

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

Xiaohong Chen
Yingyao Hu
Arthur Lewbel

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP18/07

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 July 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^* . 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 . For convenience we will interpret X as a measure of X^* that is contaminated by measurement or misclassification error, but 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. We will require that X and X^* have the same support. Equation (1.1) can 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 $m^* = m(X^*)$. Regardless of whether X is literally a flawed measure of X^* or

just some related covariate, we may still define $X - X^*$ as a measurement error, though in general this error will be *nonclassical*, that is, it will have nonzero mean not be independent of X^* .

Numerous data sets possess nonclassical measurement errors (see, e.g., Bound, Brown and Mathiowetz (2001) for a review), however, to the best of our knowledge, there is no published work that allows for nonparametric identification and estimation of nonparametric regression models with nonclassical measurement errors 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 the conditional density $f_{Y|X^*}$ (or equivalently, the function m and the density of η) just from the observed joint distribution $f_{Y,X}$, while imposing minimal restrictions on the joint distribution f_{X,X^*} .

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

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

All the existing published papers for 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 . 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 $\mathcal{X} = \{1, 2, \dots, J\}$. 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 $1(\cdot)$ is the indicator function that equals one if its argument is true and zero otherwise. This equation and independence of η yield moments

$$E[\exp(itY) | X = x] = E[\exp(it\eta)] \sum_{x^*=1}^J f_{X^*|X}(x^*|x) \exp[im(x^*)t] \quad (1.3)$$

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

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:

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

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.

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

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

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

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

We then have for any real-valued scalar t ,

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

Define

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

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

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

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

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

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

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

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

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

and

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

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

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

where

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

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 $\mathbf{t} = (0, t_2, \dots, t_J)$ 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 $D_k = \text{Diag}(0, d_{k,2}, \dots, d_{k,J})$, 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

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

and

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

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

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

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, $D_m = \text{Diag}(m(1), \dots, m(J))$, and

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

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.

Assumption 2.4 could be replaced by restrictions on f_{X,X^*} instead (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 $f_{Y|X^*}$ and f_{X,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 identify ϕ_η , $m(x^*)$ and f_{X,X^*} in equation (2.2), meaning that all the unknowns in equation (2.1) are nonparametrically identified. 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 be 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 $\mathbf{t} = (0, t_2, \dots, t_J)$ 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 $D_1 = \text{Diag}(0, d_{1,2}, \dots, d_{1,J})$, 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 $t = 0, t_2, \dots, t_J$ 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

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

With assumption 2.2, equation (2.9) simplifies to

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

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 Vandekerckhove (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 2×1 -vector $\mathbf{t} = (0, t)$,

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

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

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

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

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

Also,

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

thus

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

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 $\operatorname{Re} \phi_{Y|X=0}(t) \neq \operatorname{Re} \phi_{Y|X=1}(t)$, $\operatorname{Im} \phi_{Y|X=0}(t) \neq \operatorname{Im} \phi_{Y|X=1}(t)$, $\operatorname{Im} \phi_{Y|X=0}(t) \operatorname{Re} \phi_{Y|X=1}(t) \neq$

$\operatorname{Re} \phi_{Y|X=0}(t) \operatorname{Im} \phi_{Y|X=1}(t)$.

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 $f_{Y|X^*}$ and f_{X,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 $E(\eta^k|X^*) = E(\eta^k)$ 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 $v_j \equiv E[(Y - \mu_j)^2 | X = j]$, $s_j \equiv E[(Y - \mu_j)^3 | X = j]$,

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

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 $f_{Y|X^*}$ and $f_{X^*|X}$. 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

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

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 $m_0(\cdot)$ 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 $\alpha_0 \equiv (f_{01}, f_{02}, f_{03})^T \equiv (f_\eta, f_{X^*|X, W}, m_0)^T$ be the true parameters of interest. Then the observed likelihood of Y given (X, W) (or the likelihood for α_0) is

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

Before we present a sieve ML estimator $\hat{\alpha}$ for α_0 , we need to impose some mild smoothness restrictions on the unknown functions $\alpha_0 \equiv (f_\eta, f_{X^*|X, W}, m_0)^T$. 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 $\xi = (\xi_1, \dots, \xi_d)^T \in \mathbb{R}^d$, $\mathbf{a} = (a_1, \dots, a_d)^T$ be a vector of non-negative integers, and

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

denote the $|\mathbf{a}| = a_1 + \dots + a_d$ -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 $\underline{\gamma}$ -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:

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

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

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

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

and

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

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 $x^*, x \in \mathcal{X} \equiv \{1, \dots, J\}$; (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

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

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 $\hat{\alpha}_n = (\hat{f}_1, \hat{f}_2, \hat{f}_3)^T \in \mathcal{A}_n$ for $\alpha_0 \in \mathcal{A}$ is defined as:

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

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:

$$\begin{aligned}\mathcal{F}_1^n &= \left\{ \sqrt{f_1(\cdot)} = p_1^{k_1, n}(\cdot)^T \beta_1 \in \mathcal{F}_1 \right\}, \\ \mathcal{F}_2^n &= \left\{ \sqrt{f_2(x^*|x, \cdot)} = \sum_{k=1}^J \sum_{j=1}^J I(x^* = k) I(x = j) p_2^{k_2, n}(\cdot)^T \beta_{2, kj} \in \mathcal{F}_2 \right\}, \\ \mathcal{F}_3^n &= \left\{ f_3(x^*, \cdot) = \sum_{k=1}^J I(x^* = k) p_3^{k_3, n}(\cdot)^T \beta_{3, k} \in \mathcal{F}_3 \right\}.\end{aligned}$$

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_β} , 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 $\mathcal{F}_B = \{f_3(x^*, w) = m_0(x^*, w; \beta) : \beta \in B\}$. The sieve MLE becomes

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

3.1 Consistency

The consistency of the sieve MLE $\hat{\alpha}_n$ 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:

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

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 $\sup_{\alpha \in \mathcal{A}_n: \|\alpha - \alpha_0\|_s \leq \delta} |\ell(Z_t; \alpha) - \ell(Z_t; \alpha_0)| \leq \delta^\kappa U(Z_t)$.

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) $k_n \equiv \max\{k_{1,n}, k_{2,n}, k_{3,n}\} \rightarrow \infty$ and $k_n/n \rightarrow 0$.

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

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 $\mathcal{A}_{0s} \equiv \{\alpha \in \mathcal{A} : \|\alpha - \alpha_0\|_s = o(1), \|\alpha\|_s \leq c_0 < c\}$ and $\mathcal{A}_{0sn} \equiv \{\alpha \in \mathcal{A}_n : \|\alpha - \alpha_0\|_s = o(1), \|\alpha\|_s \leq c_0 < c\}$. 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 $\alpha_1, \alpha_2 \in \mathcal{A}_{0s}$, 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

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

and the pathwise second derivative as

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

Define the Fisher metric $\|\cdot\|$ on \mathcal{A}_{0s} as follows: for any $\alpha_1, \alpha_2 \in \mathcal{A}_{0s}$,

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

We show that $\hat{\alpha}_n$ 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 $\sup_{\tilde{\alpha} \in \mathcal{A}_{0s}} \sup_{\alpha \in \mathcal{A}_{0sn}} \left| \frac{d\ell(Z_t; \tilde{\alpha})}{d\alpha} \left[\frac{\alpha - \alpha_0}{\|\alpha - \alpha_0\|_s} \right] \right| \leq U(Z_t)$ 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 $\tilde{\alpha} \in$

\mathcal{A}_{0s} and $\alpha \in \mathcal{A}_{0sn}$, we have

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

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*

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

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

$(\overline{\mathbf{V}}, \|\cdot\|)$ 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 $\rho : \mathcal{A} \rightarrow \mathbb{R}$. 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

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

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

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

(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

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

and

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

Under Theorem 3.2, we have $\|\hat{\alpha}_n - \alpha_0\| = O_P(\delta_n)$ 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 $\lim_{\delta \rightarrow 0} \eta(\delta) = 0$ such that for all $\alpha \in \mathcal{N}_{0n}$,

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

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

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

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 $\sqrt{n}(\rho(\hat{\alpha}_n) - \rho(\alpha_0)) \xrightarrow{d} N(0, \|v^*\|^2)$.*

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 $\hat{\sigma}_n^2$:

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

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

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

Nonclassical EIV without additional information

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 identity 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:

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

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:

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

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 $\hat{\beta}_j$, $j = 1, 2, 3$ 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 $\frac{\partial}{\partial t} |\phi_\eta(0)| = 0$ and $\frac{\partial}{\partial t} a(0) = 0$. we define

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

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

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

Equation (A.1) is equivalent to

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

where

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

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

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

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

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

We then have

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

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

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

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

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

In summary, we have

$$\operatorname{Re}\{\Phi_{Y,X}(\mathbf{t})\} = F_{X,X^*} \times \operatorname{Re}\{\Phi_{Y|X^*}(\mathbf{t})\}, \quad (\text{A.5})$$

$$(\operatorname{Im}\{\Phi_{Y,X}(\mathbf{t})\} + \Upsilon_X) = F_{X,X^*} \times (\operatorname{Im}\{\Phi_{Y|X^*}(\mathbf{t})\} + \Upsilon), \quad (\text{A.6})$$

$$\left(\operatorname{Re}\frac{\partial}{\partial \mathbf{t}}\Phi_{Y,X}(\mathbf{t}) - \Upsilon_{E[Y|X]}\right) = F_{X,X^*} \times \operatorname{Re}\Phi_{m,a}(\mathbf{t}) \times D_{\partial|\phi|}(\mathbf{t}) \quad (\text{A.7})$$

$$- F_{X,X^*} \times (\operatorname{Im}\Phi_{Y|X^*}(\mathbf{t}) + \Upsilon) \times D_{\partial a}(\mathbf{t})$$

$$- F_{X,X^*} \times D_m \times (\operatorname{Im}\Phi_{Y|X^*}(\mathbf{t}) + \Upsilon),$$

$$\operatorname{Im}\frac{\partial}{\partial \mathbf{t}}\Phi_{Y,X}(\mathbf{t}) = F_{X,X^*} \times (\operatorname{Im}\Phi_{m,a}(\mathbf{t}) + \Upsilon) \times D_{\partial|\phi|}(\mathbf{t}) \quad (\text{A.8})$$

$$+ F_{X,X^*} \times \operatorname{Re}\Phi_{Y|X^*}(\mathbf{t}) \times D_{\partial a}(\mathbf{t}) + F_{X,X^*} \times D_m \times \operatorname{Re}\Phi_{Y|X^*}(\mathbf{t}).$$

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^*} , $\text{Re}\{\Phi_{m,a}(\mathbf{t})\}$ and $(\text{Im}\{\Phi_{m,a}(\mathbf{t})\} + \Upsilon)$ 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

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

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

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

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

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

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

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

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

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

where the function $\text{diag}(\cdot)$ 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:

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

Furthermore, equation (A.11) implies that

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

Define a J by 1 vector $e_1 = (1, 0, 0, \dots, 0)^T$. 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 $e_1^T \times C_{\mathbf{t}} \equiv (c_{11}, c_{12}, \dots, c_{1J})$. The vector $\text{diag} (D_{\partial a}(\mathbf{t}))$ is then uniquely determined as follows:

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

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

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

Notice that

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

where

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

Therefore, we have

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

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

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

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 $y \in \mathcal{Y}$.

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

■

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

$$\begin{aligned} m_j &= m(j), \quad \mu_j = E(Y|X = j), \\ v_j &= E \left[(Y - \mu_j)^2 | X = j \right], \quad s_j = E \left[(Y - \mu_j)^3 | X = j \right], \\ p &= f_{X^*|X}(1|0), \quad q = f_{X^*|X}(0|1), \quad f_{Y|X=j}(y) = f_{Y|X}(y|j). \end{aligned}$$

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

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

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

$$\mu_0 = (1-p)m_0 + pm_1 \quad \text{and} \quad \mu_1 = qm_0 + (1-q)m_1.$$

We may solve for p and q as follows:

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

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

$$\begin{aligned} \frac{-p}{1 - p - q} &= \frac{m_0 - \mu_0}{\mu_1 - \mu_0}, & \frac{-q}{1 - p - q} &= \frac{\mu_1 - m_1}{\mu_1 - \mu_0}, \\ \frac{1 - p}{1 - p - q} &= 1 - \frac{-q}{1 - p - q}, & \frac{1 - q}{1 - p - q} &= 1 - \frac{-p}{1 - p - q}, \end{aligned}$$

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 $j = 0, 1$

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

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 $E(\eta^k|X^*) = E(\eta^k)$ for $k = 2, 3$. For $k = 2$, we consider

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

Similarly, we have

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

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

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

That is

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

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:

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

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

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

and

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

We eliminate $E[\eta^2]$ in the two equations above to obtain,

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

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

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

Since $m_1 + m_0 = C_1$, we have

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

that is,

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

Moreover, we have

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

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

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

Finally, we have

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

where

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

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

$$\begin{aligned}
 C_2 &= \frac{3}{2}C_1^2 - 3(\mu_1 + \mu_0)C_1 + 2\frac{\mu_1^3 - \mu_0^3}{\mu_1 - \mu_0} - \frac{s_1 - s_0}{\mu_1 - \mu_0} \\
 &= \frac{3}{2}C_1^2 - 3(\mu_1 + \mu_0)C_1 + 2\frac{\mu_1^3 - \mu_0^3}{\mu_1 - \mu_0} - \frac{\kappa_1 - 3v_1\mu_1 + 2\mu_1^3 - (\kappa_0 - 3v_0\mu_0 + 2\mu_0^3)}{\mu_1 - \mu_0} \\
 &= \frac{3}{2}C_1^2 - 3(\mu_1 + \mu_0)C_1 - \frac{\kappa_1 - 3v_1\mu_1 - (\kappa_0 - 3v_0\mu_0)}{\mu_1 - \mu_0} \\
 &= \frac{3}{2}C_1^2 - 3(\mu_1 + \mu_0)\frac{v_1 - v_0}{\mu_1 - \mu_0} + \frac{3v_1\mu_1 - 3v_0\mu_0}{\mu_1 - \mu_0} - \frac{\kappa_1 - \kappa_0}{\mu_1 - \mu_0} \\
 &= \frac{3}{2}\left(\frac{v_1 - v_0}{\mu_1 - \mu_0}\right)^2 - 3\frac{v_0\mu_1 - v_1\mu_0}{\mu_1 - \mu_0} - \frac{\kappa_1 - \kappa_0}{\mu_1 - \mu_0}
 \end{aligned}$$

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 $f_{Y|X}$. ■

REFERENCES

Ai, C., and Chen, X. (2003), ‘‘Efficient Estimation of Models With Conditional Moment Restriction’’

- tions Containing Unknown Functions," *Econometrica*, 71, 1795-1843.
- Bickel, P. J.; Ritov, Y., 1987, "Efficient estimation in the errors in variables model." *Ann. Statist.* 15, no. 2, 513-540.
- Bordes, L., S. Mottelet, and P. Vandekerckhove, 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.
- 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.
- Hansen, L.P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029-1054.

Nonclassical EIV without additional information

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

Nonclassical EIV without additional information

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

Nonclassical EIV without additional information

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
Misclassification 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

Nonclassical EIV without additional information

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
Misclassification 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

Nonclassical EIV without additional information

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 \hat{f}_η and $\hat{f}_{X|X^*,W}(x|x^*, w)$ for each x and x^* .