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

Xiaohong Chen, Yingyao Hu, and Arthur Lewbel

Yale University, Johns Hopkins University, and Boston College

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 estimation results.

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

1 Introduction

We consider identification and estimation of the nonparametric regression model

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

where Y and X^* are scalars and X^* is not observed. We assume X^* is discrete, 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 $Y = m\left(X^*, W\right) + \eta$, $E[\eta|X^*, W] = 0$,

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 the same support) means that the measurement error $X - X^*$ will be *nonclassical*, in particular, the error will depend on 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 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 are binary indicators of actual versus reported pension type. Hirsch and MacPherson (2003) document misclassification in surveys of union status. 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. 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.

Many estimators and empirical analyses have been proposed to deal with misclassified discrete variables. See, e.g., Chua and Fuller (1987), Bollinger (1996), Lewbel (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 $f_{Y|X^*}$ (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 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 imply 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, a health status or life expectancy quantile for example, and X could be some observed proxy, say 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 $Y = m^* + \eta$ featuring 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 $f_{Y|X^*}$ make one of three assumptions: the measurement error structure $f_{X|X^*}$ belongs to a parametric family; 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 $\{y,x,z\}$; a secondary sample exists to provide information on f_{X,X^*} and permit recovery of $f_{Y|X^*}$ from the observed $f_{Y,X}$ in the primary sample. See Carroll, Ruppert, Stefanski and Crainiceanu (2006) and the references therein for detailed reviews on existing approaches and results.

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 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.

We utilize characteristic functions. Suppose X and X^* have support $\mathcal{X} = \{1, 2, ..., J\}$. Then by (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. This equation, and independence of η, yield moments

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

Evaluating (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 $f_{X,X^*}(x, x^*)$, $m(x^*)$, and $E[\exp(it\eta)]$ 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 $f_{Y|X^*}$.

Estimation could be based directly on (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 solve (1.3); hence GMM estimation using finitely many such t's is not efficient in general. Here we apply instead the method of sieve Maximum Likelihood (ML) of Grenander (1981), 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 (2000) and others, we get consistency and find the convergence rate of the sieve MLE, along with root-n asymptotic normality and semiparametric efficiency of such smooth functionals 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. All proofs are in the Supplement appendix, available at http://www.stat.sinica.edu.tw/statistica/submission.

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 for 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 conditional on both the true X^* and on the measurement error in X, does not depend upon the measurement error.

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 is 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 later show that this assumption can be relaxed in a couple of different ways, e.g., as noted in the introduction, it can be replaced by

$E\left[\exp\left(it\eta\right)|X^*, X\right] = E\left[\exp\left(it\eta\right)\right]$ 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 $E\left(\eta^k|X^*\right) = E\left(\eta^k\right)$ for $k = 2, 3$.

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

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

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

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

We then have for any real-valued scalar t,

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

Define

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

For a real-valued vector $\mathbf{t} = (0, t_2, ..., t_J)$, let $D_{|\phi|}(\mathbf{t}) = Diag\{1, \left|\phi_\eta(t_2)\right|, ..., \left|\phi_\eta(t_J)\right|\}$,

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

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

$$\Phi_{m,a}(\mathbf{t}) = \begin{pmatrix} 1 & \exp\left(it_2m_1 + ia\left(t_2\right)\right) & ... & \exp\left(it_Jm_1 + ia\left(t_J\right)\right) \\ 1 & \exp\left(it_2m_2 + ia\left(t_2\right)\right) & ... & \exp\left(it_Jm_2 + ia\left(t_J\right)\right) \\ ... & ... & ... & ... \\ 1 & \exp\left(it_2m_J + ia\left(t_2\right)\right) & ... & \exp\left(it_Jm_J + ia\left(t_J\right)\right) \end{pmatrix}.$$

With these matrix notations, for any real-valued vector $\mathbf{t}$, (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

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

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

Since 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, we replace (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 $C_{\mathbf{t}} \equiv (\mathrm{Re}\{\Phi_{Y,X}(\mathbf{t})\})^{-1} \times (\mathrm{Im}\{\Phi_{Y,X}(\mathbf{t})\} + \Upsilon_X)$.

ASSUMPTION **2.3.** *(rank). There is a real-valued vector* $\mathbf{t} = (0, t_2, ..., t_J)$ *such that (i)* $\mathrm{Re}\{\Phi_{Y,X}(\mathbf{t})\}$ *and* $(\mathrm{Im}\{\Phi_{Y,X}(\mathbf{t})\} + \Upsilon_X)$ *are invertible, and (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

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

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

Assumption 2.3(ii) is rather complicated, but can be replaced by some simpler sufficient alternatives, which we describe later. Given a candidate value of $\mathbf{t}$, we can test if Assumption 2.3 holds for that value, since the assumption 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. It would also be possible to set up a numerical search for sensible candidate values of $\mathbf{t}$ that one might check. For example, letting $Q(\mathbf{t})$ be an estimate of the product of the squared determinants of the matrices in Assumption 2.3(i), one could search for values of $\mathbf{t}$ that numerically maximize $Q(\mathbf{t})$. Assumption 2.3(i) is then satisfied with high probability if the maximized $Q(\mathbf{t})$ differs significantly from zero. Similarly, one could let $Q(\mathbf{t})$ be the product of the squared differences between the left and right hand sides of each inequality in Assumption 2.8, and maximize that to find values of $\mathbf{t}$ that satisfy this binary rank condition. Note also that estimation does not actually require finding an example value of $\mathbf{t}$.

In the Appendix, we show that

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

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. We may then identify $f_{X|X^*}(\cdot|x^*)$ and $m(x^*)$ under the following.

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 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 $f_X = \sum_{X^*} f_{X|X^*} f_{X^*}$ 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.** *Under Assumptions 2.1, 2.2, 2.3 and 2.4 in (1.1), 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 regressor X^* and 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 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. Let $C_{\mathbf{t}}^T$ be the transpose of $C_{\mathbf{t}}$, and "$\circ$" 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 has $C_{\mathbf{t}} \circ C_{\mathbf{t}}^T + I$ invertible, and all entries in the first row of the matrix $C_{\mathbf{t}}$ 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, 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 (1.3) for some given values of t. More formally, all that is required is that (2.4), and hence (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 use in the estimation, the requirement becomes that (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 for Theorem 2.1 can be simplified in the special case that X^* is a 0-1 dichotomous variable, i.e., $\mathcal{X} = \{0, 1\}$. Define $m_j = m(j)$ for $j = 0, 1$. Given Assumptions 2.1 and 2.2, the relationship between the observed density and the latent ones becomes

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

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) show that all the unknowns in (2.9) 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 (2.9) to obtain identification with few restrictions on the distribution f_η.

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

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

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

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

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

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

Also

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

thus

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

is always invertible. Therefore, in the dichotomous case, Assumptions 2.3 and 2.5 are the same, and can be expressed as the following

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

COROLLARY **1.** *If* $\mathcal{X} = \{0,1\}$*, (1.1) and (2.9) hold with Assumptions 2.8 and 2.9, then the density* $f_{Y,X}$ *uniquely determines* $f_{Y|X^*}$*,* $f_{X|X^*}$*, and* f_{X^*}*.*

3 Sieve Maximum Likelihood Estimation

This section considers the estimation of a nonparametric regression model $Y = m_0(X^*, W) + \eta$, where the function $m_0()$ is unknown, 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. Before we present a sieve ML estimator $\widehat{\alpha}$ for α_0, we need to impose some mild smoothness restrictions on the unknown functions α_0. The sieve method allows for unknown functions belonging to function spaces such as Sobolev, Besov and others; see e.g., Shen and Wong (1994), Wong and Shen (1995), Shen (1997) and Van de Geer (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 $\nabla^{\mathbf{a}} h(\xi) \equiv \partial^{|\mathbf{a}|} h(\xi_1, ..., \xi_d) / \partial \xi_1^{a_1} \cdots \partial \xi_d^{a_d}$ 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 $\underline{\gamma}$-th derivative is Hölder continuous with exponent $\gamma - \underline{\gamma} \in (0, 1]$. Take the Hölder norm as

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

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

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

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

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

We impose the following smoothness restrictions on the densities

Assumption **3.1.** *(i) The assumptions of 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* $x^* \in \mathcal{X}$.

Write $\mathcal{A} = \mathcal{F}_1 \times \mathcal{F}_2 \times \mathcal{F}_3$ 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 $\mathcal{A}$: 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 $\alpha_0 \in \mathcal{A}$: $\widehat{\alpha}_n = \arg\max_{\alpha \in \mathcal{A}_n} \sum_{t=1}^n \ell(Z_t; \alpha)$. For simplicity we present a finite-dimensional sieve $\mathcal{A}_n = \mathcal{F}_1^n \times \mathcal{F}_2^n \times \mathcal{F}_3^n$. 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. The sieve spaces for $\mathcal{F}_j, j = 1, 2, 3$, are

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

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

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

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 up to 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 $\widehat{\alpha}_n = \arg\max_{\alpha\in\mathcal{A}_n} \sum_{t=1}^{n} \ell(Z_t;\alpha)$ with

$$\ell(Z_t;\alpha) = \ln\left\{ \sum_{x^*\in\mathcal{X}} f_1\left(Y_t - m_0(x^*, W_t;\beta)\right) f_2\left(x^*|X_t, W_t\right) \right\}. \tag{3.2}$$

We could let $f_3\left(x^*, w\right) = f_3\left(x^*, w; \beta\right)$ be any flexible semi-nonparametric form; see, e.g., Liang and Wang (2005), Liang, Hardle and Carroll (1999), and Wang (2000).

3.1 Consistency and convergence rate

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

$$\|\alpha\|_s = \sup_{\eta} \left| f_1(\eta)\left(1+\eta^2\right)^{-\zeta/2} \right| + \sup_{x^*,x,w} \left| f_2\left(x^*|x,w\right)\right| + \sup_{x^*,w} \left| f_3\left(x^*,w\right)\right|$$

for some $\zeta > 0$.

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

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) $k_{j,n} \to \infty$ and $k_{j,n}/n \to 0$ for $j = 1, 2, 3$.*

The following consistency lemma is a direct application of Theorem 3.1 (or Remark 3.3) of Chen (2007); we omit its proof.

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

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 $\mathcal{A}_{0sn} \equiv \{\alpha \in \mathcal{A}_n : ||\alpha - \alpha_0||_s = o(1), ||\alpha||_s \leq c_0 < c\}$. For simplicity we assume that both $\mathcal{A}_{0s}$ and $\mathcal{A}_{0sn}$ are convex parameter spaces. Suppose 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, $\alpha_2 \in \mathcal{A}_{0s}$,

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

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 $\max\{(k_{1,n})^{-\gamma_1}, (k_{2,n})^{-\gamma_2/d_w}, (k_{3,n})^{-\gamma_3/d_w}\} = O\left((k_{1,n}+k_{2,n}+k_{3,n})^{-\gamma}\right)$; Assumption 3.5 makes sure that the pseudo metric $\|\alpha-\alpha_0\|$ is well defined on $\mathcal{A}_{0s}$; Assumption 3.6 imposes 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) to the local parameter space $\mathcal{A}_{0s}$ and the local sieve space $\mathcal{A}_{0sn}$; hence we omit its proof.

THEOREM **3.1.** *Under Assumptions 3.1-3.7,* $k_{1,n} = O\left(n^{\gamma/[\gamma_1(2\gamma+1)]}\right)$ *and* $k_{j,n} = O\left(n^{\gamma d_w/[\gamma_j(2\gamma+1)]}\right)$ *for* $j=2,3$, *we have* $\|\widehat{\alpha}_n-\alpha_0\| = O_P\left(n^{-\gamma/(2\gamma+1)}\right)$.

3.2 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

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

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 write

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

whenever the right hand-side limit is well defined.

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 an* $\varepsilon>0$ *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 $\langle v^*, v\rangle = \frac{d\rho(\alpha_0)}{d\alpha}[v]$ for all $v \in \mathbf{V}$, and $||v^*||^2 \equiv \|\frac{d\rho(\alpha_0)}{d\alpha}\|^2$.

Under Theorem 3.1, we have $||\widehat{\alpha}_n - \alpha_0|| = O_P(\delta_n)$ with $\delta_n = n^{\frac{-\gamma}{2\gamma+1}}$. Write $\mathcal{N}_0 = \{\alpha \in \mathcal{A}_{0s} : \|\alpha - \alpha_0\| = O(\delta_n)\}$ 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* $U(Z_t)$ *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 necessary for obtaining the $\sqrt{n}$ convergence of plug-in sieve MLE $\rho(\hat{\alpha}_n)$ to $\rho(\alpha_0)$ and its asymptotic normality; 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 obtain the following

THEOREM **3.2.** *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

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

We conclude the section by mentioning that instead of the sieve MLE method, we could also apply the random sieve MLE or more generalized sieve empirical

likelihood as proposed in Shen, Shi and Wong (1999) and Zhang and Gijbels (2003). This alternative method has the advantage of allowing for non-continuous densities.

4 Simulation

4.1 Moment-based estimation

This subsection applies the identification procedure to a simple nonlinear regression model with simulated data. The latent regression model is $y = 1 + 0.25\,(x^*)^2 + 0.1\,(x^*)^3 + \eta$, where $\eta \sim N(0,1)$ is independent of x^*. The marginal distribution $\Pr(x^*)$ is $\Pr(x^*) = 0.2[1(x^* = 1)+1(x^* = 4)]+0.3[1(x^* = 2)+1(x^* = 3)]$. We present two examples of the misclassification probability matrix $F_{x|x^*}$ in Tables 1-2. Example 1 considers a strictly diagonally dominant matrix $F_{x|x^*}$; see the true value $f_{x|x^*}(\cdot|x^*)$ in Table 1. Example 2 has $F_{x|x^*} = 0.7F_u + 0.3I$, where I is the identify matrix and $F_u = [u_{ij}/\sum_k u_{kj}]_{ij}$ with u_{ij} independently drawn from a uniform distribution on $[0,1]$; see the true value $f_{x|x^*}(\cdot|x^*)$ in Table 2.

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 $\widehat{\phi}_{Y,X=x}(t) = \sum_{j=1}^n \exp(ity_j) \times 1(x_j = x)$. Since it is directly testable, Assumption 2.3 was 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 $\mathbf{t}$ was found. The sample size was 5000 and there are 1000 repetitions. 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 performed well with the simulated data.

4.2 Sieve MLE

This subsection applies the sieve ML procedure to the semiparametric model $Y = \beta_1 W + \beta_2\,(1 - X^*)\,W^2 + \beta_3 + \eta$, 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 $f_{X^*|X,W}$.

Table 1: Simulation results, Example 1

Value of x^*:	1	2	3	4
$m(x^*)$: true value	1.3500	2.8000	5.9500	11.400
$m(x^*)$: mean estimate	1.2984	2.9146	6.0138	11.433
$m(x^*)$: standard error	0.2947	0.3488	0.2999	0.2957
$\Pr(x^*)$: true value	0.2	0.3	0.3	0.2
$\Pr(x^*)$: mean estimate	0.2159	0.2818	0.3040	0.1983
$\Pr(x^*)$: standard error	0.1007	0.2367	0.1741	0.0153
$f_{x\|x^*}(\cdot\|x^*)$: 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
$f_{x\|x^*}(\cdot\|x^*)$: 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
$f_{x\|x^*}(\cdot\|x^*)$: 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
$m(x^*)$: true value	1.3500	2.8000	5.9500	11.400
$m(x^*)$: mean estimate	1.2320	3.1627	6.1642	11.514
$m(x^*)$: standard error	0.4648	0.7580	0.7194	0.6940
$\Pr(x^*)$: true value	0.2	0.3	0.3	0.2
$\Pr(x^*)$: mean estimate	0.2244	0.3094	0.2657	0.2005
$\Pr(x^*)$: standard error	0.1498	0.1992	0.1778	0.0957
$f_{x\|x^*}(\cdot\|x^*)$: 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
$f_{x\|x^*}(\cdot\|x^*)$: 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
$f_{x\|x^*}(\cdot\|x^*)$: 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

We simulated 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 generated the covariate W as $W = (1 - 0.5X^*) \times \nu$, where $\nu \sim N(0,1)$ was independent of X^*. The observed mismeasured X was generated according to: $X = 0$ if $\Phi(\nu) \leq p(X^*)$ and $X = 1$ otherwise, where $p(0) = 0.5$ and $p(1) = 0.3$.

The Monte Carlo simulation consisted of 400 repetitions. In each repetition, we randomly drew 3000 observations of (Y, X, W), and then applied three ML estimators to compute the parameter of interest β. All three estimators assumed that the true density f_η of the regression error was unknown. The first estimator used 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 was 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 the "infeasible MLE" since X_i^* is not observed in practice. The third estimator was the sieve MLE (3.2) presented in Section 3, using the sample $\{Y_i, X_i, W_i\}_{i=1}^n$ and allowing for arbitrary measurement error by assuming $f_{X|X^*,W}$ unknown. In this simulation study, all three estimators were computed by approximating the unknown $\sqrt{f_\eta}$ using the same Hermite polynomial sieve with $k_{1,n} = 3$; for the third estimator (the sieve MLE) we also approximated $\sqrt{f_{X|X^*,W}}$ by another Hermite polynomial sieve with $k_{2,n} = 3$ for each x and x^* value. In applications, the sieve MLE method needs to specify the order of the sieve terms. Our experience is that the estimation of the finite dimensional parameters is not very sensitive to the order of sieves. Of course if one cares about estimation of the nonparametric density function itself, then one could apply the covariance penalty methods suggested in Efron (2004) and Shen and Huang (2006), among others. The Monte Carlo results in Table 3 show that the sieve MLE had 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 performed well in this Monte Carlo simulation.

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

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

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. It may be possible to extend the identification result to continuously distributed nonclassically mismeasured regressors, by replacing many of our matrix-related assumptions and calculations with corresponding linear operators.

Acknowledgment We thank participants at June 2007 North American Summer Meetings of the Econometric Society at Duke for helpful comments. Chen acknowledges support from NSF Grant SES-0631613.

References

Ai, C., and Chen, X. (2003) Efficient estimation of models with conditional moment restrictions containing unknown functions. Econometrica 71, 1795-1843.

Balke, A. and Pearl, J. (1997) Bounds on treatment fffects from studies with imperfect compliance. Journal of the American Statistical Association 92, 1171-1176.

Bollinger, C. R. (1996) Bounding mean regressions when a binary regressor is mismeasured. Journal of Econometrics 73, 387-399.

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

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

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

Chen, X. (2007) Large sample sieve estimation of semi-nonparametric models. In Handbook of Econometrics, Vol. 6B (ed. by J.J. Heckman and E.E. Leamer) North-Holland.

Chua, T. C. and Fuller, W. A. (1987) A model for multinomial response error applied to labor flows. Journal of the American Statistical Association 82, 46-51.

Efron, B. (2004) The estimation of prediction error: covariance penalties and cross-validation. Journal of the American Statistical Association 99, 619-642.

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

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

Gustman, A. L. and Steinmeier, T. L. (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.

Hirsch, B.T. and Macpherson, D. A. (2003) Union membership and coverage database from the current population survey: note. Industrial and Labor Relations Review 56, 349-354.

Hu, Y. (2006) Identification and estimation of nonlinear models with misclassification error using instrumental variables: a general solution. Working Paper, University of Texas at Austin.

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

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

Liang, H., Hardle, W., and Carroll, R. (1999) Estimation in a semiparametric partially linear errors-in-variables model. The Annals of Statistics 27, 1519-1535.

Liang, H., Wang, N. (2005) Partially linear single-index measurement error models. Statist. Sinica 15, 99–116.

Mahajan, A. (2006) Identification and estimation of regression models with misclassification. Econometrica 74, 631-665.

Shen, X. (1997) On methods of sieves and penalization. The Annals of Statistics 25, 2555-2591.

Shen, X. and Huang, H. (2006) Optimal model assessment, selection, and combination. Journal of the American Statistical Association 101, 554-568.

Shen, X., Shi, J., and Wong, W. (1999) Random sieve likelihood and general regression models. Journal of the American Statistical Association 94, 835-846.

Shen, X., and Wong, W. (1994) Convergence rate of sieve estimates. The Annals of Statistics 22, 580-615.

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

Wang, C.Y., (2000) Flexible regression calibration for covariate measurement error with longitudinal surrogate variables. Statist. Sinica 10, 905–921.

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.

Zhang, J. and Gijbels, I. (2003) Sieve empirical likelihood and extensions of the generalized least squares. Scandinavian Journal of Statistics 30, 1-24.

Department of Economics, Yale University, New Haven, CT 06520, USA
E-mail: (xiaohong.chen@yale.edu)
Department of Economics, Johns Hopkins University, Baltimore, MD 21218, USA

E-mail: (yhu@jhu.edu)

Department of Economics, Boston College, Boston, MA 02467, USA

E-mail: (lewbel@bc.edu)