^*) = f_\eta(y - m(x^*))$. The relationship between the observed density and the latent ones are then:

$$f_{Y,X}(y, x) = \sum_{x^*=1}^{J} f_\eta(y - m(x^*)) f_{X,X^*}(x, x^*). \tag{2.1}$$

Assumption 2.2 rules out heteroskedasticity or other heterogeneity of the regression error η, but allows its density f_η to be completely unknown and nonparametric. The regression error η is not required to be continuously distributed, but the rank condition discussed below does place a lower bound on the number of points in the support of η. We will later show that this assumption can be relaxed in a couple of different ways, e.g., as noted in the introduction this assumption can be replaced with $E[\exp(it\eta)|X^*, X] = E[\exp(it\eta)]$ for a certain finite set of values of t. For dichotomous (binary) X^*, we show Assumption 2.2 can alternatively be weakened to just requiring $E(\eta^k|X^*) = E(\eta^k)$ for $k = 2, 3$.

Let ϕ denote a characteristic function (ch.f.). Equation (2.1) is equivalent to

$$\phi_{Y,X=x}(t) = \phi_\eta(t) \sum_{x^*=1}^{J} \exp(itm(x^*)) f_{X,X^*}(x, x^*), \tag{2.2}$$

for all real-valued t, where $\phi_{Y,X=x}(t) = \int \exp(ity) f_{Y,X}(y, x) dy$ and $x \in \mathcal{X}$. Since η may not be symmetric, $\phi_\eta(t) = \int \exp(it\eta) f_\eta(\eta) d\eta$ need not be real-valued. We therefore let

$$\phi_\eta(t) \equiv |\phi_\eta(t)| \exp(ia(t)),$$

where

$$|\phi_\eta(t)| \equiv \sqrt{[\mathrm{Re}\{\phi_\eta(t)\}]^2 + [\mathrm{Im}\{\phi_\eta(t)\}]^2}, \quad a(t) \equiv \arccos \frac{\mathrm{Re}\{\phi_\eta(t)\}}{|\phi_\eta(t)|}.$$

We then have for any real-valued scalar t,

$$\phi_{Y,X=x}(t) = \left|\phi_{\eta}(t)\right| \sum_{x^*=1}^{J} \exp\left(itm(x^*) + ia(t)\right) f_{X,X^*}(x, x^*). \tag{2.3}$$

Define

$$F_{X,X^*} = \begin{pmatrix} f_{X,X^*}(1,1) & f_{X,X^*}(1,2) & ... & f_{X,X^*}(1,J) \\ f_{X,X^*}(2,1) & f_{X,X^*}(2,2) & ... & f_{X,X^*}(2,J) \\ ... & ... & ... & ... \\ f_{X,X^*}(J,1) & f_{X,X^*}(J,2) & ... & f_{X,X^*}(J,J) \end{pmatrix},$$

and for any real-valued vector $\mathbf{t} = (0, t_2, ..., t_J)$, we define

$$\Phi_{Y,X}(\mathbf{t}) = \begin{pmatrix} f_X(1) & \phi_{Y,X=1}(t_2) & ... & \phi_{Y,X=1}(t_J) \\ f_X(2) & \phi_{Y,X=2}(t_2) & ... & \phi_{Y,X=2}(t_J) \\ ... & ... & ... & ... \\ f_X(J) & \phi_{Y,X=J}(t_2) & ... & \phi_{Y,X=J}(t_J) \end{pmatrix}, D_{|\phi|}(\mathbf{t}) = \begin{pmatrix} 1 & 0 & ... & 0 \\ 0 & \left|\phi_{\eta}(t_2)\right| & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \left|\phi_{\eta}(t_J)\right| \end{pmatrix},$$

and define $m_j = m(j)$ for $j = 1, 2, ..., J$,

$$\Phi_{m,a}(\mathbf{t}) = \begin{pmatrix} 1 & \exp\left(it_2 m_1 + ia(t_2)\right) & ... & \exp\left(it_J m_1 + ia(t_J)\right) \\ 1 & \exp\left(it_2 m_2 + ia(t_2)\right) & ... & \exp\left(it_J m_2 + ia(t_J)\right) \\ ... & ... & ... & ... \\ 1 & \exp\left(it_2 m_J + ia(t_2)\right) & ... & \exp\left(it_J m_J + ia(t_J)\right) \end{pmatrix}.$$

With these matrix notations, for any real-valued vector $\mathbf{t}$ equation (2.3) is equivalent to

$$\Phi_{Y,X}(\mathbf{t}) = F_{X,X^*} \times \Phi_{m,a}(\mathbf{t}) \times D_{|\phi|}(\mathbf{t}). \tag{2.4}$$

Equation (2.4) relates the known parameters $\Phi_{Y,X}(\mathbf{t})$ (which may be interpreted as reduced form parameters of the model) to the unknown structural parameters F_{X,X^*}, $\Phi_{m,a}(\mathbf{t})$, and $D_{|\phi|}(\mathbf{t})$. Equation (2.4) provides a sufficient number of equality constraints to identify the structural parameters given the reduced form parameters, so what is required are sufficient invertibility or rank restrictions to rule out multiple solutions of these equations.

To provide these conditions, consider both the real and imaginary parts of $\Phi_{Y,X}(\mathbf{t})$. Since $D_{|\phi|}(\mathbf{t})$ is real by definition, we have

$$\mathrm{Re}\{\Phi_{Y,X}(\mathbf{t})\} = F_{X,X^*} \times \mathrm{Re}\{\Phi_{m,a}(\mathbf{t})\} \times D_{|\phi|}(\mathbf{t}), \tag{2.5}$$

and

$$\mathrm{Im}\{\Phi_{Y,X}(\mathbf{t})\} = F_{X,X^*} \times \mathrm{Im}\{\Phi_{m,a}(\mathbf{t})\} \times D_{|\phi|}(\mathbf{t}). \tag{2.6}$$

The matrices $\mathrm{Im}\{\Phi_{Y,X}(\mathbf{t})\}$ and $\mathrm{Im}\{\Phi_{m,a}(\mathbf{t})\}$ are not invertible because their first columns are zeros, so we replace equation (2.6) with

$$(\mathrm{Im}\{\Phi_{Y,X}(\mathbf{t})\} + \Upsilon_X) = F_{X,X^*} \times (\mathrm{Im}\{\Phi_{m,a}(\mathbf{t})\} + \Upsilon) \times D_{|\phi|}(\mathbf{t}), \tag{2.7}$$

where

$$\Upsilon_X = \begin{pmatrix} f_X(1) & 0 & ... & 0 \\ f_X(2) & 0 & ... & 0 \\ ... & ... & ... & ... \\ f_X(J) & 0 & ... & 0 \end{pmatrix} \text{ and } \Upsilon = \begin{pmatrix} 1 & 0 & ... & 0 \\ 1 & 0 & ... & 0 \\ ... & ... & ... & ... \\ 1 & 0 & ... & 0 \end{pmatrix}.$$

Equation (2.7) holds because $F_{X,X^*} \times \Upsilon = \Upsilon_X$ and $\Upsilon \times D_{|\phi|}(\mathbf{t}) = \Upsilon$. Let $C_{\mathbf{t}} \equiv (\text{Re}\{\Phi_{Y,X}(\mathbf{t})\})^{-1} \times (\text{Im}\{\Phi_{Y,X}(\mathbf{t})\} + \Upsilon_X)$.

Assumption 2.3 *(rank). There is a real-valued vector* $\mathbf{t} = (0, t_2, ..., t_J)$ *such that: (i)* $\text{Re}\{\Phi_{Y,X}(\mathbf{t})\}$ *and* $(\text{Im}\{\Phi_{Y,X}(\mathbf{t})\} + \Upsilon_X)$ *are invertible; (ii) For any real-valued* $J \times J-$*diagonal matrices* $D_k = Diag\,(0, d_{k,2}, ..., d_{k,J})$*, if* $D_1 + C_{\mathbf{t}} \times D_1 \times C_{\mathbf{t}} + D_2 \times C_{\mathbf{t}} - C_{\mathbf{t}} \times D_2 = 0$ *then* $D_k = 0$ *for* $k = 1, 2$.

We call Assumption 2.3 the rank condition, because it is analogous to the rank condition for identification in linear models, and in particular implies identification of the two diagonal matrices

$$D_{\partial \ln|\phi|}(\mathbf{t}) = Diag\left(0, \frac{\partial}{\partial t} \ln\left|\phi_\eta(t_2)\right|, ..., \frac{\partial}{\partial t} \ln\left|\phi_\eta(t_J)\right|\right)$$

and

$$D_{\partial a}(\mathbf{t}) = Diag\left(0, \frac{\partial}{\partial t} a(t_2), ..., \frac{\partial}{\partial t} a(t_J)\right).$$

Assumption 2.3(ii) is rather complicated, but can be replaced by some simpler sufficient alternatives, which we will describe later. Note also that the rank condition, Assumption 2.3, is testable, since it is expressed entirely in terms of f_X and the matrix $\Phi_{Y,X}(\mathbf{t})$, which, given a vector $\mathbf{t}$, can be directly estimated from data.

In the appendix, we show that

$$\text{Re}\,\Phi_{Y,X}(\mathbf{t}) \times A_{\mathbf{t}} \times (\text{Re}\,\Phi_{Y,X}(\mathbf{t}))^{-1} = F_{X|X^*} \times D_m \times \left(F_{X|X^*}\right)^{-1}, \tag{2.8}$$

where $A_{\mathbf{t}}$ on the left-hand side is identified when $D_{\partial \ln|\phi|}(\mathbf{t})$ and $D_{\partial a}(\mathbf{t})$ are identified, $D_m = Diag\,(m(1), ..., m(J))$, and

$$F_{X|X^*} = \begin{pmatrix} f_{X|X^*}(1|1) & f_{X|X^*}(1|2) & ... & f_{X|X^*}(1|J) \\ f_{X|X^*}(2|1) & f_{X|X^*}(2|2) & ... & f_{X|X^*}(2|J) \\ ... & ... & ... & ... \\ f_{X|X^*}(J|1) & f_{X|X^*}(J|2) & ... & f_{X|X^*}(J|J) \end{pmatrix}.$$

Equation (2.8) implies that $f_{X|X^*}(\cdot|x^*)$ and $m(x^*)$ are eigenfunctions and eigenvalues of an identified $J \times J-$matrix on the left-hand. We may then identify $f_{X|X^*}(\cdot|x^*)$ and $m(x^*)$ under the following assumption:

Assumption 2.4 *(i)* $m(x^*) < \infty$ *and* $m(x^*) \neq 0$ *for all* $x^* \in \mathcal{X}$*; (ii)* $m(x^*)$ *is strictly increasing in* $x^* \in \mathcal{X}$.

Assumption 2.4(i) implies that each possible value of X^* is relevant for Y, and the monotonicity assumption 2.4(ii) allows us to assign each eigenvalue $m(x^*)$ to its corresponding value

x^*. If we only wish to identify the support of the latent factor $m^* = m(X^*)$ and not the regression function $m(\cdot)$ itself, then this monotonicity assumption can be dropped.

Given identification and invertibility of $F_{X|X^*}$, identification of f_{X^*} (the marginal distribution of X^*) immediately follows because f_{X^*} can solved from equation $f_X = \sum_{X^*} f_{X|X^*} f_{X^*}$ given the invertibility of $F_{X|X^*}$.

Assumption 2.4 could be replaced by restrictions on $f_{X|X^*}$ (e.g., by exploiting knowledge about the eigenfunctions rather than eigenvalues to properly assign each $m(x^*)$ to its corresponding value x^*), but assumption 2.4 is more in line with our other assumptions, which assume that we have information about our regression model but know very little about the relationship of the unobserved X^* to the proxy X.

Theorem 2.1 *Suppose that assumptions 2.1, 2.2, 2.3 and 2.4 hold in equation (1.1). Then the density $f_{Y,X}$ uniquely determines $f_{Y|X^*}$, $f_{X|X^*}$, and f_{X^*}.*

Given our model, defined by assumptions 2.1 and 2.2, Theorem 2.1 shows that assumptions 2.3 and 2.4 guarantee that the sample of (Y, X) is informative enough to nonparametrically identify ϕ_η, $m(x^*)$ and f_{X,X^*}, which correspond respectively to the regression error distribution, the regression function, and the joint distribution of the unobserved refressor X^* and of the measurement error. This identification is obtained without additional sample information such as an instrumental variable or a secondary sample. Of course, if we have additional covariates such as instruments or repeated measures, they could exploited along with Theorem 2.1. Our results can also be immediately applied if we observe an additional covariate vector W that appears in the regression function, so $Y = m(X^*, W) + \eta$, since our assumptions and results can all be restated as conditioned upon W.

Now consider some simpler sufficient conditions for assumption 2.3(ii) in Theorem 2.1. Denote $C_{\mathbf{t}}^T$ as the transpose of $C_{\mathbf{t}}$. Let the notation "$\circ$" stand for the Hadamard product, i.e., the element-wise product of two matrices.

Assumption 2.5 *The real-valued vector $\mathbf{t} = (0, t_2, ..., t_J)$ satisfying assumption 2.3(i) also satisfies: $C_{\mathbf{t}} \circ C_{\mathbf{t}}^T + I$ is invertible and all the entries in the first row of the matrix $C_{\mathbf{t}}$ are nonzero.*

Assumption 2.5 implies assumption 2.3(ii), and is in fact stronger than assumption 2.3(ii), since if it holds then we may explicitly solve for $D_{\partial \ln|\phi|}(\mathbf{t})$ and $D_{\partial a}(\mathbf{t})$ in simple closed form. Another alternative to assumption 2.3(ii) is the following:

Assumption 2.6 *(symmetric rank) $a(t) = 0$ for all t and for any real-valued $J \times J-$diagonal matrix $D_1 = Diag(0, d_{1,2}, ..., d_{1,J})$, if $D_1 + C_{\mathbf{t}} \times D_1 \times C_{\mathbf{t}} = 0$ then $D_1 = 0$.*

The condition in assumption 2.6 that $a(t) = 0$ for all t is the same as assuming that the distribution of the error term η is symmetric. We call assumption 2.6 the symmetric rank condition because it implies our previous rank condition when η is symmetrically distributed.

Finally, as noted in the introduction, the assumption that the measurement error is independent of the regression error, assumption 2.2, is stronger than necessary. All independence is used for is to obtain equation (1.3) for some given values of t. More formally, all that is required is that equation (2.4), and hence that equations (2.6) and (2.7) hold for the vector

$\mathbf{t}$ in assumption 2.3. When there are covariates W in the regression model, which we will use in the estimation, the requirement becomes that equation (2.4) hold for the vector $\mathbf{t}$ in assumption 2.3 conditional on W. Therefore, Theorem 2.1 holds replacing assumption 2.2 with the following, strictly weaker assumption.

Assumption 2.7 *For the known $t = 0, t_2, ..., t_J$ that satisfies assumption 2.3, $\phi_{\eta|X^*=x^*}(t) = \phi_{\eta|X^*=1}(t)$ and $\frac{\partial}{\partial t}\phi_{\eta|X^*=x^*}(t) = \frac{\partial}{\partial t}\phi_{\eta|X^*=1}(t)$ for all $x^* \in \mathcal{X}$.*

This condition permits some correlation of the proxy X with the regression error η, and allows some moments of η to correlate with X^*

2.1 THE DICHOTOMOUS CASE

We now show how the assumptions required for Theorem 2.1 can be relaxed and simplified in the special case where X^* is a 0-1 dichotomous variable, i.e., $\mathcal{X} = \{0, 1\}$. Define $m_j = m(j)$ for $j = 0, 1$. Given just assumption 2.1, the relationship between the observed density and the latent ones becomes

$$f_{Y|X}(y|j) = f_{X^*|X}(0|j)\, f_{\eta|X^*}(y - m_0|j) + f_{X^*|X}(1|j)\, f_{\eta|X^*}(y - m_1|j) \quad \text{for } j = 0, 1. \tag{2.9}$$

With assumption 2.2, equation (2.9) simplifies to

$$f_{Y|X}(y|j) = f_{X^*|X}(0|j)\, f_{\eta}(y - m_0) + f_{X^*|X}(1|j)\, f_{\eta}(y - m_1) \quad \text{for } j = 0, 1, \tag{2.10}$$

which says that the observed density $f_{Y|X}(y|j)$ is a mixture of two distributions that only differ in their means. Studies on mixture models focus on parametric or nonparametric restrictions on f_η for a single value of j that suffice to identify all the unknowns in this equation. For example, Bordes, Mottelet and Vandekerkhove (2006) shows that all the unknowns in equation (2.10) are identified for each j when the distribution of η is symmetric. In contrast, errors-in-variables models typically impose restrictions on $f_{X^*|X}$ (or exploit additional information regarding $f_{X^*|X}$ such as instruments or validation data) along with equation (2.9) or (2.10) to obtain identification with few restrictions on the distribution f_η.

Now consider assumptions 2.3 or 2.5 in the dichotomous case. We then have for any real-valued $2 \times 1-$vector $\mathbf{t} = (0, t)$,

$$\Phi_{Y,X}(\mathbf{t}) = \begin{pmatrix} f_X(0) & \phi_{Y|X=0}(t) f_X(0) \\ f_X(1) & \phi_{Y|X=1}(t) f_X(1) \end{pmatrix}$$

$$\mathrm{Re}\{\Phi_{Y,X}(\mathbf{t})\} = \begin{pmatrix} f_X(0) & \mathrm{Re}\,\phi_{Y|X=0}(t) f_X(0) \\ f_X(1) & \mathrm{Re}\,\phi_{Y|X=1}(t) f_X(1) \end{pmatrix}$$

$$\det\left(\mathrm{Re}\{\Phi_{Y,X}(\mathbf{t})\}\right) = f_X(0) f_X(1)\left[\mathrm{Re}\,\phi_{Y|X=1}(t) - \mathrm{Re}\,\phi_{Y|X=0}(t)\right]$$

$$\mathrm{Im}\{\Phi_{Y,X}(\mathbf{t})\} + \Upsilon_X = \begin{pmatrix} f_X(0) & \mathrm{Im}\,\phi_{Y|X=0}(t) f_X(0) \\ f_X(1) & \mathrm{Im}\,\phi_{Y|X=1}(t) f_X(1) \end{pmatrix}$$

$$\det\left(\mathrm{Im}\{\Phi_{Y,X}(\mathbf{t})\} + \Upsilon_X\right) = f_X(0) f_X(1)\left[\mathrm{Im}\,\phi_{Y|X=1}(t) - \mathrm{Im}\,\phi_{Y|X=0}(t)\right]$$

Also,

$$\begin{aligned} C_{\mathbf{t}} &\equiv \left(\mathrm{Re}\{\Phi_{Y,X}(\mathbf{t})\}\right)^{-1} \times \left(\mathrm{Im}\{\Phi_{Y,X}(\mathbf{t})\} + \Upsilon_X\right) \\ &= \frac{1}{\det\left(\mathrm{Re}\{\Phi_{Y,X}(\mathbf{t})\}\right)} \begin{bmatrix} \mathrm{Re}\,\phi_{Y|X=1}(t) f_X(1) & -\mathrm{Re}\,\phi_{Y|X=0}(t) f_X(0) \\ -f_X(1) & f_X(0) \end{bmatrix} \begin{pmatrix} f_X(0) & \mathrm{Im}\,\phi_{Y|X=0}(t) f_X(0) \\ f_X(1) & \mathrm{Im}\,\phi_{Y|X=1}(t) f_X(1) \end{pmatrix} \\ &= \begin{bmatrix} 1 & \frac{f_X(0)f_X(1)\left[\mathrm{Im}\,\phi_{Y|X=0}(t)\,\mathrm{Re}\,\phi_{Y|X=1}(t) - \mathrm{Re}\,\phi_{Y|X=0}(t)\,\mathrm{Im}\,\phi_{Y|X=1}(t)\right]}{\det\left(\mathrm{Re}\{\Phi_{Y,X}(\mathbf{t})\}\right)} \\ 0 & \frac{\det\left(\mathrm{Im}\{\Phi_{Y,X}(\mathbf{t})\} + \Upsilon_X\right)}{\det\left(\mathrm{Re}\{\Phi_{Y,X}(\mathbf{t})\}\right)} \end{bmatrix}, \end{aligned}$$

thus

$$\left(C_{\mathbf{t}} \circ C_{\mathbf{t}}^T\right) + I = \begin{bmatrix} 2 & 0 \\ 0 & \left(\frac{\det\left(\mathrm{Im}\{\Phi_{Y,X}(\mathbf{t})\} + \Upsilon_X\right)}{\det\left(\mathrm{Re}\{\Phi_{Y,X}(\mathbf{t})\}\right)}\right)^2 + 1 \end{bmatrix},$$

which is always invertible. Therefore, for the dichotomous case, assumption 2.3 and assumption 2.5 become the same, and can be expressed as the following rank condition for binary data:

Assumption 2.8 *(binary rank) (i) $f_X(0)f_X(1) > 0$; (ii) there exist a real-valued scalar t such that $\mathrm{Re}\,\phi_{Y|X=0}(t) \neq \mathrm{Re}\,\phi_{Y|X=1}(t)$, $\mathrm{Im}\,\phi_{Y|X=0}(t) \neq \mathrm{Im}\,\phi_{Y|X=1}(t)$, $\mathrm{Im}\,\phi_{Y|X=0}(t)\,\mathrm{Re}\,\phi_{Y|X=1}(t) \neq \mathrm{Re}\,\phi_{Y|X=0}(t)\,\mathrm{Im}\,\phi_{Y|X=1}(t)$.*

It should be generally easy to find a real-valued scalar t that satisfies this binary rank condition.

In the dichotomous case, instead of imposing Assumption 2.4, we may obtain the ordering of m_j from that of observed $\mu_j \equiv E(Y|X=j)$ under the following assumption:

Assumption 2.9 *(i) $\mu_1 > \mu_0$; (ii) $f_{X^*|X}(1|0) + f_{X^*|X}(0|1) < 1$.*

Assumption 2.9(i) is not restrictive because one can always redefine X as $1 - X$ if needed. Assumption 2.9(ii) reveals the ordering of m_1 and m_0, by making it the same as that of μ_1 and μ_0 because

$$1 - f_{X^*|X}(1|0) - f_{X^*|X}(0|1) = \frac{\mu_1 - \mu_0}{m_1 - m_0},$$

so $m_1 \geq \mu_1 > \mu_0 \geq m_0$. Assumption 2.9(ii) says that the sum of misclassification probabilities is less than one, meaning that, on average, the observations X are more accurate predictions of X^* than pure guesses. See Lewbel (2007) for further discussion of this assumption.

The following Corollary is a direct application of Theorem 2.1; hence we omit its proof.

Corollary 2.2 *Suppose that $\mathcal{X} = \{0, 1\}$, equations (1.1) and (2.10), assumptions 2.8 and 2.9 hold. Then the density $f_{Y,X}$ uniquely determines $f_{Y|X^*}$, $f_{X|X^*}$, and f_{X^*}.*

2.1.1 IDENTIFICATION WITHOUT INDEPENDENCE

We now show how to obtain identification in the dichotomous case without the independent regression error assumption 2.2. Given just assumption 2.1, Equation (2.9) implies that the observed density $f_{Y|X}(y|j)$ is a mixture of two conditional densities $f_{\eta|X^*}(y-m_0|j)$ and $f_{\eta|X^*}(y-m_1|j)$.

Instead of assuming the independence between X^* and η, we impose the following weaker assumption:

Assumption 2.10 $E\left(\eta^k|X^*\right) = E\left(\eta^k\right)$ *for* $k = 2, 3$.

Only these two moment restrictions are needed because we only need to solve for two unknowns, m_0 and m_1. Identification could also be obtained using other, similar restrictions such as quantiles or modes. For example, one of the moments in this assumption 2.10 might be replaced with assuming that the density $f_{\eta|X^*=0}$ has zero median. Equation (2.9) then implies that

$$0.5 = \frac{\mu_1 - m_0}{\mu_1 - \mu_0}\int_{-\infty}^{m_0} f_{Y|X=0}(y)dy + \frac{m_0 - \mu_0}{\mu_1 - \mu_0}\int_{-\infty}^{m_0} f_{Y|X=1}(y)dy$$

which may uniquely identify m_0 under some testable assumptions. An advantage of Assumption 2.10 is that we obtain a closed-form solution for m_0 and m_1.

Define $v_j \equiv E\left[\left(Y-\mu_j\right)^2|X=j\right]$, $s_j \equiv E\left[\left(Y-\mu_j\right)^3|X=j\right]$,

$$C_1 \equiv \frac{(v_1+\mu_1^2)-(v_0+\mu_0^2)}{\mu_1-\mu_0},\ C_2 \equiv \frac{1}{2}(\mu_1-\mu_0)^2 + \frac{3}{2}\left(\frac{v_1-v_0}{\mu_1-\mu_0}\right)^2 - \frac{s_1-s_0}{\mu_1-\mu_0}.$$

We leave the detailed proof to the appendix and present the result as follows:

Theorem 2.3 *Suppose that* $\mathcal{X} = \{0,1\}$*, equations (1.1) and (2.9), assumptions 2.9 and 2.10 hold. Then the density* $f_{Y|X}$ *uniquely determines* $f_{Y|X^*}$*,* $f_{X|X^*}$*, and* f_{X^*}*. To be specific, we have*

$$m_0 = \frac{1}{2}C_1 - \sqrt{\frac{1}{2}C_2},\quad m_1 = \frac{1}{2}C_1 + \sqrt{\frac{1}{2}C_2},$$

$$f_{X^*|X}(1|0) = \frac{\mu_0 - \frac{1}{2}C_1}{\sqrt{2C_2}} - \frac{1}{2},\quad f_{X^*|X}(0|1) = \frac{\frac{1}{2}C_1 - \mu_1}{\sqrt{2C_2}} - \frac{1}{2},$$

and

$$f_{Y|X^*=j}(y) = \frac{\mu_1 - m_j}{\mu_1-\mu_0}f_{Y|X=0}(y) + \frac{m_j-\mu_0}{\mu_1-\mu_0}f_{Y|X=1}(y).$$

Note that $f_{X|X^*}$ and f_{X^*} can be immediately recovered from $f_{X^*|X}$ and f_X.

3 SIEVE MAXIMUM LIKELIHOOD ESTIMATION

This section considers the estimation of a nonparametric regression model as follows:

$$Y = m_0(X^*, W) + \eta,$$

where the function $m_0()$ is unknown and W is a vector of error-free covariates and η is independent of (X^*, W). Let $\{Z_t \equiv (Y_t, X_t, W_t)\}_{t=1}^n$ denote a random sample of $Z \equiv (Y, X, W)$. We have shown that $f_{Y|X^*,W}$ and $f_{X^*|X,W}$ are identified from $f_{Y|X,W}$. Let $\alpha_0 \equiv (f_{01}, f_{02}, f_{03})^T \equiv \left(f_\eta, f_{X^*|X,W}, m_0\right)^T$ be the true parameters of interest. Then the observed likelihood of Y given (X, W) (or the likelihood for α_0) is

$$\prod_{t=1}^n f_{Y|X,W}\left(Y_t|X_t, W_t\right) = \prod_{t=1}^n \left\{ \sum_{x^* \in \mathcal{X}} f_\eta\left(Y_t - m_0(x^*, W_t)\right) f_{X^*|X,W}\left(x^*|X_t, W_t\right) \right\}.$$

Before we present a sieve ML estimator $\widehat{\alpha}$ for α_0, we need to impose some mild smoothness restrictions on the unknown functions $\alpha_0 \equiv \left(f_\eta, f_{X^*|X,W}, m_0\right)^T$. The sieve method allows for unknown functions belonging to many different function spaces such as Sobolev space, Besov space and others; see e.g., Shen and Wong (1994), Wong and Shen (1995), Shen (1997) and Van de Geer (1993, 2000). But, for the sake of concreteness and simplicity, we consider the widely used Hölder space of functions. Let $\xi = (\xi_1, ..., \xi_d)^T \in \mathbb{R}^d$, $\mathbf{a} = (a_1, ..., a_d)^T$ be a vector of non-negative integers, and

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

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

$$\|h\|_{\Lambda^\gamma} = \max_{|\mathbf{a}| \leq \underline{\gamma}} \sup_\xi \left|\nabla^{\mathbf{a}} h(\xi)\right| + \max_{|\mathbf{a}| = \underline{\gamma}} \sup_{\xi \neq \xi'} \frac{\left|\nabla^{\mathbf{a}} h(\xi) - \nabla^{\mathbf{a}} h(\xi')\right|}{\left(\|\xi - \xi'\|_E\right)^{\gamma - \underline{\gamma}}} < \infty.$$

Denote $\Lambda_c^\gamma(\mathcal{V}) \equiv \{h \in \Lambda^\gamma(\mathcal{V}) : \|h\|_{\Lambda^\gamma} \leq c < \infty\}$ as a Hölder ball. Let $\eta \in \mathbb{R}$, $W \in \mathcal{W}$ with $\mathcal{W}$ a compact convex subset in $\mathbb{R}^{d_w}$. Also denote

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

$$\mathcal{F}_2 = \left\{ \sqrt{f_2\left(x^*|x, \cdot\right)} \in \Lambda_c^{\gamma_2}\left(\mathcal{W}\right) :\ f_2\left(\cdot|\cdot, \cdot\right) > 0,\ \int_{\mathcal{X}} f_2\left(x^*|x, w\right) dx^* = 1 \text{ for all } x \in \mathcal{X},\, w \in \mathcal{W} \right\},$$

and

$$\mathcal{F}_3 = \{f_3\left(x^*, \cdot\right) \in \Lambda_c^{\gamma_3}\left(\mathcal{W}\right) :\ f_3\left(i, w\right) > f_3\left(j, w\right) \text{ for all } i > j,\, i, j \in \mathcal{X},\, w \in \mathcal{W}\}$$

We impose the following smoothness restrictions on the densities:

Assumption 3.1 *(i) all the assumptions in Theorem 2.1 hold; (ii) $f_\eta(\cdot) \in \mathcal{F}_1$ with $\gamma_1 > 1/2$; (iii) $f_{X^*|X,W}(x^*|x, \cdot) \in \mathcal{F}_2$ with $\gamma_2 > d_w/2$ for all $x^*, x \in \mathcal{X} \equiv \{1, ..., J\}$; (iv) $m_0(x^*, \cdot) \in \mathcal{F}_3$ with $\gamma_3 > d_w/2$ for all $x^* \in \mathcal{X}$.*

Denote $\mathcal{A} = \mathcal{F}_1 \times \mathcal{F}_2 \times \mathcal{F}_3$ and $\alpha = (f_1, f_2, f_3)^T$. Let $E[\cdot]$ denote the expectation with respect to the underlying true data generating process for Z_t. Then $\alpha_0 \equiv (f_{01}, f_{02}, f_{03})^T = \arg\max_{\alpha\in\mathcal{A}} E[\ell(Z_t;\alpha)]$, where

$$\ell(Z_t;\alpha) \equiv \ln\left\{\sum_{x^*\in\mathcal{X}} f_1\left(Y_t - f_3(x^*, W_t)\right) f_2\left(x^*|X_t, W_t\right)\right\}. \tag{3.1}$$

Let $\mathcal{A}_n = \mathcal{F}_1^n \times \mathcal{F}_2^n \times \mathcal{F}_3^n$ 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 defined as:

$$\widehat{\alpha}_n = \arg\max_{\alpha\in\mathcal{A}_n} \sum_{t=1}^{n} \ell(Z_t;\alpha). \tag{3.2}$$

We could apply infinite-dimensional approximating spaces as sieves $\mathcal{F}_j^n$ for $\mathcal{F}_j, j = 1, 2, 3$. However, in applications, we shall use finite-dimensional sieve spaces since they are easier to implement. For $j = 1, 2, 3$, let $p_j^{k_{j,n}}(\cdot)$ be a $k_{j,n} \times 1-$vector of known basis functions, such as power series, splines, Fourier series, etc. Then we denote the sieve space for $\mathcal{F}_j, j = 1, 2, 3$ as follows:

$$\mathcal{F}_1^n = \left\{\sqrt{f_1(\cdot)} = p_1^{k_{1,n}}(\cdot)^T\beta_1 \in \mathcal{F}_1\right\},$$

$$\mathcal{F}_2^n = \left\{\sqrt{f_2\left(x^*|x,\cdot\right)} = \sum_{k=1}^{J}\sum_{j=1}^{J} I\left(x^* = k\right) I\left(x = j\right) p_2^{k_{2,n}}(\cdot)^T\beta_{2,kj} \in \mathcal{F}_2\right\},$$

$$\mathcal{F}_3^n = \left\{f_3\left(x^*,\cdot\right) = \sum_{k=1}^{J} I\left(x^* = k\right) p_3^{k_{2,n}}(\cdot)^T\beta_{3,k} \in \mathcal{F}_3\right\}.$$

We note that the method of sieve MLE is very flexible and we can easily impose prior information on the parameter space ($\mathcal{A}$) and the sieve space ($\mathcal{A}_n$). For example, if the functional form of the true regression function $m_0(x^*, w)$ is known upto some finite-dimensional parameters $\beta_0 \in B$, where B is a compact subset of $\mathbb{R}^{d_\beta}$, then we can take $\mathcal{A} = \mathcal{F}_1 \times \mathcal{F}_2 \times \mathcal{F}_B$ and $\mathcal{A}_n = \mathcal{F}_1^n \times \mathcal{F}_2^n \times \mathcal{F}_B$ with $\mathcal{F}_B = \{f_3\left(x^*, w\right) = m_0(x^*, w;\beta) : \beta \in B\}$. The sieve MLE becomes

$$\widehat{\alpha}_n = \arg\max_{\alpha\in\mathcal{A}_n} \sum_{t=1}^{n} \ell(Z_t;\alpha), \quad \text{with } \ell(Z_t;\alpha) = \ln\left\{\sum_{x^*\in\mathcal{X}} f_1\left(Y_t - m_0(x^*, W_t;\beta)\right) f_2\left(x^*|X_t, W_t\right)\right\}. \tag{3.3}$$

3.1 Consistency

The consistency of the sieve MLE $\widehat{\alpha}_n$ can be established by applying either Geman and Hwang (1982) or lemma A.1 of Newey and Powell (2003). First we define a norm on $\mathcal{A}$ as follows:

$$\|\alpha\|_s = \sup_{\eta}\left|h(\eta)\left(1+\eta^2\right)^{-\zeta/2}\right| + \sup_{x^*,x,w}\left|f_2\left(x^*|x,w\right)\right| + \sup_{x^*,w}\left|f_3\left(x^*,w\right)\right| \quad \text{for some } \zeta > 0.$$

We assume

Assumption 3.2 *(i)* $-\infty < E[\ell(Z_t;\alpha_0)] < \infty$, $E[\ell(Z_t;\alpha)]$ *is upper semicontinuous on* $\mathcal{A}$ *under the metric* $\|\cdot\|_s$*; (ii) there are a finite* $\kappa > 0$ *and a random variable* $U(Z_t)$ *with* $E\{U(Z_t)\} < \infty$ *such that* $\sup_{\alpha\in\mathcal{A}_n:\|\alpha-\alpha_0\|_s\le\delta} |\ell(Z_t;\alpha) - \ell(Z_t;\alpha_0)| \le \delta^\kappa U(Z_t)$.

Assumption 3.3 *(i)* $p_1^{k_{1,n}}(\cdot)$ *is a* $k_{1,n}\times 1-$*vector of spline wavelet basis functions on* $\mathbb{R}$*, and for* $j = 2, 3$, $p_j^{k_{j,n}}(\cdot)$ *is a* $k_{j,n}\times 1-$*vector of tensor product of spline basis functions on* $\mathcal{W}$*; (ii)* $k_n \equiv \max\{k_{1,n}, k_{2,n}, k_{3,n}\} \to \infty$ *and* $k_n/n \to 0$.

The following consistency lemma is a direct application of lemma A.1 of Newey and Powell (2003) or theorem 3.1 (or remark 3.1(4), remark 3.3) of Chen (2006); hence we omit its proof.

Lemma 3.1 *Let* $\widehat{\alpha}_n$ *be the sieve MLE. Under assumptions 3.1-3.3, we have* $\|\widehat{\alpha}_n - \alpha_0\|_s = o_p(1)$.

3.2 Convergence rate under the Fisher metric

Given Lemma 3.1, we can now restrict our attention to a shrinking $||\cdot||_s-$neighborhood around α_0. Let $\mathcal{A}_{0s} \equiv \{\alpha \in \mathcal{A} : ||\alpha - \alpha_0||_s = o(1), ||\alpha||_s \le c_0 < c\}$ and $\mathcal{A}_{0sn} \equiv \{\alpha \in \mathcal{A}_n : ||\alpha - \alpha_0||_s = o(1), ||\alpha||_s \le c_0 < c\}$. Then, for the purpose of establishing a convergence rate under a pseudo metric that is weaker than $||\cdot||_s$, we can treat $\mathcal{A}_{0s}$ as the new parameter space and $\mathcal{A}_{0sn}$ as its sieve space, and assume that both $\mathcal{A}_{0s}$ and $\mathcal{A}_{0sn}$ are convex parameter spaces. For any α_1, $\alpha_2 \in \mathcal{A}_{0s}$, we consider a continuous path $\{\alpha(\tau) : \tau \in [0,1]\}$ in $\mathcal{A}_{0s}$ such that $\alpha(0) = \alpha_1$ and $\alpha(1) = \alpha_2$. For simplicity we assume that for any α, $\alpha + v \in \mathcal{A}_{0s}$, $\{\alpha + \tau v : \tau \in [0,1]\}$ is a continuous path in $\mathcal{A}_{0s}$, and that $\ell(Z_t;\alpha+\tau v)$ is twice continuously differentiable at $\tau = 0$ for almost all Z_t and any direction $v \in \mathcal{A}_{0s}$. Define the pathwise first derivative as

$$\frac{d\ell(Z_t;\alpha)}{d\alpha}[v] \equiv \frac{d\ell(Z_t;\alpha+\tau v)}{d\tau}|_{\tau=0} \text{ a.s. } Z_t,$$

and the pathwise second derivative as

$$\frac{d^2\ell(Z_t;\alpha)}{d\alpha d\alpha^T}[v,v] \equiv \frac{d^2\ell(Z_t;\alpha+\tau v)}{d\tau^2}|_{\tau=0} \quad \text{a.s. } Z_t.$$

Define the Fisher metric $\|\cdot\|$ on $\mathcal{A}_{0s}$ as follows: for any α_1, $\alpha_2 \in \mathcal{A}_{0s}$,

$$\|\alpha_1 - \alpha_2\|^2 \equiv E\left\{\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2]\right)^2\right\}.$$

We show that $\widehat{\alpha}_n$ converges to α_0 at a rate faster than $n^{-1/4}$ under the Fisher metric $\|\cdot\|$ with the following assumptions:

Assumption 3.4 *(i)* $\zeta > \gamma_1$*; (ii)* $\gamma \equiv \min\{\gamma_1, \gamma_2/d_w, \gamma_3/d_w\} > 1/2$.

Assumption 3.5 *(i) $\mathcal{A}_{0s}$ is convex at α_0; (ii) $\ell(Z_t;\alpha)$ is twice continuously pathwise differentiable with respect to $\alpha \in \mathcal{A}_{0s}$.*

Assumption 3.6 $\sup_{\widetilde{\alpha}\in\mathcal{A}_{0s}} \sup_{\alpha\in\mathcal{A}_{0sn}} \left| \frac{d\ell(Z_t;\widetilde{\alpha})}{d\alpha}\left[\frac{\alpha-\alpha_0}{\|\alpha-\alpha_0\|_s}\right]\right| \leq U(Z_t)$ *for a random variable $U(Z_t)$ with $E\{[U(Z_t)]^2\} < \infty$.*

Assumption 3.7 *(i)* $\sup_{v\in\mathcal{A}_{0s}:||v||_s=1} E\left\{\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[v]\right)^2\right\} \leq c < \infty$*; (ii) uniformly over $\widetilde{\alpha} \in \mathcal{A}_{0s}$ and $\alpha \in \mathcal{A}_{0sn}$, we have*

$$-E\left(\frac{d^2\ell(Z_t;\widetilde{\alpha})}{d\alpha d\alpha^T}[\alpha-\alpha_0,\alpha-\alpha_0]\right) = \|\alpha-\alpha_0\|^2 \times \{1+o(1)\}.$$

Assumption 3.4 guarantees that the sieve approximation error under the strong norm $||\cdot||_s$ goes to zero at the rate of $(k_n)^{-\gamma}$. Assumption 3.5 makes sure that the twice pathwise derivatives are well defined with respect to $\alpha \in \mathcal{A}_{0s}$, hence the pseudo metric $\|\alpha-\alpha_0\|$ is well defined on $\mathcal{A}_{0s}$. Assumption 3.6 impose an envelope condition. Assumption 3.7(i) implies that $\|\alpha-\alpha_0\| \leq \sqrt{c}\,\|\alpha-\alpha_0\|_s$ for all $\alpha \in \mathcal{A}_{0s}$. Assumption 3.7(ii) implies that there are positive finite constants c_1 and c_2 such that for all $\alpha \in \mathcal{A}_{0sn}$, $c_1\|\alpha-\alpha_0\|^2 \leq E[\ell(Z_t;\alpha_0)-\ell(Z_t;\alpha)] \leq c_2\|\alpha-\alpha_0\|^2$, that is, $\|\alpha-\alpha_0\|^2$ is equivalent to the Kullback-Leibler discrepancy on the local sieve space $\mathcal{A}_{0sn}$. The following convergence rate theorem is a direct application of theorem 3.2 of Shen and Wong (2004) or theorem 3.2 of Chen (2006) to the local parameter space $\mathcal{A}_{0s}$ and the local sieve space $\mathcal{A}_{0sn}$; hence we omit its proof.

Theorem 3.2 *Under assumptions 3.1-3.7, we have*

$$\|\widehat{\alpha}_n-\alpha_0\| = O_P\left(\max\left\{k_n^{-\gamma},\sqrt{\frac{k_n}{n}}\right\}\right) = O_P\left(n^{\frac{-\gamma}{2\gamma+1}}\right) \text{ if } k_n = O\left(n^{\frac{1}{2\gamma+1}}\right).$$

3.3 Asymptotic normality and semiparametric efficiency

Let $\overline{\mathbf{V}}$ denote the closure of the linear span of $\mathcal{A}_{0s}-\{\alpha_0\}$ under the Fisher metric $\|\cdot\|$. Then $\left(\overline{\mathbf{V}},\|\cdot\|\right)$ is a Hilbert space with the inner product defined as

$$\langle v_1,v_2\rangle \equiv E\left\{\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[v_1]\right)\left(\frac{d\ell(Z_t;\alpha_0)}{d\alpha}[v_2]\right)\right\}.$$

We are interested in estimation of a functional $\rho(\alpha_0)$, where $\rho:\mathcal{A}\to\mathbb{R}$. It is known that the asymptotic properties of $\rho(\hat{\alpha}_n)$ depend on the smoothness of the functional ρ and the rate of convergence of the sieve MLE $\hat{\alpha}_n$. For any $v \in \mathbf{V}$, we denote

$$\frac{d\rho(\alpha_0)}{d\alpha}[v] \equiv \lim_{\tau\to 0}[(\rho(\alpha_0+\tau v)-\rho(\alpha_0))/\tau]$$

whenever the right hand-side limit is well defined.

We impose the following additional conditions for asymptotic normality of plug-in sieve MLE $\rho(\hat{\alpha}_n)$:

Assumption 3.8 *(i) for any $v \in \mathbf{V}$, $\rho(\alpha_0 + \tau v)$ is continuously differentiable in $\tau \in [0,1]$ near $\tau = 0$, and*

$$\|\frac{d\rho(\alpha_0)}{d\alpha}\| \equiv \sup_{v \in \mathbf{V}: ||v|| > 0} \frac{\left|\frac{d\rho(\alpha_0)}{d\alpha}[v]\right|}{||v||} < \infty;$$

(ii) there exist constants $c > 0, \omega > 0$, and a small $\varepsilon > 0$ such that for any $v \in \mathbf{V}$ with $||v|| \leq \varepsilon$, we have

$$\left|\rho(\alpha_0 + v) - \rho(\alpha_0) - \frac{d\rho(\alpha_0)}{d\alpha}[v]\right| \leq c||v||^{\omega}.$$

Under Assumption 3.8 (i), by the Riesz representation theorem, there exists $v^* \in \overline{\mathbf{V}}$ such that

$$\langle v^*, v \rangle = \frac{d\rho(\alpha_0)}{d\alpha}[v] \quad \text{for all } v \in \mathbf{V} \tag{3.4}$$

and

$$||v^*||^2 \equiv \|\frac{d\rho(\alpha_0)}{d\alpha}\|^2 \equiv \sup_{v \in \mathbf{V}: ||v|| > 0} \frac{\left|\frac{d\rho(\alpha_0)}{d\alpha}[v]\right|^2}{||v||^2} < \infty. \tag{3.5}$$

Under Theorem 3.2, we have $||\widehat{\alpha}_n - \alpha_0|| = O_P(\delta_n)$ with $\delta_n = n^{\frac{-\gamma}{2\gamma+1}}$. In the following we denote $\mathcal{N}_0 = \{\alpha \in \mathcal{A}_{0s} : \|\alpha - \alpha_0\| = O(\delta_n)\}$ and $\mathcal{N}_{0n} = \{\alpha \in \mathcal{A}_{0sn} : \|\alpha - \alpha_0\| = O(\delta_n)\}$.

Assumption 3.9 *(i) $(\delta_n)^{\omega} = o(n^{-1/2})$; (ii) there is a $v_n^* \in \mathcal{A}_n - \{\alpha_0\}$ such that $||v_n^* - v^*|| = o(1)$ and $\delta_n \times \|v_n^* - v^*\| = o(n^{-1/2})$.*

Assumption 3.10 *there is a random variable $U(Z_t)$ with $E\{[U(Z_t)]^2\} < \infty$ and a non-negative measurable function η with $\lim_{\delta \to 0} \eta(\delta) = 0$ such that for all $\alpha \in \mathcal{N}_{0n}$,*

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

Assumption 3.11 *Uniformly over $\overline{\alpha} \in \mathcal{N}_0$ and $\alpha \in \mathcal{N}_{0n}$,*

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

Assumption 3.8(i) is critical for obtaining the $\sqrt{n}$ convergence of plug-in sieve MLE $\rho(\hat{\alpha}_n)$ to $\rho(\alpha_0)$ and its asymptotic normality. If Assumption 3.8(i) is not satisfied, then the plug-in sieve MLE $\rho(\hat{\alpha}_n)$ is still consistent for $\rho(\alpha_0)$, but the best achievable convergence rate is slower than the $\sqrt{n}-$rate. Assumption 3.9 implies that the asymptotic bias of the Riesz representer is negligible. Assumptions 3.10 and 3.11 control the remainder term.

Applying theorems 1 and 4 of Shen (1997), we immediately obtain

Theorem 3.3 *Suppose that assumptions 3.1-3.11 hold. Then the plug-in sieve MLE $\rho(\hat{\alpha}_n)$ is semiparametrically efficient, and $\sqrt{n}\left(\rho(\hat{\alpha}_n) - \rho(\alpha_0)\right) \xrightarrow{d} N\left(0, ||v^*||^2\right)$.*

Following Ai and Chen (2003), the asymptotic efficient variance, $||v^*||^2$, of the plug-in sieve MLE $\rho(\hat{\alpha}_n)$ can be consistently estimated by $\widehat{\sigma}_n^2$:

$$\widehat{\sigma}_n^2 = \max_{v \in \mathcal{A}_n} \frac{\left| \frac{d\rho(\widehat{\alpha}_n)}{d\alpha}[v] \right|^2}{\frac{1}{n}\sum_{t=1}^n \left(\frac{d\ell(Z_t;\widehat{\alpha}_n)}{d\alpha}[v] \right)^2}.$$

Instead of estimating this asymptotic variance, one could also construct confidence intervals by applying the likelihood ratio inference as in Murphy and Van der Vaart (1996, 2000).

4 Simulation

This section presents two small simulation studies: the first one corresponds to the identification strategy, and the second one checks the performance of sieve MLE.

4.1 Moment-based estimation

This subsection applies the identification procedure to a simple nonlinear regression model with simulated data. We consider the following regression model

$$y = 1 + 0.25\,(x^*)^2 + 0.1\,(x^*)^3 + \eta,$$

where $\eta \sim N(0,1)$ is independent of x^*. The marginal distribution $\Pr(x^*)$ is as follows:

$$\Pr(x^*) = 0.2 \times [1(x^* = 1) + 1(x^* = 4)] + 0.3 \times [1(x^* = 2) + 1(x^* = 3)]$$

and the misclassification probability matrix $F_{x|x^*}$ are in Tables 1-2. We consider two examples of the misclassification probability matrix. Example 1 considers a strictly diagonally dominant matrix $F_{x|x^*}$ as follows:

$$F_{x|x^*} = \begin{pmatrix} 0.6 & 0.2 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.7 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.7 \end{pmatrix}.$$

Example 2 has a misclassification probability matrix

$$F_{x|x^*} = 0.7F_u + 0.3I,$$

where I is an identify matrix and $F_u = \left[\frac{u_{ij}}{\sum_k u_{kj}} \right]_{ij}$ with u_{ij} independently drawn from a uniform distribution on $[0,1]$.

In each repetition, we directly follow the identification procedure shown in the proof of theorem 2.1. The matrix $\Phi_{Y,X}$ is estimated by replacing the function $\phi_{Y,X=x}(t)$ with its

corresponding empirical counterpart as follows:

$$\widehat{\phi}_{Y,X=x}(t) = \sum_{j=1}^{n} \exp(ity_j) \times 1(x_j = x).$$

Since it is directly testable, assumption 2.3 is verified with t_j in the vector $\mathbf{t} = (0, t_2, t_3, t_4)$ independently drawn from a uniform distribution on $[-1, 1]$ until a desirable $\mathbf{t}$ is found. The sample size is 5000 and the repetition times is 1000. The simulation results in Tables 1-2 include the estimates of regression function $m(x^*)$, the marginal distribution $\Pr(x^*)$, and the estimated misclassification probability matrix $F_{x|x^*}$, together with standard errors of each element. As shown in Tables 1-2, the estimator following the identification procedure performs well with the simulated data.

4.2 Sieve MLE

This subsection applies the sieve ML procedure to a semiparametric model as follows:

$$Y = \beta_1 W + \beta_2 (1 - X^*) W^2 + \beta_3 + \eta,$$

where η is independent of $X^* \in \{0, 1\}$ and W. The unknowns include the parameter of interest $\beta = (\beta_1, \beta_2, \beta_3)$ and the nuisance functions f_η and $f_{X^*|X,W}$.

We simulate the model from $\eta \sim N(0, 1)$ and $X^* \in \{0, 1\}$ according to the marginal distribution $f_{X^*}(x^*) = 0.4 \times 1(x^* = 0) + 0.6 \times 1(x^* = 1)$. We generate the covariate W as $W = (1 - 0.5X^*) \times \nu$, where $\nu \sim N(0, 1)$ is independent of X^*. The observed mismeasured X is generated as follows:

$$X = \begin{cases} 0 & \Phi(\nu) \leq p(X^*) \\ 1 & otherwise \end{cases},$$

where $p(0) = 0.5$ and $p(1) = 0.3$.

The Monte Carlo simulation consists of 400 repetitions. In each repetition, we randomly draw 3000 observations of (Y, X, W), and then apply three ML estimators to compute the parameter of interest β. All three estimators assume that the true density f_η of the regression error is unknown. The first estimator uses the contaminated sample $\{Y_i, X_i, W_i\}_{i=1}^n$ as if it were accurate; this estimator is inconsistent and its bias should dominate the squared root of mean square error (root MSE). The second estimator is the sieve MLE using uncontaminated data $\{Y_i, X_i^*, W_i\}_{i=1}^n$; this estimator is consistent and most efficient. However, we call it "infeasible MLE" since X_i^* is not observed in practice. The third estimator is the sieve MLE (3.3) presented in Section 3, using the sample $\{Y_i, X_i, W_i\}_{i=1}^n$ and allowing for arbitrary measurement error by assuming $f_{X|X^*,W}$ is unknown. In this simulation study, all three estimators are computed by approximating the unknown $\sqrt{f_\eta}$ using the same Hermite polynomial sieve; for the third estimator (the sieve MLE) we also approximate $\sqrt{f_{X|X^*,W}}$ by another Hermite polynomial sieve. The Monte Carlo results in Table 3 show that the sieve MLE has a much smaller bias than the first estimator ignoring measurement error. Since the sieve MLE has to estimate the additional unknown function $f_{X|X^*,W}$, its $\widehat{\beta}_j$, $j = 1, 2, 3$ estimate may have larger standard error compared to the other two estimators. In summary,

our sieve MLE performs well in this Monte Carlo simulation.

5 Discussion

We have provided nonparametric identification and estimation of a regression model in the presence of a mismeasured discrete regressor, without the use of additional sample information such as instruments, repeated measurements or validation data, and without parameterizing the distributions of the measurement error or of the regression error.

Identification mainly comes from the monotonicity of the regression function, the limited support of the mismeasured regressor, sufficient variation in the dependent variable, and from some independence related assumptions regarding the regression model error. It may be possible to extend these results to continuously distributed nonclassically mismeasured regressors, by replacing many of our matrix related assumptions and calculations with corresponding linear operators.

APPENDIX. MATHEMATICAL PROOFS

Proof. (Theorem 2.1) Notice that $\frac{\partial}{\partial t}\left|\phi_\eta(0)\right| = 0$ and $\frac{\partial}{\partial t}a(0) = 0$. we define

$$\frac{\partial}{\partial \mathbf{t}}\Phi_{Y,X}(\mathbf{t}) = \begin{pmatrix} iE[Y|X=1]f_X(1) & \frac{\partial}{\partial t}\phi_{Y,X=1}(t_2) & ... & \frac{\partial}{\partial t}\phi_{Y,X=1}(t_J) \\ iE[Y|X=2]f_X(2) & \frac{\partial}{\partial t}\phi_{Y,X=2}(t_2) & ... & \frac{\partial}{\partial t}\phi_{Y,X=2}(t_J) \\ ... & ... & ... & ... \\ iE[Y|X=J]f_X(J) & \frac{\partial}{\partial t}\phi_{Y,X=J}(t_2) & ... & \frac{\partial}{\partial t}\phi_{Y,X=J}(t_J) \end{pmatrix}.$$

By taking the derivative with respect to scalar t, we have from equation (2.3)

$$\begin{aligned} \frac{\partial}{\partial t}\phi_{Y,X=x}(t) &= \left(\frac{\partial}{\partial t}\left|\phi_\eta(t)\right|\right)\sum_{x^*}\exp\left(itm(x^*)+ia(t)\right)f_{X,X^*}(x,x^*) \\ &\quad +i\left(\frac{\partial}{\partial t}a(t)\right)\left|\phi_\eta(t)\right|\sum_{x^*}\exp\left(itm(x^*)+ia(t)\right)f_{X,X^*}(x,x^*) \\ &\quad +i\left|\phi_\eta(t)\right|\sum_{x^*}\exp\left(itm(x^*)+ia(t)\right)m(x^*)f_{X,X^*}(x,x^*). \end{aligned} \tag{A.1}$$

Equation (A.1) is equivalent to

$$\begin{aligned} \frac{\partial}{\partial \mathbf{t}}\Phi_{Y,X}(\mathbf{t}) &= F_{X,X^*}\times\Phi_{m,a}(\mathbf{t})\times D_{\partial|\phi|}(\mathbf{t}) \\ &\quad +i\times F_{X,X^*}\times\Phi_{m,a}(\mathbf{t})\times D_{|\phi|}(\mathbf{t})\times D_{\partial a}(\mathbf{t}) + i\times F_{X,X^*}\times D_m\times\Phi_{m,a}(\mathbf{t})\times D_{|\phi|}(\mathbf{t}), \end{aligned} \tag{A.2}$$

where

$$D_{\partial|\phi|}(\mathbf{t}) = \begin{pmatrix} 0 & 0 & ... & 0 \\ 0 & \frac{\partial}{\partial t}\left|\phi_\eta(t_2)\right| & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \frac{\partial}{\partial t}\left|\phi_\eta(t_J)\right| \end{pmatrix},$$

$$D_{\partial a}(\mathbf{t}) = \begin{pmatrix} 0 & 0 & ... & 0 \\ 0 & \frac{\partial}{\partial t}a(t_2) & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \frac{\partial}{\partial t}a(t_J) \end{pmatrix}, \quad D_m = \begin{pmatrix} m_1 & 0 & ... & 0 \\ 0 & m_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & m_J \end{pmatrix}.$$

Since by definition, $D_{\partial|\phi|}(\mathbf{t})$ and $D_{\partial a}(\mathbf{t})$ are real-valued, we also have from equation (A.2)

$$\begin{aligned} \mathrm{Re}\{\frac{\partial}{\partial \mathbf{t}}\Phi_{Y,X}(\mathbf{t})\} &= F_{X,X^*} \times \mathrm{Re}\{\Phi_{m,a}(\mathbf{t})\} \times D_{\partial|\phi|}(\mathbf{t}) \\ &\quad -F_{X,X^*} \times \mathrm{Im}\{\Phi_{m,a}(\mathbf{t})\} \times D_{|\phi|}(\mathbf{t}) \times D_{\partial a}(\mathbf{t}) \\ &\quad -F_{X,X^*} \times D_m \times \mathrm{Im}\{\Phi_{m,a}(\mathbf{t})\} \times D_{|\phi|}(\mathbf{t}). \end{aligned} \tag{A.3}$$

In order to replace the singular matrix $\mathrm{Im}\{\Phi_{m,a}(\mathbf{t})\}$ with the invertible $(\mathrm{Im}\{\Phi_{m,a}(\mathbf{t})\} + \Upsilon)$, we define

$$\Upsilon_{E[Y|X]} = \begin{pmatrix} E\left[Y|X=1\right] f_X(1) & 0 & ... & 0 \\ E\left[Y|X=2\right] f_X(2) & 0 & ... & 0 \\ ... & ... & ... & ... \\ E\left[Y|X=J\right] f_X(J) & 0 & ... & 0 \end{pmatrix} = F_{X,X^*} \times D_m \times \Upsilon.$$

We then have

$$\begin{aligned} \left(\mathrm{Re}\{\frac{\partial}{\partial \mathbf{t}}\Phi_{Y,X}(\mathbf{t})\} - \Upsilon_{E[Y|X]}\right) &= F_{X,X^*} \times \mathrm{Re}\{\Phi_{m,a}(\mathbf{t})\} \times D_{\partial|\phi|}(\mathbf{t}) \\ &\quad -F_{X,X^*} \times (\mathrm{Im}\{\Phi_{m,a}(\mathbf{t})\} + \Upsilon) \times D_{|\phi|}(\mathbf{t}) \times D_{\partial a}(\mathbf{t}) \\ &\quad -F_{X,X^*} \times D_m \times (\mathrm{Im}\{\Phi_{m,a}(\mathbf{t})\} + \Upsilon) \times D_{|\phi|}(\mathbf{t}), \end{aligned} \tag{A.4}$$

where $\Upsilon \times D_{|\phi|}(\mathbf{t}) \times D_{\partial a}(\mathbf{t}) = 0$ and $\Upsilon = \Upsilon \times D_{|\phi|}(\mathbf{t})$. Similarly, we have

$$\begin{aligned} \mathrm{Im}\{\frac{\partial}{\partial \mathbf{t}}\Phi_{Y,X}(\mathbf{t})\} &= F_{X,X^*} \times \mathrm{Im}\{\Phi_{m,a}(\mathbf{t})\} \times D_{\partial|\phi|}(\mathbf{t}) \\ &\quad +F_{X,X^*} \times \mathrm{Re}\{\Phi_{m,a}(\mathbf{t})\} \times D_{|\phi|}(\mathbf{t}) \times D_{\partial a}(\mathbf{t}) + F_{X,X^*} \times D_m \times \mathrm{Re}\{\Phi_{m,a}(\mathbf{t})\} \times D_{|\phi|}(\mathbf{t}) \\ &= F_{X,X^*} \times (\mathrm{Im}\{\Phi_{m,a}(\mathbf{t})\} + \Upsilon) \times D_{\partial|\phi|}(\mathbf{t}) \\ &\quad +F_{X,X^*} \times \mathrm{Re}\{\Phi_{m,a}(\mathbf{t})\} \times D_{|\phi|}(\mathbf{t}) \times D_{\partial a}(\mathbf{t}) + F_{X,X^*} \times D_m \times \mathrm{Re}\{\Phi_{m,a}(\mathbf{t})\} \times D_{|\phi|}(\mathbf{t}), \end{aligned}$$

where $\Upsilon \times D_{\partial|\phi|}(\mathbf{t}) = 0$. Define $\Phi_{Y|X^*}(\mathbf{t}) = \Phi_{m,a}(\mathbf{t}) \times D_{|\phi|}(\mathbf{t})$. We then have

$$\mathrm{Re}\{\Phi_{Y|X^*}(\mathbf{t})\} = \mathrm{Re}\{\Phi_{m,a}(\mathbf{t})\} \times D_{|\phi|}(\mathbf{t}), \quad \left(\mathrm{Im}\{\Phi_{Y|X^*}(\mathbf{t})\} + \Upsilon\right) = (\mathrm{Im}\{\Phi_{m,a}(\mathbf{t})\} + \Upsilon) \times D_{|\phi|}(\mathbf{t}).$$

In summary, we have

$$\begin{aligned}
\mathrm{Re}\{\Phi_{Y,X}(\mathbf{t})\} &= F_{X,X^*} \times \mathrm{Re}\{\Phi_{Y|X^*}(\mathbf{t})\}, && \text{(A.5)}\\
(\mathrm{Im}\{\Phi_{Y,X}(\mathbf{t})\} + \Upsilon_X) &= F_{X,X^*} \times \left(\mathrm{Im}\{\Phi_{Y|X^*}(\mathbf{t})\} + \Upsilon\right), && \text{(A.6)}\\
\left(\mathrm{Re}\,\frac{\partial}{\partial \mathbf{t}}\Phi_{Y,X}(\mathbf{t}) - \Upsilon_{E[Y|X]}\right) &= F_{X,X^*} \times \mathrm{Re}\,\Phi_{m,a}(\mathbf{t}) \times D_{\partial|\phi|}(\mathbf{t}) && \text{(A.7)}\\
&\quad -F_{X,X^*} \times \left(\mathrm{Im}\,\Phi_{Y|X^*}(\mathbf{t}) + \Upsilon\right) \times D_{\partial a}(\mathbf{t})\\
&\quad -F_{X,X^*} \times D_m \times \left(\mathrm{Im}\,\Phi_{Y|X^*}(\mathbf{t}) + \Upsilon\right),\\
\mathrm{Im}\,\frac{\partial}{\partial \mathbf{t}}\Phi_{Y,X}(\mathbf{t}) &= F_{X,X^*} \times (\mathrm{Im}\,\Phi_{m,a}(\mathbf{t}) + \Upsilon) \times D_{\partial|\phi|}(\mathbf{t}) && \text{(A.8)}\\
&\quad +F_{X,X^*} \times \mathrm{Re}\,\Phi_{Y|X^*}(\mathbf{t}) \times D_{\partial a}(\mathbf{t}) + F_{X,X^*} \times D_m \times \mathrm{Re}\,\Phi_{Y|X^*}(\mathbf{t}).
\end{aligned}$$

The left-hand sides of these equations are all observed, while the right-hand sides contain all the unknowns. Assumption 2.3(i) also implies that F_{X,X^*}, $\mathrm{Re}\{\Phi_{m,a}(\mathbf{t})\}$ and $(\mathrm{Im}\{\Phi_{m,a}(\mathbf{t})\} + \Upsilon)$ are invertible in equations (2.5) and (2.7). Recall the definition of the observed matrix $C_{\mathbf{t}}$, which by equations (A.5) and (A.6) equals

$$C_{\mathbf{t}} \equiv (\mathrm{Re}\,\Phi_{Y,X}(\mathbf{t}))^{-1} \times (\mathrm{Im}\,\Phi_{Y,X}(\mathbf{t}) + \Upsilon_X) = \left(\mathrm{Re}\,\Phi_{Y|X^*}(\mathbf{t})\right)^{-1} \times \left(\mathrm{Im}\,\Phi_{Y|X^*}(\mathbf{t}) + \Upsilon\right).$$

Denote $A_{\mathbf{t}} \equiv \left(\mathrm{Re}\,\Phi_{Y|X^*}(\mathbf{t})\right)^{-1} \times D_m \times \mathrm{Re}\,\Phi_{Y|X^*}(\mathbf{t})$. With equations (A.5) and (A.7), we consider

$$\begin{aligned}
B_R &\equiv (\mathrm{Re}\,\Phi_{Y,X}(\mathbf{t}))^{-1} \times \left(\mathrm{Re}\,\frac{\partial}{\partial \mathbf{t}}\Phi_{Y,X}(\mathbf{t}) - \Upsilon_{E[Y|X]}\right)\\
&= \left(\mathrm{Re}\,\Phi_{m,a}(\mathbf{t}) \times D_{|\phi|}(\mathbf{t})\right)^{-1} \times \mathrm{Re}\,\Phi_{m,a}(\mathbf{t}) \times D_{\partial|\phi|}(\mathbf{t})\\
&\quad - \left(\mathrm{Re}\,\Phi_{Y|X^*}(\mathbf{t})\right)^{-1} \times \left(\mathrm{Im}\,\Phi_{Y|X^*}(\mathbf{t}) + \Upsilon\right) \times D_{\partial a}(\mathbf{t}) - \left(\mathrm{Re}\,\Phi_{Y|X^*}(\mathbf{t})\right)^{-1} \times D_m \times \left(\mathrm{Im}\,\Phi_{Y|X^*}(\mathbf{t}) + \Upsilon\right)\\
&= [D_{|\phi|}(\mathbf{t})]^{-1} \times D_{\partial|\phi|}(\mathbf{t}) - C_{\mathbf{t}} \times D_{\partial a}(\mathbf{t}) - \left(\left(\mathrm{Re}\,\Phi_{Y|X^*}(\mathbf{t})\right)^{-1} \times D_m \times \mathrm{Re}\,\Phi_{Y|X^*}(\mathbf{t})\right) \times C_{\mathbf{t}}\\
&\equiv D_{\partial \ln|\phi|}(\mathbf{t}) - C_{\mathbf{t}} \times D_{\partial a}(\mathbf{t}) - A_{\mathbf{t}} \times C_{\mathbf{t}}. && \text{(A.9)}
\end{aligned}$$

Similarly, we have by equations (A.6) and (A.8)

$$\begin{aligned}
B_I &\equiv (\mathrm{Im}\,\Phi_{Y,X}(\mathbf{t}) + \Upsilon_X)^{-1} \times \left(\mathrm{Im}\,\frac{\partial}{\partial \mathbf{t}}\Phi_{Y,X}(\mathbf{t})\right)\\
&= \left((\mathrm{Im}\,\Phi_{m,a}(\mathbf{t}) + \Upsilon) \times D_{|\phi|}(\mathbf{t})\right)^{-1} \times (\mathrm{Im}\,\Phi_{m,a}(\mathbf{t}) + \Upsilon) \times D_{\partial|\phi|}(\mathbf{t})\\
&\quad + \left(\mathrm{Im}\,\Phi_{Y|X^*}(\mathbf{t}) + \Upsilon\right)^{-1} \times \mathrm{Re}\,\Phi_{Y|X^*}(\mathbf{t}) \times D_{\partial a}(\mathbf{t}) + \left(\mathrm{Im}\,\Phi_{Y|X^*}(\mathbf{t}) + \Upsilon\right)^{-1} \times D_m \times \mathrm{Re}\,\Phi_{Y|X^*}(\mathbf{t}),\\
&= D_{\partial \ln|\phi|}(\mathbf{t}) + C_{\mathbf{t}}^{-1} \times D_{\partial a}(\mathbf{t}) + C_{\mathbf{t}}^{-1} \times A_{\mathbf{t}} && \text{(A.10)}
\end{aligned}$$

We eliminate the matrix $A_{\mathbf{t}}$ in equations (A.9) and (A.10) to have

$$\begin{aligned}
&B_R + C_{\mathbf{t}} \times B_I \times C_{\mathbf{t}}\\
&= D_{\partial \ln|\phi|}(\mathbf{t}) + C_{\mathbf{t}} \times D_{\partial \ln|\phi|}(\mathbf{t}) \times C_{\mathbf{t}} + D_{\partial a}(\mathbf{t}) \times C_{\mathbf{t}} - C_{\mathbf{t}} \times D_{\partial a}(\mathbf{t}). && \text{(A.11)}
\end{aligned}$$

Notice that both $D_{\partial \ln|\phi|}(\mathbf{t})$ and $D_{\partial a}(\mathbf{t})$ are diagonal, Assumption 2.3(ii) implies that $D_{\partial \ln|\phi|}(\mathbf{t})$ and $D_{\partial a}(\mathbf{t})$ are uniquely identified from equation (A.11).

Further, since the diagonal terms of $(D_{\partial a}(\mathbf{t}) \times C_{\mathbf{t}} - C_{\mathbf{t}} \times D_{\partial a}(\mathbf{t}))$ are zeros, we have

$$\begin{aligned} diag\left(B_R + C_{\mathbf{t}} \times B_I \times C_{\mathbf{t}}\right) &= diag\left(D_{\partial \ln|\phi|}(\mathbf{t})\right) + \left(C_{\mathbf{t}} \circ C_{\mathbf{t}}^T\right) \times diag\left(D_{\partial \ln|\phi|}(\mathbf{t})\right) \\ &\quad + D_{\partial a}(\mathbf{t}) \times diag\left(C_{\mathbf{t}}\right) - D_{\partial a}(\mathbf{t}) \times diag\left(C_{\mathbf{t}}\right) \\ &= \left[\left(C_{\mathbf{t}} \circ C_{\mathbf{t}}^T\right) + I\right] \times diag\left(D_{\partial \ln|\phi|}(\mathbf{t})\right), \end{aligned}$$

where the function $diag(\cdot)$ generates a vector of the diagonal entries of its argument and the notation "$\circ$" stands for the Hadamard product or the element-wise product. By assumption 2.5(i), we may solve $D_{\partial \ln|\phi|}(\mathbf{t})$ as follows:

$$diag\left(D_{\partial \ln|\phi|}(\mathbf{t})\right) = \left\{\left(C_{\mathbf{t}} \circ C_{\mathbf{t}}^T\right) + I\right\}^{-1} \times diag\left(B_R + C_{\mathbf{t}} \times B_I \times C_{\mathbf{t}}\right). \tag{A.12}$$

Furthermore, equation (A.11) implies that

$$\begin{aligned} U &\equiv B_R + C_{\mathbf{t}} \times B_I \times C_{\mathbf{t}} - D_{\partial \ln|\phi|}(\mathbf{t}) - C_{\mathbf{t}} \times D_{\partial \ln|\phi|}(\mathbf{t}) \times C_{\mathbf{t}} \\ &= D_{\partial a}(\mathbf{t}) \times C_{\mathbf{t}} - C_{\mathbf{t}} \times D_{\partial a}(\mathbf{t}), \end{aligned} \tag{A.13}$$

Define a J by 1 vector $e_1 = (1, 0, 0, ..., 0)^T$. The definition of $D_{\partial a}(\mathbf{t})$ implies that $e_1^T \times D_{\partial a}(\mathbf{t}) = 0$. Therefore, equation A.13 implies

$$e_1^T \times U = -e_1^T \times C_{\mathbf{t}} \times D_{\partial a}(\mathbf{t}).$$

Assumption 2.5(ii) implies that all the entries in the row vector $e_1^T \times C_{\mathbf{t}}$ are nonzero. Let $e_1^T \times C_{\mathbf{t}} \equiv (c_{11}, c_{12}, ..., c_{1J})$. The vector $diag\left(D_{\partial a}(\mathbf{t})\right)$ is then uniquely determined as follows:

$$diag\left(D_{\partial a}(\mathbf{t})\right) = -\begin{pmatrix} c_{11} & 0 & ... & 0 \\ 0 & c_{12} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & c_{1J} \end{pmatrix}^{-1} \times U^T \times e_1.$$

After $D_{\partial \ln|\phi|}(\mathbf{t})$ and $D_{\partial a}(\mathbf{t})$ are identified, we may then identify the matrix $A_{\mathbf{t}} \equiv \left(\mathrm{Re}\,\Phi_{Y|X^*}(\mathbf{t})\right)^{-1} \times D_m \times \mathrm{Re}\,\Phi_{Y|X^*}(\mathbf{t})$ from equation (A.10)

$$A_{\mathbf{t}} = C_{\mathbf{t}} \times \left(B_I - D_{\partial \ln|\phi|}(\mathbf{t})\right) - D_{\partial a}(\mathbf{t}).$$

Notice that

$$\mathrm{Re}\,\Phi_{Y|X^*}(\mathbf{t}) = \left(F_{X,X^*}\right)^{-1} \times \mathrm{Re}\,\Phi_{Y,X}(\mathbf{t}) = \left(F_{X|X^*} \times F_{X^*}\right)^{-1} \times \mathrm{Re}\,\Phi_{Y,X}(\mathbf{t})$$

where

$$\begin{aligned} F_{X,X^*} &= F_{X|X^*} \times F_{X^*}, \\ F_{X|X^*} &= \begin{pmatrix} f_{X|X^*}(1|1) & f_{X|X^*}(1|2) & ... & f_{X|X^*}(1|J) \\ f_{X|X^*}(2|1) & f_{X|X^*}(2|2) & ... & f_{X|X^*}(2|J) \\ ... & ... & ... & ... \\ f_{X|X^*}(J|1) & f_{X|X^*}(J|2) & ... & f_{X|X^*}(J|J) \end{pmatrix}, \\ F_{X^*} &= \begin{pmatrix} f_{X^*}(1) & 0 & ... & 0 \\ 0 & f_{X^*}(2) & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & f_{X^*}(J) \end{pmatrix}. \end{aligned}$$

Therefore, we have

$$\begin{aligned} \operatorname{Re}\Phi_{Y,X}(\mathbf{t}) \times A_{\mathbf{t}} \times (\operatorname{Re}\Phi_{Y,X}(\mathbf{t}))^{-1} &= \left(F_{X|X^*} \times F_{X^*}\right) \times D_m \times \left(F_{X|X^*} \times F_{X^*}\right)^{-1} \\ &= F_{X|X^*} \times D_m \times \left(F_{X|X^*}\right)^{-1}. \end{aligned} \tag{A.14}$$

Equation (A.14) implies that the unknowns m_j in matrix D_m are eigenvalues of a directly estimatable matrix on the left-hand side, and each column in the matrix $F_{X|X^*}$ is an eigenvector. Assumption 2.4 guarantees that all the eigenvalues are distinctive and nonzero in the diagonalization in equation (A.14). We may then identify m_j as the roots of

$$\det\left(A_{\mathbf{t}} - m_j I\right) = 0.$$

To be specific, m_j may be identified as the j-th smallest root. Equation (A.14) also implies that the j-th column in the matrix $F_{X|X^*}$ is the eigenvector corresponding to the eigenvalue m_j. Notice that each eigenvector is already normalized because each column of $F_{X|X^*}$ is a conditional density and the sum of entries in each column equals one. Therefore, each column of $F_{X|X^*}$ is identified as normalized eigenvectors corresponding to each eigenvalue m_j. Finally, we may identify f_{Y,X^*} through equation (2.1) as follows, for any $y \in \mathcal{Y}$.

$$\begin{aligned} &\begin{pmatrix} f_{Y,X^*}(y,1) & f_{Y,X^*}(y,2) & ... & f_{Y,X^*}(y,J) \end{pmatrix}^T \\ = \; & F_{X|X^*}^{-1} \times \begin{pmatrix} f_{Y,X}(y,1) & f_{Y,X}(y,2) & ... & f_{Y,X}(y,J) \end{pmatrix}^T. \end{aligned}$$

The identification of the joint distribution f_{Y,X^*} implies that both the latent model $f_{Y|X^*}$ and the marginal distribution of X^*, i.e., f_{X^*}, are identified. ■

Proof. (Theorem 2.3) First, we introduce notations as follows: for $j = 0, 1$

$$m_j = m(j), \quad \mu_j = E(Y|X=j),$$

$$v_j = E\left[\left(Y-\mu_j\right)^2 |X=j\right], \quad s_j = E\left[\left(Y-\mu_j\right)^3 |X=j\right],$$

$$p = f_{X^*|X}(1|0), \quad q = f_{X^*|X}(0|1), \quad f_{Y|X=j}(y) = f_{Y|X}(y|j).$$

We start the proof with equation (2.9), which is equivalent to

$$\begin{pmatrix} f_{Y|X}(y|0) \\ f_{Y|X}(y|1) \end{pmatrix} = \begin{pmatrix} f_{X^*|X}(0|0) & f_{X^*|X}(1|0) \\ f_{X^*|X}(0|1) & f_{X^*|X}(1|1) \end{pmatrix} \begin{pmatrix} f_{\eta|X^*=0}(y-m_0) \\ f_{\eta|X^*=1}(y-m_1) \end{pmatrix}. \tag{A.15}$$

Using the notations above, we have

$$\begin{pmatrix} f_{Y|X=0}(y) \\ f_{Y|X=1}(y) \end{pmatrix} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} \begin{pmatrix} f_{\eta|X^*=0}(y-m_0) \\ f_{\eta|X^*=1}(y-m_1) \end{pmatrix}.$$

Since $E[\eta|X^*] = 0$, we have

$$\mu_0 = (1-p)\, m_0 + p m_1 \text{ and } \mu_1 = q m_0 + (1-q)\, m_1.$$

We may solve for p and q as follows:

$$p = \frac{\mu_0 - m_0}{m_1 - m_0} \text{ and } q = \frac{m_1 - \mu_1}{m_1 - m_0}. \tag{A.16}$$

We also have

$$1 - p - q = 1 - \left(\frac{m_1 - m_0 + \mu_0 - \mu_1}{m_1 - m_0} \right) = \frac{\mu_1 - \mu_0}{m_1 - m_0}.$$

Assumption 2.9 implies that

$$m_1 \geq \mu_1 > \mu_0 \geq m_0.$$

and

$$\begin{pmatrix} f_{\eta|X^*=0}(y-m_0) \\ f_{\eta|X^*=1}(y-m_1) \end{pmatrix} = \frac{1}{1-p-q} \begin{pmatrix} 1-q & -p \\ -q & 1-p \end{pmatrix} \begin{pmatrix} f_{Y|X=0}(y) \\ f_{Y|X=1}(y) \end{pmatrix}.$$

Plug-in the expression of p and q in equation (A.16), we have

$$\frac{-p}{1-p-q} = \frac{m_0 - \mu_0}{\mu_1 - \mu_0}, \quad \frac{-q}{1-p-q} = \frac{\mu_1 - m_1}{\mu_1 - \mu_0},$$

$$\frac{1-p}{1-p-q} = 1 - \frac{-q}{1-p-q}, \quad \frac{1-q}{1-p-q} = 1 - \frac{-p}{1-p-q},$$

and

$$\begin{aligned} \begin{pmatrix} f_{\eta|X^*=0}(y-m_0) \\ f_{\eta|X^*=1}(y-m_1) \end{pmatrix} &= \begin{pmatrix} 1 - \frac{m_0-\mu_0}{\mu_1-\mu_0} & \frac{m_0-\mu_0}{\mu_1-\mu_0} \\ \frac{\mu_1-m_1}{\mu_1-\mu_0} & 1 - \frac{\mu_1-m_1}{\mu_1-\mu_0} \end{pmatrix} \begin{pmatrix} f_{Y|X=0}(y) \\ f_{Y|X=1}(y) \end{pmatrix} \\ &= \begin{pmatrix} \frac{\mu_1-m_0}{\mu_1-\mu_0} & \frac{m_0-\mu_0}{\mu_1-\mu_0} \\ \frac{\mu_1-m_1}{\mu_1-\mu_0} & \frac{m_1-\mu_0}{\mu_1-\mu_0} \end{pmatrix} \begin{pmatrix} f_{Y|X=0}(y) \\ f_{Y|X=1}(y) \end{pmatrix}. \end{aligned}$$

In other words, we have for $j = 0, 1$

$$f_{\eta|X^*=j}(y) = \frac{\mu_1 - m_j}{\mu_1 - \mu_0} f_{Y|X=0}(y + m_j) + \frac{m_j - \mu_0}{\mu_1 - \mu_0} f_{Y|X=1}(y + m_j). \tag{A.17}$$

In summary, $f_{X^*|X}$ (or p and q) and $f_{\eta|X^*}$ are identified if we can identify m_0 and m_1. Next, we show that m_0 and m_1 are indeed identified. By assumption 2.10, we have $E\left(\eta^k|X^*\right) = E\left(\eta^k\right)$ for $k = 2, 3$. For $k = 2$, we consider

$$\begin{aligned} v_1 &= E\left[\left(m(X^*) - \mu_1\right)^2 |X = 1\right] + E\left(\eta^2\right) \\ &= E\left[m(X^*)^2 | X = 1\right] - \mu_1^2 + E\left(\eta^2\right) = qm_0^2 + (1-q)\, m_1^2 - \mu_1^2 + E\left(\eta^2\right). \end{aligned}$$

Similarly, we have

$$v_0 = (1-p)\, m_0^2 + pm_1^2 - \mu_0^2 + E\left(\eta^2\right).$$

We eliminate $E\left(\eta^2\right)$ to obtain,

$$(1-p)\, m_0^2 + pm_1^2 - \left(v_0 + \mu_0^2\right) = qm_0^2 + (1-q)\, m_1^2 - \left(v_1 + \mu_1^2\right).$$

That is

$$\left(v_1 + \mu_1^2\right) - \left(v_0 + \mu_0^2\right) = (1 - p - q)\left(m_1^2 - m_0^2\right),$$

We have shown that

$$1 - p - q = \frac{\mu_1 - \mu_0}{m_1 - m_0}.$$

Thus, m_1 and m_0 satisfy the following linear equation:

$$m_1 + m_0 = \frac{\left(v_1 + \mu_1^2\right) - \left(v_0 + \mu_0^2\right)}{\mu_1 - \mu_0} \equiv C_1.$$

This means we need one more restriction to identify m_1 and m_0.We consider

$$\begin{aligned} s_1 &= E\left[(Y - \mu_1)^3 |X = 1\right] = E\left[\left(m\left(X^*\right) - \mu_1\right)^3 |X = 1\right] + E\left[\eta^3\right] \\ &= q\left(m_0 - \mu_1\right)^3 + (1-q)\left(m_1 - \mu_1\right)^3 + E\left[\eta^3\right] \end{aligned}$$

and

$$s_0 = (1-p)\left(m_0 - \mu_0\right)^3 + p\left(m_1 - \mu_0\right)^3 + E\left[\eta^3\right].$$

We eliminate $E\left(\eta^2\right)$ in the two equations above to obtain,

$$(1-p)\left(m_0 - \mu_0\right)^3 + p\left(m_1 - \mu_0\right)^3 - s_0 = q\left(m_0 - \mu_1\right)^3 + (1-q)\left(m_1 - \mu_1\right)^3 - s_1$$

Plug in the expression of p and q in equation (A.16), we have

$$-\left(m_1 - \mu_0\right)\left(m_0 - \mu_0\right)\left(m_1 + m_0 - 2\mu_0\right) - s_0 = -\left(m_1 - \mu_1\right)\left(m_0 - \mu_1\right)\left(m_1 + m_0 - 2\mu_1\right) - s_1,$$

Since $m_1 + m_0 = C_1$, we have

$$\left(C_1 - m_0 - \mu_0\right)\left(m_0 - \mu_0\right)\left(C_1 - 2\mu_0\right) + s_0 = \left(C_1 - m_0 - \mu_1\right)\left(m_0 - \mu_1\right)\left(C_1 - 2\mu_1\right) + s_1,$$

that is,

$$\begin{aligned} & -\left(m_0^2-\mu_0^2\right)\left(C_1-2\mu_0\right)+\left(m_0-\mu_0\right)C_1\left(C_1-2\mu_0\right)+s_0 \\ = \; & -\left(m_0^2-\mu_1^2\right)\left(C_1-2\mu_1\right)+\left(m_0-\mu_1\right)C_1\left(C_1-2\mu_1\right)+s_1 \end{aligned}$$

Moreover, we have

$$\begin{aligned} & -2m_0^2\left(\mu_1-\mu_0\right)+2C_1\left(\mu_1-\mu_0\right)m_0 \\ = \; & \mu_1^2\left(C_1-2\mu_1\right)-\mu_0^2\left(C_1-2\mu_0\right)-\mu_1C_1\left(C_1-2\mu_1\right)+\mu_0C_1\left(C_1-2\mu_0\right)+s_1-s_0 \\ = \; & \left(\mu_1^2-\mu_0^2\right)C_1-2\left(\mu_1^3-\mu_0^3\right)-\left(\mu_1-\mu_0\right)C_1^2+2\left(\mu_1^2-\mu_0^2\right)C_1+s_1-s_0 \end{aligned}$$

Since $(\mu_1-\mu_0)>0$, we have

$$-2m_0^2+2C_1m_0=3\left(\mu_1+\mu_0\right)C_1-2\frac{\mu_1^3-\mu_0^3}{\mu_1-\mu_0}-C_1^2+\frac{s_1-s_0}{\mu_1-\mu_0}.$$

Finally, we have

$$-2\left(m_0-\frac{1}{2}C_1\right)^2+C_2=0$$

where

$$\begin{aligned} C_2 &= \frac{3}{2}C_1^2-3\left(\mu_1+\mu_0\right)C_1+2\frac{\mu_1^3-\mu_0^3}{\mu_1-\mu_0}-\frac{s_1-s_0}{\mu_1-\mu_0} \\ &= \frac{3}{2}\left[C_1-\left(\mu_1+\mu_0\right)\right]^2-\frac{3}{2}\left(\mu_1+\mu_0\right)^2+2\left(\mu_1^2+\mu_1\mu_0+\mu_0^2\right)-\frac{s_1-s_0}{\mu_1-\mu_0} \\ &= \frac{3}{2}\left[C_1-\left(\mu_1+\mu_0\right)\right]^2+\frac{1}{2}\left(\mu_1-\mu_0\right)^2-\frac{s_1-s_0}{\mu_1-\mu_0} \\ &= \frac{1}{2}\left(\mu_1-\mu_0\right)^2+\frac{3}{2}\left(\frac{v_1-v_0}{\mu_1-\mu_0}\right)^2-\frac{s_1-s_0}{\mu_1-\mu_0} \end{aligned}$$

$$\begin{aligned} s_j &= E\left[\left(Y-\mu_j\right)^3|X=j\right] \\ &= E\left[Y^3|X=j\right]-3E\left[Y^2|X=j\right]\mu_j+3\mu_j^3-\mu_j^3 \\ &= E\left[Y^3|X=j\right]-3E\left[Y^2|X=j\right]\mu_j+2\mu_j^3 \\ &\equiv \kappa_j-3\upsilon_j\mu_j+2\mu_j^3, \end{aligned}$$

$$\begin{aligned}
C_2 &= \frac{3}{2}C_1^2 - 3(\mu_1+\mu_0)C_1 + 2\frac{\mu_1^3-\mu_0^3}{\mu_1-\mu_0} - \frac{s_1-s_0}{\mu_1-\mu_0} \\
&= \frac{3}{2}C_1^2 - 3(\mu_1+\mu_0)C_1 + 2\frac{\mu_1^3-\mu_0^3}{\mu_1-\mu_0} - \frac{\kappa_1 - 3\upsilon_1\mu_1 + 2\mu_1^3 - (\kappa_0 - 3\upsilon_0\mu_0 + 2\mu_0^3)}{\mu_1-\mu_0} \\
&= \frac{3}{2}C_1^2 - 3(\mu_1+\mu_0)C_1 - \frac{\kappa_1 - 3\upsilon_1\mu_1 - (\kappa_0 - 3\upsilon_0\mu_0)}{\mu_1-\mu_0} \\
&= \frac{3}{2}C_1^2 - 3(\mu_1+\mu_0)\frac{\upsilon_1-\upsilon_0}{\mu_1-\mu_0} + \frac{3\upsilon_1\mu_1 - 3\upsilon_0\mu_0}{\mu_1-\mu_0} - \frac{\kappa_1-\kappa_0}{\mu_1-\mu_0} \\
&= \frac{3}{2}\left(\frac{\upsilon_1-\upsilon_0}{\mu_1-\mu_0}\right)^2 - 3\frac{\upsilon_0\mu_1 - \upsilon_1\mu_0}{\mu_1-\mu_0} - \frac{\kappa_1-\kappa_0}{\mu_1-\mu_0}
\end{aligned}$$

Notice that we also have

$$-2\left(m_1 - \frac{1}{2}C_1\right)^2 + C_2 = 0,$$

which implies that m_1 and m_0 are two roots of this quadratic equation. Since $m_1 > m_0$, we have

$$m_0 = \frac{1}{2}C_1 - \sqrt{\frac{1}{2}C_2}, \quad m_1 = \frac{1}{2}C_1 + \sqrt{\frac{1}{2}C_2}.$$

After we have identified m_0 and m_1, p and q (or $f_{X^*|X}$) are identified from equation (A.16), and the density f_η(or $f_{Y|X^*}$) is also identified from equation (A.17). Thus, we have identified the latent densities $f_{Y|X^*}$ and $f_{X^*|X}$ from the observed density $f_{Y|X}$, and summing $f_{X^*|X}$ over X gives f_{X^*}. ■

REFERENCES

Ai, C., and Chen, X. 2003, "Efficient Estimation of Models With Conditional Moment Restrictions Containing Unknown Functions," Econometrica, 71, 1795-1843.

Aigner, D. J. 1973, "Regression With a Binary Independent Variable Subject to Errors of Obervation," Journal of Econometrics, 1, 249-60.

Balke, A. and J. Pearl 1997, "Bounds on Treatment Effects from Studies with Imperfect Compliance," Journal of the American Statistical Association, 92, 1171-1176.

Bickel, P. J.; Ritov, Y., 1987, "Efficient estimation in the errors in variables model." Ann. Statist. 15, no. 2, 513–540.

Bollinger, C. R. 1996, "Bounding Mean Regressions When a Binary Regressor is Mismeasured," Journal of Econometrics, 73, 387-399.

Bordes, L., S. Mottelet, and P. Vandekerkhove, 2006, "Semiparametric estimation of a two-component mixture model," Annals of Statistics, 34, 1204-232.

Bound, J. C. Brown and N. Mathiowetz, 2001, "Measurement error in survey data", in Handbook of Econometrics, Vol. 5, ed. by J. Heckman and E. Leamer. North Holland.

Carroll, R.J., D. Puppert, C. Crainiceanu, T. Tostenson and M. Karagas, 2004, "Nonlinear and Nonparametric Regression and Instrumental Variables," Journal of the American Statistical Association, 99 467, pp. 736-750.

Carroll, R.J., D. Puppert, L. Stefanski and C. Crainiceanu, 2006, *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*, CRI.

Carroll, R.J. and L.A. Stefanski, 1990, "Approximate quasi-likelihood estimation in models with surrogate predictors," Journal of the American Statistical Association 85, pp. 652-663.

Chen, X. 2006: "Large Sample Sieve Estimation of Semi-nonparametric Models", in J.J. Heckman and E.E. Leamer (eds.), The Handbook of Econometrics, vol. 6. North-Holland, Amsterdam, forthcoming.

Chen, X., H. Hong, and D. Nekipelov, 2007, "Measurement error models," working paper of New York University and Stanford University, a survey prepared for the Journal of Economic Literature.

Chen, X., H. Hong, and E. Tamer, 2005, "Measurement error models with auxiliary data," Review of Economic Studies, 72, pp. 343-366.

Chen, X., H. Hong, and A. Tarozzi 2007: "Semiparametric Efficiency in GMM Models with Auxiliary Data," Annals of Statistics, forthcoming.

Cheng, C. L., Van Ness, J. W., 1999, *Statistical Regression with Measurement Error*, Arnold, London.

Chua, T. C. and W. A Fuller, 1987, A Model For Multinomial Response Error Applied to Labor Flows," Journal of the American Statistical Association, 82, 46-51.

Finney, D. J. 1964Statistical Method in Biological Assay. Havner: New York.

Fuller, W., 1987, *Measurement error models*. New York: John Wiley & Sons.

Geman, S., and Hwang, C. 1982, "Nonparametric Maximum Likelihood Estimation by the Method of Sieves," The Annals of Statistics, 10, 401-414.

Grenander, U. 1981, *Abstract Inference*, New York: Wiley Series.

Gustman, A. L. and T. L. Steinmeier, 2004, "Social security, pensions and retirement behaviour within the family," Journal of Applied Econometrics, 19, 723-737.

Hansen, L.P. 1982: "Large Sample Properties of Generalized Method of Moments Estimators," Econometrica, 50, 1029-1054.

Hausman, J. A., J. Abrevaya, and F. M. Scott-Morton 1998, "Misclassification of the Dependent Variable in a Discrete-Response Setting," Journal of Econometrics, 87, 239-269.

Hausman, J. A., Ichimura, H., Newey, W., and Powell, J., 1991, "Identification and estimation of polynomial errors-in-variables models," Journal of Econometrics, 50, pp. 273-295.

Hirsch, B.T. and D. A. Macpherson 2003, "Union Membership and Coverage Database from the Current Population Survey: Note," Industrial and Labor Relations Review, 56, 349-354.

Hsiao, C., 1991, "Identification and estimation of dichotomous latent variables models using panel data," Review of Economic Studies 58, pp. 717-731.

Hu, Y. 2006: "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables," Working Paper, University of Texas at Austin.

Hu, Y, and G. Ridder, 2006, "Estimation of Nonlinear Models with Measurement Error Using Marginal Information," Working Paper, University of Southern California.

Hu, Y, and S. Schennach, 2006,"Identification and estimation of nonclassical nonlinear errors-in-variables models with continuous distributions using instruments," Cemmap Working Papers CWP17/06.

Kane, T. J., and C. E. Rouse, 1995, "Labor market returns to two- and four- year college," American Economic Review, 85, 600-614

Kendall, M. and A. Stuart, 1979, The Advanced Theory of Statistics, Macmillan, New York, 4th edition.

Lee, L.-F., and J.H. Sepanski, 1995, "Estimation of linear and nonlinear errors-in-variables models using validation data," Journal of the American Statistical Association, 90 (429).

Lewbel, A., 1997, "Constructing Instruments for Regressions With Measurement Error When No Additional Data are Available, With an Application to Patents and R&D," Econometrica, 65, 1201-1213.

Lewbel, A., 2000, "Identification of the Binary Choice Model With Misclassification," Econometric Theory, 16, 603-609.

Lewbel, A., 2007, "Estimation of average treatment effects with misclassification," Econometrica, 2007, 75, 537-551.

Li, T., 2002, "Robust and consistent estimation of nonlinear errors-in-variables models," Journal of Econometrics, 110, pp. 1-26.

Li, T., and Q. Vuong, 1998, "Nonparametric estimation of the measurement error model using multiple indicators," Journal of Multivariate Analysis, 65, pp. 139-165.

Liang, H., W. Hardle, and R. Carroll, 1999, "Estimation in a Semiparametric Partially Linear Errors-in-Variables Model," The Annals of Statistics, Vol. 27, No. 5, 1519-1535.

Liang, H.; Wang, N., 2005, "Partially linear single-index measurement error models," Statist. Sinica 15, no. 1, 99–116.

Mahajan, A. 2006: "Identification and estimation of regression models with misclassification," Econometrica, vol. 74, pp. 631-665.

Murphy, S. A. and Van der Vaart, A. W. 1996, "Likelihood inference in the errors-in-variables model." J. Multivariate Anal. 59, no. 1, 81-08.

Murphy, S. A. and Van der Vaart, A. W. 2000, "On Profile Likelihood", Journal of the American Statistical Association, 95, 449-486.

Newey, W.K. and J. Powell 2003: "Instrumental Variables Estimation for Nonparametric Models," Econometrica, 71, 1565-1578.

Poterba, J. M. and L. H. Summers 1995"Unemployment Benefits and Labor Market Transitions: A Multinomial Logit Model With Errors in Classification," Review of Economics and Statistics, 77, 207-216.

Reiersol, O. 1950: "Identifiability of a Linear Relation between Variables Which Are Subject to Error," Econometrica, 18, 375-389.

Schennach, S. 2004: "Estimation of Nonlinear Models with Measurement Error," Econometrica, 72, 33-75.

Shen, X. 1997, "On Methods of Sieves and Penalization," The Annals of Statistics, 25, 2555-2591.

Shen, X., and Wong, W. 1994, "Convergence Rate of Sieve Estimates," The Annals of Statistics, 22, 580-615.

Taupin, M. L., 2001, "Semi-parametric estimation in the nonlinear structural errors-in-variables model," Annals of Statistics, 29, pp. 66-93.

Van de Geer, S. 1993, "Hellinger-Consistency of Certain Nonparametric Maximum Likelihood Estimators," The Annals of Statistics, 21, 14-44.

Van de Geer, S. 2000, *Empirical Processes in M-estimation*, Cambridge University Press.

Van der Vaart, A. and J. Wellner 1996: Weak Convergence and Empirical Processes: with Applications to Statistics. New York: Springer-Verlag.

Wang, L., 2004, "Estimation of nonlinear models with Berkson measurement errors," The Annals of Statistics 32, no. 6, 2559–2579.

Wang, N., X. Lin, R. Gutierrez, and R. Carroll, 1998, "Bias analysis and SIMEX approach in generalized linear mixed measurement error models," J. Amer. Statist. Assoc. 93, no. 441, 249–261.

Wansbeek, T. and E. Meijer, 2000, *Measurement Error and Latent Variables in Econometrics*. New York: North Holland.

Wong, W., and Shen, X. 1995, "Probability Inequalities for Likelihood Ratios and Convergence Rates for Sieve MLE's," The Annals of Statistics, 23, 339-362.

Table 1: Simulation results

Example 1 Value of x^*:	1	2	3	4
Regression function $m(x^*)$:				
– true value	1.3500	2.8000	5.9500	11.400
– mean estimate	1.2984	2.9146	6.0138	11.433
– standard error	0.2947	0.3488	0.2999	0.2957
Marginal distribution $\Pr(x^*)$:				
– true value	0.2	0.3	0.3	0.2
– mean estimate	0.2159	0.2818	0.3040	0.1983
– standard error	0.1007	0.2367	0.1741	0.0153
Misclasification Prob. $f_{x\|x^*}(\cdot\|x^*)$:				
– true value	0.6	0.2	0.1	0.1
	0.2	0.6	0.1	0.1
	0.1	0.1	0.7	0.1
	0.1	0.1	0.1	0.7
– mean estimate	0.5825	0.2008	0.0991	0.0986
	0.2181	0.5888	0.1012	0.0974
	0.0994	0.1137	0.6958	0.0993
	0.1001	0.0967	0.1039	0.7047
– standard error	0.0788	0.0546	0.0201	0.0140
	0.0780	0.0788	0.0336	0.0206
	0.0387	0.0574	0.0515	0.0281
	0.0201	0.0192	0.0293	0.0321

Table 2: Simulation results

Example 2 Value of x^*:	1	2	3	4
Regression function $m(x^*)$:				
– true value	1.3500	2.8000	5.9500	11.400
– mean estimate	1.2320	3.1627	6.1642	11.514
– standard error	0.4648	0.7580	0.7194	0.6940
Marginal distribution $\Pr(x^*)$:				
– true value	0.2	0.3	0.3	0.2
– mean estimate	0.2244	0.3094	0.2657	0.2005
– standard error	0.1498	0.1992	0.1778	0.0957
Misclasification Prob. $f_{x\|x^*}(\cdot\|x^*)$:				
– true value	0.5220	0.1262	0.2180	0.2994
	0.1881	0.4968	0.1719	0.2489
	0.1829	0.1699	0.4126	0.0381
	0.1070	0.2071	0.1976	0.4137
– mean estimate	0.4761	0.1545	0.2214	0.2969
	0.2298	0.4502	0.1668	0.2455
	0.1744	0.1980	0.4063	0.0437
	0.1197	0.1973	0.2056	0.4140
– standard error	0.1053	0.0696	0.0343	0.0215
	0.0806	0.0771	0.0459	0.0262
	0.0369	0.0528	0.0573	0.0313
	0.0327	0.0221	0.0327	0.0238

Table 3: Simulation results ($n = 3000, reps = 400$)

true value of β:	$\beta_1 = 1$	$\beta_2 = 1$	$\beta_3 = 1$
ignoring meas. error:			
– mean estimate	2.280	1.636	0.9474
– standard error	0.1209	0.1145	0.07547
– root mse	1.286	0.6461	0.09197
infeasible MLE:			
– mean estimate	0.9950	1.012	0.9900
– standard error	0.05930	0.08263	0.07048
– root mse	0.05950	0.08346	0.07118
sieve MLE:			
– mean estimate	0.9760	0.9627	0.9834
– standard error	0.1366	0.06092	0.1261
– root mse	0.1387	0.07145	0.1272

note: $K_n = 3$ in $\widehat{f_\eta}$ and $\widehat{f_{X|X^*,W}}(x|x^*, w)$ for each x and x^*.

Nonparametric identification of the classical errors-in-variables model without side information.

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

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

A. Lewbel
Department of Economics
Boston College
140 Commonwealth Avenue
Chestnut Hill, MA 02467
lewbel@bc.edu

July 16, 2007

Abstract

This note establishes that the fully nonparametric classical errors-in-variables model is identifiable from data on the regressor and the dependent variable alone, unless the specification is a member of a very specific parametric family. This family includes the linear specification with normally distributed variables as a special case. This result relies on standard primitive regularity conditions taking the form of smoothness and monotonicity of the regression function and nonvanishing characteristic functions of the disturbances.

*Corresponding author. S. M. Schennach acknowledges support from the National Science Foundation via grant SES-0452089.

1 Introduction

The identification of regression models in which both the dependent and independent variables are measured with error has received considerable attention over the last few decades. 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 such that*

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

where only x and y are observed while all remaining variables are not and where x^*, Δx, Δy, are mutually independent, $E[\Delta x] = 0$ and $E[\Delta y] = 0$.

A well-known result is that when $g(x^*)$ is linear while x^*, Δx and Δy are normal, the model is not identified, although the regression coefficients can often be consistently bounded (Klepper and Leamer (1984)).[1] This negative result 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), Schennach (2004a), Schennach (2004b), Schennach (2007), Hu and Schennach (2006), Hu and Ridder (2004), among many others).

Nevertheless, since the seminal work of Geary (1942), a large number of authors (e.g. 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 (2006), and the many references therein) have suggested alternative methods to identify a linear regression with nonnormally distributed regressors based on the idea that higher order moments of x and y then provide additional information that can be exploited. However, the question of characterizing the set of identifiable models in fully nonparametric settings while exploiting the joint distribution of all the observable variables remains wide open.

[1] Chesher (1998) suggests some settings where a polynomial regression is not identified based on the knowledge of *some* of the moments of the observed data.

We demonstrate that the answer to this 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.

2 Identification result

We need a few basic regularity conditions.

Assumption 1 *$E\left[e^{i\xi\Delta x}\right]$ 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 Schennach (2004a) and the references therein). Without it, the measurement error effectively masks information regarding the true variables that cannot be recovered.[2] The only commonly encountered distributions with a vanishing characteristic function are the uniform and the triangular distributions.

Assumption 2 *The distribution of x^* admits a finite density $f_{x^*}(x^*)$ with respect to the Lebesgue measure.*

This assumption rules out pathological case such as fractal-like distributions. It also rules out discrete distributions.[3]

Assumption 3 *The regression function $g(x^*)$ has a continuous, finite and nonvanishing first derivative at each point*[4] *in the interior of the support of x^*.*

[2] Although our approach could probably be extended to the case of characteristic functions vanishing at isolated points in $\mathbb{R}$ along the lines of Hu and Ridder (2004).

[3] An extension of our result to purely discrete distributions is straightforward, although such a result would not be very useful in the context of classical measurement error.

[4] It need not be *uniformly* bounded above and below.

This is a smoothness and monotonicity constraint. Without it, it is difficult to rule out extremely complex and pathological joint distributions of x and y (including, once again, fractal-like distributions). In particular, one could imagine an extremely rapidly oscillating $g(x^*)$, where nearly undetectable changes in x^* yield changes in y that are almost observationally indistinguishable from genuine errors in y. Even relaxing this assumption to include less pathological functional forms that oscillate a finite number of times is difficult, due to the overlap between the measurement error distributions in regions where the regression function is not one-to-one and due to the appearance of divergences in some of the densities entering the model. Many recent nonparametric identification results also rely on monotonicity assumptions, as discussed, for instance, in the Handbook of Econometrics chapter by Matzkin (2007).

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

Definition 1 *We say that a random variable r is* decomposable with 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.*

Theorem 1 *Let Assumptions 1, 2 and 3 hold.*

1. *If $g(x^*)$ 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 $f_{x^*}(x^*)$) in Model 1 are ***identified****.*

2. *If $g(x^*)$* ***is*** *of the form (1) with[5] $d > 0$, 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[6] $C \in \mathbb{R}$, $A, B, D, E, F \in [0, \infty[$ and Δx and Δy are decomposable with a type I extreme value factor.[7]

[5] A case where $d < 0$ can be converted into a case with $d > 0$ by permuting the roles of x and y.

[6] The constants A, B, C, D, E, F depend on a, b, c, d, although this dependence is omitted here for simplicity. Constants yielding a valid density can be found for any a, b, c, d (with $d > 0$).

[7] A type I extreme value distribution has a density of the general form $f(u) = K_1\exp(K_2\exp(K_3u) + K_4u)$. Here, the constant K_1, K_2, K_3, K_4 are such that $f(u)$ integrates to 1 and has zero mean and may depend on a, b, c, d, although this dependence is omitted here for simplicity.

3. *If $g(x^*)$ **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 is decomposable with a normal factor.*

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 identifiablity out have different regression functions in all of the unidentified cases.

The proof of this result (provided in the Appendix) proceeds in four steps:

1. We reduce the identification problem of a model with errors along x and y into the equivalent problem of finding two observationally 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 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.

4. 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 also surprising that, even in the presumably difficult case of normally distributed regressors, most nonlinear specifications are actually identified. 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.

Theorem 1 can be extended in various useful directions. For instance, perfectly observed covariates w can be included simply by conditioning all densities (and expectations) on these covariates. We then establish identification of $f_{x^*|w}(x^*|w)$ and $g(x^*, w) \equiv E[y|x^*, w]$ and therefore of $f_{x^*,w}(x^*, w) = f_{x^*|w}(x^*|w) f_w(w)$. The above results do not yet establish identification of the measurement error distributions, but this can be trivially achieved by deconvolution techniques (once $g(x^*)$ and $f_{x^*}(x^*)$ have been determined) under the additional assumption that $E\left[e^{i\xi x^*}\right]$ and $E\left[e^{i\gamma g(x^*)}\right]$ do not vanish.

3 Conclusion

This note 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. Our identification result is agnostic regarding the type of estimator to be used in practice. One could use higher-order moment equalities, characteristic function equalities, or nonparametric sieve-type likelihoods. Finding the most convenient and statistically powerful method remains a nontrivial and important avenue of future research. It would also be useful to investigate whether these results extend to the case of nonclassical measurement error (i.e. relaxing some of the independence assumptions), where the dimensionality of the unknown distributions is greater or equal to the dimensionality of the observable distributions.

A Proof of Theorem 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).

Consider an alternative observationally equivalent model defined as:

Model 2 *Similar to Model 1 with* $x^*, \Delta x, \Delta y, g(\cdot)$ *replaced, respectively, by* $\tilde{x}^*, \Delta\tilde{x}, \Delta\tilde{y}, \tilde{g}(\cdot)$.

It is clear that any assumptions (including regularity conditions) made regarding Model 1 must hold for this alternative model as well.

We first reduce the identification problem to a simpler but equivalent problem involving only one error term. Consider the following two models:

Model 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), 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[8] of the distribution of Δx in Model 1.

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

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

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

Note that, given the above definitions, $\Delta x = \Delta\bar{x} + \Delta\tilde{x}$. This assumes, without loss of generality, that the distribution of $\Delta\tilde{x}$ is a factor of the distribution of Δx (otherwise, one can just permute the role of Models 1 and 2, which interchanges the role of tilded and non tilded symbols).

Lemma 1 *Under Assumptions 1-3, there exist two distinct observationally equivalent Models 1 and 2 iff there exist two distinct observationally equivalent models of the form of Models 3 and 4. Moreover, when two such models exist, the distributions of $\bar{x}, \bar{y}, \Delta\bar{x}$ and $\Delta\bar{y}$ all admit a density with respect to the Lebesgue measure and are supported on all of $\mathbb{R}$.*

Proof. **(1)** The joint characteristic function 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. Under Model 1,

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

The independence conditions stated in Model 1 then imply that

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

[8] A distribution F is said to be a *factor* of a distribution H if there exists a distribution G (which may be degenerate) such that the random variable $h = f + g$ has distribution H, where f, g are independent random variables drawn from F, G respectively.

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

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

Define

$$\alpha(\xi) \equiv \frac{E\left[e^{i\xi\Delta x}\right]}{E\left[e^{i\xi\Delta\tilde{x}}\right]} \qquad \beta(\gamma) \equiv \frac{E\left[e^{i\gamma\Delta\tilde{y}}\right]}{E\left[e^{i\gamma\Delta y}\right]}$$

and note that $\alpha(\xi)$ and $\beta(\gamma)$ are everywhere continuous, nonvanishing and finite.[9] Also, $\alpha(0) = 1$, $\alpha'(0) = 0$ and $\beta(0) = 1$, $\beta'(0) = 0$. Rearranging, we obtain

$$\begin{aligned} E\left[e^{i\xi\Delta\tilde{x}}\right] &= \frac{E\left[e^{i\xi\Delta x}\right]}{\alpha(\xi)} \\ E\left[e^{i\gamma\Delta\tilde{y}}\right] &= \beta(\gamma)E\left[e^{i\gamma\Delta y}\right]. \end{aligned}$$

Substituting these expressions into (4), yields

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

But, by (3), this is also equal to $E\left[e^{i\xi x^*}e^{i\gamma g(x^*)}\right]E\left[e^{i\xi\Delta x}\right]E\left[e^{i\gamma\Delta y}\right]$ and therefore

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

Since $E\left[e^{i\xi\Delta x}\right]$, $E\left[e^{i\gamma\Delta y}\right]$ and $\alpha(\xi)$ are finite and nonvanishing, we can multiply each side by $\alpha(\xi)/\left(E\left[e^{i\xi\Delta x}\right]E\left[e^{i\gamma\Delta y}\right]\right)$ to yield:

$$E\left[e^{i\xi x^*}e^{i\gamma g(x^*)}\right]\alpha(\xi) = E\left[e^{i\xi\tilde{x}^*}e^{i\gamma\tilde{g}(\tilde{x}^*)}\right]\beta(\gamma) \tag{5}$$

or

$$E\left[e^{i\xi x^*}e^{i\gamma g(x^*)}\right]\alpha(\xi)E\left[e^{i\gamma 0}\right] = E\left[e^{i\xi\tilde{x}^*}e^{i\gamma\tilde{g}(\tilde{x}^*)}\right]E\left[e^{i\xi 0}\right]\beta(\gamma).$$

In other words, Models 1 and 2 are observationally equivalent iff there exists a model with errors only in the regressor (Model 3, where $\alpha(\xi)$ is the characteristic function of $\Delta\bar{x}$) that is observationally equivalent to a model with errors in the dependent variable (Model 4, where $\beta(\gamma)$ is the characteristic function of $\Delta\bar{y}$). This completes the first part the proof.

[9] That is, finite at each point, though not necessarily uniformly bounded.

(2) It remains to be shown that we can indeed limit ourselves to $\alpha(\xi)$ and $\beta(\gamma)$ that are valid characteristic functions and, more specifically, to characteristic functions of densities supported on $\mathbb{R}$. Define $y^* \equiv g(x^*)$ and $h(y^*) \equiv g^{-1}(y^*)$ and note that y^* admits a density.$f_{y^*}(y^*) = f_{x^*}(h(y^*))/g'(h(y^*))$ since $g'(x^*) \neq 0$ by assumption. We can then rewrite Equation (5) as

$$E\left[e^{i\xi h(y^*)}e^{i\gamma y^*}\right]\alpha(\xi) = E\left[e^{i\xi\tilde{x}^*}e^{i\gamma\tilde{g}(\tilde{x}^*)}\right]\beta(\gamma). \tag{6}$$

We now calculate the inverse Fourier transform (FT) of each side using the convolution theorem. To this effect, we calculate the FT of each term individually. Since we can write

$$E\left[e^{i\xi h(y^*)}e^{i\gamma y^*}\right] = \int\int \delta(x^* - h(y^*))\, f_{y^*}(y^*)\, e^{i\xi x^*}e^{i\gamma y^*}dx^*dy^*$$

$$E\left[e^{i\xi\tilde{x}^*}e^{i\gamma\tilde{g}(\tilde{x}^*)}\right] = \int\int \delta(\tilde{y}^* - \tilde{g}(\tilde{x}^*))\, \tilde{f}_{\tilde{x}^*}(\tilde{x}^*)\, e^{i\xi\tilde{x}^*}e^{i\gamma\tilde{y}^*}d\tilde{x}^*d\tilde{y}^*,$$

the inverse FT of $E\left[e^{i\xi h(y^*)}e^{i\gamma y^*}\right]$ and $E\left[e^{i\xi\tilde{x}^*}e^{i\gamma\tilde{g}(\tilde{x}^*)}\right]$ are, respectively, $\delta(x^* - h(y^*))\, f_{y^*}(y^*)$ and $\delta(\tilde{y}^* - \tilde{g}(\tilde{x}^*))\, \tilde{f}_{\tilde{x}^*}(\tilde{x}^*)$, where $\delta(\cdot)$ denotes a delta function.

Let $\mathcal{W}_{\Delta\bar{x}}$ denote the set where the inverse FT of $\alpha(\xi)$ is well-defined and finite[10] and let $f_{\Delta\bar{x}}(\Delta\bar{x})$ denote this inverse FT for $\Delta\bar{x} \in \mathcal{W}_{\Delta\bar{x}}$. Similarly define $\mathcal{W}_{\Delta\bar{y}}$ and $\tilde{f}_{\Delta\bar{y}}(\Delta\bar{y})$ for $\beta(\gamma)$. Note that the sets $\mathcal{W}_{\Delta\bar{x}}$ and $\mathcal{W}_{\Delta\bar{y}}$ cannot be empty since it would then be impossible for $\alpha(\xi)$ and $\beta(\gamma)$ to be finite everywhere.[11] By (6) and the convolution theorem, we have

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

where we have used the equivalence $y^* = \bar{y}$ (under Model 3) and $\bar{x} = \tilde{x}^*$ (under Model 4). Using the properties of the delta function $\delta(\cdot)$,

$$f_{\bar{y}}(\bar{y})\, f_{\Delta\bar{x}}(\bar{x} - h(\bar{y})) = f_{\bar{x}}(\bar{x})\, \tilde{f}_{\Delta\bar{y}}(\bar{y} - \tilde{g}(\bar{x})) \tag{7}$$

an equality which holds for $(\bar{x}, \bar{y})$ such that $\bar{x} - h(\bar{y}) \in \mathcal{W}_{\Delta\bar{x}}$ and $\bar{y} - \tilde{g}(\bar{x}) \in \mathcal{W}_{\Delta\bar{y}}$.

Suppose that at some point $(\bar{x}_0, \bar{y}_0)$ in the interior of the support of $(\bar{x}, \bar{y})$, we have that $\tilde{f}_{\Delta\bar{y}}(\bar{y}_0 - \tilde{g}(\bar{x}_0))$ changes sign, becomes zero, infinite or undefined. Then the same behavior must necessarily occur in $f_{\Delta\bar{x}}(\bar{x}_0 - h(\bar{y}_0))$ at the same point $(\bar{x}_0, \bar{y}_0)$

[10] That is, for a given $\Delta\bar{x}$, $\lim_{t\to\infty}\int_{-t}^{t}\alpha(\xi)\, e^{i\xi\Delta\bar{x}}d\xi$ exists in $\mathbb{C}$.

[11] If $\mathcal{W}_{\Delta\bar{x}}$ is empty, $\tilde{f}_{\Delta\bar{x}}(\Delta\bar{x})$ would be undefined or infinite for all points in $\mathbb{R}$, hence its Fourier transform $\alpha(\xi)$ could not exist.

because multiplication by a bounded positive number (here, $f_{\bar{y}}(\bar{y}_0)$ and $f_{\bar{x}}(\bar{x}_0)$ are finite by assumption) does not affect whether a quantity is well-defined, positive, nonzero or finite. Furthermore, the same behavior would occur along the whole curve $(\bar{x}, \bar{y})$ giving the same value of $v \equiv \bar{y}_0 - \tilde{g}(\bar{x}_0) = \bar{y} - \tilde{g}(\bar{x})$ or the same value of $u \equiv \bar{x}_0 - h(\bar{y}_0) = \bar{x} - h(\bar{y})$. If the curves

$$\mathcal{V}_v = \{(\tilde{x}^*, \tilde{g}(\tilde{x}^*) + v) : \tilde{x}^* \in \mathcal{S}_{\tilde{x}^*}\} \text{ and } \mathcal{U}_u = \{(h(y^*) + u, y^*) : y^* \in \mathcal{S}_{y^*}\} \quad (8)$$

did not coincide, then it would be possible to recursively construct the following sequence of sets

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

that is such that $\mathcal{V}^n \to \mathcal{S}_{\bar{x}\bar{y}}$ and $\mathcal{U}^n \to \mathcal{S}_{\bar{x}\bar{y}}$. This implies that $f_{\Delta\bar{x}}$ and $\tilde{f}_{\Delta\bar{y}}$ are either everywhere zero, everywhere changing sign, everywhere infinite or everywhere undefined. None of these situations are possible, since the FT of $f_{\Delta\bar{x}}$ and $\tilde{f}_{\Delta\bar{y}}$, respectively, $\alpha(\xi)$ and $\beta(\gamma)$, are everywhere well-defined and nonzero.

Hence the curves in (8) would have to coincide. We can reparametrize the right-hand side curve, letting $y^* = g(x^*)$, to yield $\{(x^* + u, g(x^*)) : x^* \in \mathcal{S}_{x^*}\}$ and we must then have the equality.

$$(\tilde{x}^*, \tilde{g}(\tilde{x}^*) + v) = (x^* + u, g(x^*))$$

implying that

$$\tilde{g}(x^* + u) + v = g(x^*),$$

i.e., $\tilde{g}(\cdot)$ 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.[12] Hence, for any pair of valid models 3 and 4, we must have $\tilde{g}(x^*) = g(x^*)$. The density of x^* can then be determined (up to

[12] 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.

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 $u \in \mathcal{W}_{\Delta\bar{y}}$.

This means that if there are any points where $\tilde{f}_{\Delta\bar{y}}$ or $f_{\Delta\bar{x}}$ are ill-defined, change sign, become zero or are infinite, then Model 3 and 4 are such that $\tilde{g}(x^*) = g(x^*)$ and $\tilde{f}_{\tilde{x}^*}(x^*) = f_{x^*}(x^*)$. So any pair of *distinct* but *observationally equivalent* models must be such that $\tilde{f}_{\Delta\bar{y}}$ and $f_{\Delta\bar{x}}$ are well-defined densities with respect to the Lebesgue measure that are nonzero, finite and never change sign (and are positive, since $\alpha(0) = 1$ and $\beta(0) = 1$). Since $\tilde{f}_{\Delta\bar{y}}$ and $f_{\Delta\bar{x}}$ are supported on $\mathbb{R}$, so must $f_{\bar{x}}$ and $f_{\bar{y}}$, in light of Equation (7). ■

Now, continuing the proof of Theorem 1: Under Model 3, the joint density of $\bar{x}$ and $\bar{y}$ can be written as:

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

where $h(y) \equiv g^{-1}(y)$ (which exists by Assumption 3), while under Model 4, we have

$$f_{\bar{x}\bar{y}}(\bar{x},\bar{y}) = \tilde{f}_{\Delta\bar{y}}(\bar{y} - \tilde{g}(\bar{x}))\, f_{\bar{x}}(\bar{x}) \tag{10}$$

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 (9) and (10):

$$f_{\Delta\bar{x}}(\bar{x} - h(\bar{y}))\, f_{\bar{y}}(\bar{y}) = \tilde{f}_{\Delta\bar{y}}(\bar{y} - \tilde{g}(\bar{x}))\, f_{\bar{x}}(\bar{x}). \tag{11}$$

After rearranging (11) and taking logs, we obtain:

$$\ln \tilde{f}_{\Delta\bar{y}}(\bar{y} - \tilde{g}(\bar{x})) - \ln f_{\Delta\bar{x}}(\bar{x} - h(\bar{y})) = \ln f_{\bar{y}}(\bar{y}) - \ln f_{\bar{x}}(\bar{x}), \tag{12}$$

where these densities are always positive (by Lemma 1), so that the $\ln(\cdot)$ are always well-defined.

We will find necessary conditions for Equation (12) 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.

We use the following Lemma:

Lemma 2 *A twice-continuously differentiable function $c(x,y)$ is such that $\partial^2 c(x,y)/\partial x\partial y = 0\ \forall x,y$ iff it can be written as $c(x,y) = a(x) + b(y)$.*

Proof. We may write

$$c(x,y) = c(0,0) + \int_0^x \frac{\partial c(u,0)}{\partial x} du + \int_0^y \frac{\partial c(x,v)}{\partial y} dv$$

where

$$\frac{\partial c(x,v)}{\partial y} = \frac{\partial c(0,v)}{\partial y} + \int_0^x \frac{\partial^2 c(u,v)}{\partial x \partial y} du = \frac{\partial c(0,y)}{\partial y} + 0$$

if $\partial^2 c(x,y)/\partial x\partial y = 0$. Hence,

$$c(x,y) = \underbrace{c(0,0) + \int_0^x \frac{\partial c(u,0)}{\partial x} du}_{a(x)} + \underbrace{\int_0^y \frac{\partial c(0,v)}{\partial y} dv}_{b(y)}.$$

Conversely,

$$\frac{\partial^2 c(x,y)}{\partial x \partial y} = \frac{\partial^2 a(x)}{\partial x \partial y} + \frac{\partial^2 b(y)}{\partial x \partial y} = 0.$$

■

Note that differentiability of $g(x^*)$, combined with $g'(x^*) \neq 0$ 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 (12) as

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

By Lemma 2, we must then have

$$\begin{aligned} \frac{\partial^2}{\partial\bar{x}\partial\bar{y}} \tilde{F}_{\Delta\bar{y}}(\bar{y} - \tilde{g}(\bar{x})) - \frac{\partial^2}{\partial\bar{x}\partial\bar{y}} F_{\Delta\bar{x}}(\bar{x} - h(\bar{y})) &= 0 \\ \tilde{F}''_{\Delta\bar{y}}(\bar{y} - \tilde{g}(\bar{x}))\,\tilde{g}'(\bar{x}) - F''_{\Delta\bar{x}}(\bar{x} - h(\bar{y}))\,h'(\bar{y}) &= 0 \end{aligned} \tag{13}$$

In the above, we have assumed differentiability of $\tilde{F}_{\Delta\bar{y}}$ and $\tilde{F}_{\Delta\bar{x}}$, but if this fails to hold, we can show that the model is actually identified: The functions $\tilde{g}'(\bar{x})$ and $h'(\bar{y})$ are bounded, continuous and nonzero by Assumption 3. Hence, the points $(\bar{x},\bar{y})$ where $\tilde{F}_{\Delta\bar{y}}(\bar{y} - \tilde{g}(\bar{x}))$ and $\tilde{F}_{\Delta\bar{x}}(\bar{x} - h(\bar{y}))$ and not twice continuously differentiable must coincide. By the same reasoning as in the second part of the proof of Lemma

1, the alternative model would have to be identical to the true model.[13] We can therefore rule out insufficient continuous differentiability for the purpose of finding models that are not identified. To proceed, we need the following Lemma.

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

Proof. Note that, by Lemma 1, $\{(\bar{y} - \tilde{g}(\bar{x}), \bar{x} - h(\bar{y})) : \forall (\bar{x}, \bar{y}) \in \mathbb{R}^2\} = \mathbb{R}^2$. It is therefore possible to vary $\bar{x}$ and $\bar{y}$ so that $\Delta\bar{y} = \bar{y} - \tilde{g}(\bar{x})$ 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 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. ■

Continuing with the proof of Theorem 1, we can rearrange Equation (13) to yield

$$\tilde{F}''_{\Delta\bar{y}}(\bar{y} - \tilde{g}(\bar{x})) = \frac{h'(\bar{y})}{\tilde{g}'(\bar{x})} F''_{\Delta\bar{x}}(\bar{x} - h(\bar{y})), \tag{14}$$

where the ratio $h'(\bar{y}) / \tilde{g}'(\bar{x})$ is nonzero and finite by assumption. Hence if $F''_{\Delta\bar{x}}(\bar{x} - h(\bar{y}))$ is zero, then so is $\tilde{F}''_{\Delta\bar{y}}(\bar{y} - \tilde{g}(\bar{x}))$ and vice versa. If either of those two functions vanishes at a point, by Lemma 3, they must vanish everywhere. It would follows that $\tilde{F}_{\Delta\bar{y}}(\Delta\bar{y})$ and $F_{\Delta\bar{x}}(\Delta\bar{x})$ would be linear and that the corresponding densities $\tilde{f}_{\Delta\bar{y}}(\Delta\bar{y})$ 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 $\tilde{F}''_{\Delta\bar{y}}(\bar{y} - \tilde{g}(\bar{x}))$ vanish at some point is incorrect.

[13] Note that even if a function is nowhere differentiable to some given order, the singularities cannot be fully translation-invariant. Informally, if a derivative is "$+\infty$" at every point, then the function would be infinite everywhere, a situation already ruled out in Lemma 1. Divergence in the derivatives must change sign to maintain the density finite. These changes in derivative sign could be exploited to gain identification as in Lemma 1.

Hence we may assume that $F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))$ and $\tilde{F}''_{\Delta\bar{y}}(\bar{y}-\tilde{g}(\bar{x}))$ do not vanish. Since these functions are continuous, this means they never change sign. Also note that, by assumption, $h'(\bar{y})$ and $\tilde{g}'(\bar{x})$ never change sign or vanish either. We can thus, without loss of generality, rewrite Equation (13) as:

$$\frac{\left|\tilde{F}''_{\Delta\bar{y}}(\bar{y}-\tilde{g}(\bar{x}))\right|}{|F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))|}=\frac{|h'(\bar{y})|}{|\tilde{g}'(\bar{x})|} \tag{15}$$

or

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

Again, since the right-hand side is a difference of functions of $\bar{y}$ and $\bar{x}$, respectively, we must have[14] (by Lemma 2)

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

By the same argument as before, if $\left(\ln\left|\tilde{F}''_{\Delta\bar{y}}(\bar{y}-\tilde{g}(\bar{x}))\right|\right)''=0$ or $(\ln|F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))|)''=0$ at a point then they must vanish everywhere, a situation covered in Case 2 below. (We can also re-use the argument that lack of sufficient continuous differentiability implies identification, hence we can assume sufficient continuous differentiability.)

Case 1 If $\left(\ln\left|\tilde{F}''_{\Delta\bar{y}}(\bar{y}-\tilde{g}(\bar{x}))\right|\right)''$ and $(\ln|F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))|)''$ do not vanish, we may write

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

combined with Equation (15) this implies:

$$\begin{aligned}\frac{\left|\tilde{F}''_{\Delta\bar{y}}(\bar{y}-\tilde{g}(\bar{x}))\right|}{|F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))|} &= \frac{\left|\left(\ln\tilde{F}''_{\Delta\bar{y}}(\bar{y}-\tilde{g}(\bar{x}))\right)''\right|}{\left|(\ln F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y})))''\right|}\\ \frac{\left|\left(\ln\tilde{F}''_{\Delta\bar{y}}(\bar{y}-\tilde{g}(\bar{x}))\right)''\right|}{\left|\tilde{F}''_{\Delta\bar{y}}(\bar{y}-\tilde{g}(\bar{x}))\right|} &= \frac{\left|(\ln F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y})))''\right|}{|F''_{\Delta\bar{x}}(\bar{x}-h(\bar{y}))|}\end{aligned} \tag{16}$$

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

By Lemma 3, 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 $\left|\tilde{F}''_{\Delta\bar{y}}\left(\bar{y}-\tilde{g}\left(\bar{x}\right)\right)\right| / \left|F''_{\Delta\bar{x}}\left(\bar{x}-h\left(\bar{y}\right)\right)\right|$ can actually be written as a ratio of a function of $\bar{y}$ and a function of $\bar{x}$, as required by Equation (15). This will need to be subsequently checked.

We now find densities such that the left-hand (or right-hand) side of Equation (16) is constant. Letting $u=\bar{y}-\tilde{g}\left(\bar{x}\right)$ and $F\left(\cdot\right)\equiv\tilde{F}_{\Delta y}\left(\cdot\right)$ (or similarly, $u=\bar{x}-h\left(\bar{y}\right)$ and $F\left(\cdot\right)\equiv F_{\Delta x}\left(\cdot\right)$), we must have that

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

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.

Lemma 4 *The solution $F\left(u\right)$ to*

$$F''\left(u\right)=-\exp\left(AF\left(u\right)+Bu+C\right) \tag{17}$$

is:

$$F\left(u\right)=-\frac{B}{A}u-\frac{C}{A}+\frac{1}{A}\ln\left(\frac{2D^{2}}{A}\rho\left(D\left(u-u_{0}\right)\right)\right) \tag{18}$$

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.

Proof. This solution can be verified by substitution into the differential equation and 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}. ■

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 is this is valid solution, we first calculate what the implied forms of $\tilde{g}(\bar{x})$ and $h(\bar{y})$ are. From Equation (14), we know that

$$\frac{\tilde{F}''_{\Delta\bar{y}}\left(\bar{y}-\tilde{g}\left(\bar{x}\right)\right)}{F''_{\Delta\bar{x}}\left(\bar{x}-h\left(\bar{y}\right)\right)} = \frac{h'\left(\bar{y}\right)}{\tilde{g}'\left(\bar{x}\right)} \tag{19}$$

where we can find an expression for $F''_{\Delta\bar{x}}(\cdot)$ and $\tilde{F}''_{\Delta\bar{y}}(\cdot)$, generically denoted $F''(\cdot)$ using Equations (17) and (18):

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

The constants D and u_0 may differ for $F''_{\Delta\bar{x}}(\cdot)$ and $\tilde{F}''_{\Delta\bar{y}}(\cdot)$ 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:

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

and note that it cannot be written as a ratio of a function of $\bar{y}$ and a function of $\bar{x}$ (unless $\tilde{g}(\bar{x})$ or $h(\bar{y})$ are constant, a situation ruled out by Assumption 3). Hence Equation (15) 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 3 and 4.

Case 2 We now consider the (so far excluded) case where $\left(\ln F''(u)\right)'' = 0$ for $F =$

$F_{\Delta\bar{x}}$ and $\tilde{F}_{\Delta\bar{y}}$. We have

$$\begin{aligned}
\left(\ln\left|F''\left(u\right)\right|\right)'' &= 0 \\
\left|F''\left(u\right)\right| &= \exp\left(Au+B\right) \qquad (20) \\
F''\left(u\right) &= \pm\exp\left(Au+B\right) \\
F'\left(u\right) &= \pm A^{-1}\exp\left(Au+B\right)+C \\
F\left(u\right) &= -A^{-2}\exp\left(Au+B\right)+Cu+D \qquad (21)
\end{aligned}$$

for some adjustable constants A, B, C, D with $A \neq 0$ (the case $A = 0$ is covered in case 3 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 (21) is of the form

$$f\left(u\right) = \exp\left(-A^{-2}\exp\left(Au+B\right)+Cu+D\right) \qquad (22)$$

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 $h\left(\bar{y}\right)$ and $g\left(\bar{x}\right)$ through relationship (15):

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

Rearranging, we must have

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

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

$$h'\left(\bar{y}\right) = \pm A_{hg}\exp\left(A_{\Delta\bar{x}}h\left(\bar{y}\right)+A_{\Delta\bar{y}}\bar{y}+B_{\Delta\bar{y}}\right) \qquad (23)$$

is

$$h\left(y\right) = -\frac{B_{\Delta\bar{y}}}{A_{\Delta\bar{x}}} - \frac{1}{A_{\Delta\bar{x}}}\ln\left(\pm\frac{A_{\Delta\bar{x}}A_{hg}}{A_{\Delta\bar{y}}}\left(e^{A_{\Delta\bar{y}}\bar{y}}+C_{1\Delta\bar{y}}A_{\Delta\bar{y}}\right)\right), \qquad (24)$$

where $C_{1\Delta\bar{y}}$ is a constant. (This can be shown by substitution of (24) into (23) and by noting that any initial condition $h\left(0\right)$ can be accomodated by adjusting $C_{1\Delta\bar{y}}$.) Similarly,

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

and

$$\tilde{g}\left(\bar{x}\right)=-\frac{B_{\Delta\bar{x}}}{A_{\Delta\bar{y}}}-\frac{1}{A_{\Delta\bar{y}}}\ln\left(\pm\frac{A_{\Delta\bar{y}}A_{hg}}{A_{\Delta\bar{x}}}\left(e^{A_{\Delta\bar{x}}\bar{x}}+C_{1\Delta\bar{x}}A_{\Delta\bar{x}}\right)\right) \tag{25}$$

where $C_{1\Delta\bar{x}}$ is a constant. From Equations (11), (22) (24) and (25), we have

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

$$\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 \\
&\quad\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}-\tilde{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}}\tilde{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 \\
&\quad\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 3 In the special case where $A = 0$ in Equation (20) (not included in Case 2), we let $B_2 = \exp(B)$ and write, for $F = F_{\Delta x}, \tilde{F}_{\Delta y}$:

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

for some constants B_2, C, D (that differ for $F_{\Delta x}$ and $\tilde{F}_{\Delta y}$) to conclude that $f(u)$ is a normal and therefore that $\Delta\bar{x}$ and $\Delta\bar{y}$ are normally distributed. Since under Model 3 the distribution of $\Delta\bar{x}$ is a factor of the distribution of Δx and under model 4 the distribution of $\Delta\bar{y}$ is a factor of the distribution of Δy, we conclude that either Δx must have a normal factor.or Δy must have a normal factor. Next,

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

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

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

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

References

Bonhomme, S., and J.-M. Robin (2006): "Using High-Order Moments to Estimate Linear Independent Factor Models," Working Paper, University College London.

Chesher, A. (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.

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

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 (2004): "Estimation of Nonlinear Models with Measurement Error Using Marginal Information," Working Paper, University of Southern California, Department of Economics.

Hu, Y., and S. M. Schennach (2006): "Identification and estimation of nonclassical nonlinear errors-in-variables models with continuous distributions," Working Paper, University of Chicago.

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

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.

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

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

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.