Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables

Yingyao Hu[1]

The University of Texas at Austin

(This version: May 2005)

Abstract

This paper provides a general solution to the problem of identification and estimation of nonlinear models with misclassification error in a general discrete explanatory variable using instrumental variables. The misclassification error is allowed to be correlated with all the explanatory variables in the model. It is not enough to identify the model by simply generalizing the identification in the binary case with a claim that the number of restrictions is no less than that of unknowns. Such a claim requires solving a complicated nonlinear system of equations, which may have multiple solutions. Finding the unique solution of this complicated nonlinear system becomes manageable only when noticing that the problem can be phrased in terms of matrix diagonalization shown in this paper. The solution shows that the latent model can be expressed as an explicit function of directly observed distribution functions. Therefore, the latent model can be nonparametrically identifiable and directly estimable using instrumental variables. The results show that certain monotonicity restrictions on the latent model may lead to its identification with virtually no restrictions on the misclassification probabilities. An alternative identification condition suggests that the nonparametric identification may rely on the belief that people always have a higher probability of telling the truth than of misreporting. The nonparametric identification in this paper directly leads to a $\sqrt{n}$ consistent semiparametric estimator. The Monte Carlo simulation and empirical illustration show that the estimator performs well with a finite sample and real data.

JEL classification: C14, C41.

Keywords: nonlinear errors-in-variables model, instrumental variable, misclassification error.

[1]This paper combines and supercedes two earlier papers previously circulated under the titles "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: a general solution" (first version: August 2004) and "Misclassification Error and Instrumental Variables" (first version: April 2004). Contact information: Department of Economics, University of Texas at Austin, 1 University Station C3100, BRB 1.116, Austin, TX 78712, hu@eco.utexas.edu, http://www.eco.utexas.edu/~hu/.

1 Introduction

Estimation of a general nonlinear model with measurement error in the covariates is a notoriously difficult problem that has received considerable attention in the recent econometrics literature (for example, relevant studies using repeated measurements or instrumental variables include Hausman et al (1991), Wang and Hsiao (1995), Newey (2001), Li (2002), and Schennach (2004a, 2004b)). The measurement error in a continuous variable, such as wage or income, is considered to be continuous, and the error in a discrete variable, such as education, marital status, or union status, is believed to be discrete. The second type of error is also called misclassification error. Most studies on misclassification error focus on the dichotomous case, avoiding clarifying identification and estimation in the general discrete case, while many discrete variables have more than two possible values (see Bound, Brown, and Mathiowetz, 2001, for a survey). Using an instrumental variable, this paper achieves the nonparametric identification of a nonlinear model with a general discrete explanatory variable which is subject to misclassification error. The error is allowed to be correlated with all the explanatory variables in the model. The identification may lead to a semiparametric estimator.

In general, a nonlinear model cannot be identified using misreported data without any further restrictions or additional sample information. Some weak assumptions include the restriction that the misclassification error is independent of the dependent variable conditional on the true value of the variable, and the restriction that the misclassification error is not very large so that the misreported variable can still be positively correlated with the true value. There are more restrictive assumptions, such as the restriction that the misclassification probabilities are independent of other explanatory variables, and therefore, are constants. These assumptions are widely used in relevant studies (for example, Aigner, 1973; Bollinger, 1996; Kane et al., 1999; Mahajan 2003). This paper adopts those weak assumptions and allows the error to be correlated with all the explanatory variables.

The misclassification error in a dichotomous explanatory variable has been analyzed in a few studies. Aigner (1973) and Bollinger (1996) consider the issue of misclassified binary regressors. Freeman (1984) investigates the misclassification error in the union status in

a longitudinal sample. Ramalho (2002) deals with the presence of misclassification in the response variable in choice-based samples. Black, Berger, and Scott (2000) estimate the slope coefficient in a regression model when a secondary measurement is available. Kane, Rouse, and Staiger (1999) and Lewbel (2003) also use instruments to solve misclassification in treatment effect models. Mahajan (2003) provides point estimators for binary choice models with misclassified dichotomous regressors. He uses the sieve method to avoid the specification of misclassification probabilities as functions of the covariates. Mahajan (2004) uses an instrument to identify part of the coefficients in a semiparametric single index model that includes a mismeasured binary regressor. It is not clear how to extend the identification of the existing results in the dichotomous case to the multi-value discrete case. For example, suppose the latent variable has k possible values. The misclassification probability will have $k \times (k-1)$ unknown parameters if the misclassification error is independent of all other variables conditional on the latent variable. Without that independence assumption, there will be $k \times (k-1)$ unknown density functions needed to be identified and estimated. A simple generalization of the identification in the binary case such as in Mahajan (2003, 2004) is to claim that the number of restrictions is no less than that of unknowns. Such a claim is not enough to identify the model because it requires solving a complicated nonlinear system of equations, which may have multiple solutions. Finding the unique solution of this complicated nonlinear system becomes manageable only when noticing that the problem can be phrased in terms of matrix diagonalization shown in this paper. In the general discrete case, Molinari (2004) provides the interval identification of parameters of interest. However, it is not clear when and how the partial identification will become the point identification, which is more feasible in estimation. This paper shows that these density functions are nonparametrically point-identified and directly estimable when an instrumental variable is available.

The additional sample information used in this paper is an instrumental variable, which can also be treated as a secondary measurement of the latent variable (Li, 2002; Schennach, 2004a). Amemiya (1985) shows that IV estimators are generally biased in nonlinear models. Under the assumption that the measurement error vanishes if the sample size increases, Amemiya and Fuller (1988) and Carroll and Stefanski (1990) obtain a consistent IV esti-

mator in nonlinear models. The IV estimator of a polynomial regression model is discussed in Hausman, Ichimura, Newey, and Powell (1991) and Hausman, Newey, and Powell (1995). Buzas (1997) derives an instrumental variable estimator that is approximately consistent for general nonlinear models. Lewbel (1998) shows a consistent estimator for a specially specified latent variable model with instrumental variables and a strong exclusion restriction. Newey (2001) and Schennach (2004b) consider the nonlinear regression model using a prediction equation with instrumental variables independent of the prediction error. Most studies on IV estimators focus on the continuous measurement error, on which certain independence restrictions are imposed. This study considers the measurement error in a discrete explanatory variable so that the error cannot be independent of the latent true value anymore. This paper assumes that the instrumental variable is independent of the dependent variable and the measurement error conditional on all the explanatory variables. These restrictions on the instrumental variable are also widely used in relevant studies such as Kane et al. (1999), Newey (2001), and Schennach (2004a, 2004b).

This study shows that a nonlinear model with misclassification error is nonparametrically identified and directly estimable when instrumental variables are available. One identification condition is that the latent model satisfies certain monotonicity condition, which holds in many popular models. An advantage of this identification condition is that the restrictions on the misclassification probabilities are very weak. An alternative identification condition suggests that the nonparametric identification may rely on the belief that people always have a higher probability of telling the truth than of misreporting. The nonparametric identification provides an explicit expression of the latent model as a function of the observed model, and therefore directly leads to a nonparametric or semiparametric "plug-in" estimator, introduced in Newey and McFadden (1994).

The model considered in this paper contains three variables, y, x^*, and w. The variable y is a dependent variable, x^* is the latent true discrete variable which is subject to measurement error, and w is a vector of other accurately measured independent variables. The misclassification error may be correlated with all the independent variables. Suppose the

conditional density of the dependent variable y on x^* is

$$f_{y|x^*w}(y|x^*, w). \tag{1}$$

In an i.i.d. sample, we observe y, x, and z, where x is a proxy of x^* and z is an instrumental variable satisfying:

Assumption 1: $f_{y|x^*xzw}(y|x^*, x, z, w) = f_{y|x^*w}(y|x^*, w)$.
Assumption 2: $f_{x|x^*zw}(x|x^*, z, w) = f_{x|x^*w}(x|x^*, w)$.

In this paper, we consider general nonlinear models with misclassification errors. As shown in Assumption 2, the misclassification error can be correlated with not only the true latent variable but also with all the other explanatory variables. (A detailed comparison between these assumptions and those in Mahajan's work is in remark 1 in the appendix.) This paper shows that the distribution function of the dependent variable conditional on explanatory variables $f_{y|x^*w}$ can be expressed as a known function of densities of observed distributions, and, therefore, is nonparametrically identified. To be specific, the latent model $f_{y|x^*w}$ with a different value of x^* is the eigenvalue of a matrix composed of the observed density $f_{yx|zw}$. And the matrix of eigenvectors can be a matrix composed of the misclassification probability $f_{x|x^*w}$ or the conditional density of the true value on observables $f_{x^*|zw}$. The estimator in this paper is a simple "plug-in" semiparametric MLE. The parameter of interest is estimated through the maximization of the likelihood whose unknown parts only contain the parameter of interest and unknown density functions, i.e. the "plug-in" part, whose corresponding sample distributions are directly available in the data. This feature makes the estimator highly applicable. With this expression of $f_{y|x^*w}$, we can extend the method in this paper to a GMM framework, such as $E_{y|x^*w}\left[m(y, x^*, w; \theta_0)\right] = 0$.

As discussed before, these assumptions are widely used in the relevant literature. Assumption 1 means that the misclassified variable x and the instrumental variable z do not contain any useful information about the dependent variable y beyond the true value x^* and w. It also implies that the misclassification error in x is independent of the dependent variable y conditional on the true value x^* and w. That means that the measurement error

is nondifferential. As discussed in Bound, Brown, and Mathiowetz (2001, page 3725), the nondifferential assumption is popular but strong. The correlation between the dependent variable y and the measurement error may have two sources. One is the correlation between the measurement error and other observables w, and the other is the correlation between the measurement error and the unobservables, such as regression error in a regression model. Because assumption 1 allows the first type of correlation, i.e., the correlation between the measurement error and other observables w, our assumption is weaker than the one discussed in Bound, Brown, and Mathiowetz (2001). Assumption 2 means that the misclassification error between x^* and x is independent of the instrumental variable z conditional on the true value x^* and w. This assumption is also discussed in Bound, Brown, and Mathiowetz (2001, page 3732) without considering other explanatory variables w. Under a similar assumption, Kane, Rouse, and Staiger (1999) obtain a consistent estimator using GMM methods. The assumption 2 is weaker than those in previous studies because we do not have any parametric specification for the misclassification probabilities, and these probabilities can be functions of other covariates. The results in this paper suggest that such a specification is not necessary because the misclassification probabilities are nonparametrically identified. An important advantage of assumption 2 is that it allows the correlation between the misclassification error and all the explanatory variables. This extension is important because previous studies have shown that such a correlation is significant in the data. For example, Levine (1993) compares the contemporaneous and retrospective reports of labor force status and finds that the rate of underreporting is significant and related to demographic characteristics of the individual.

This paper is organized as follows. Section 2 shows nonparametric identification of the model. Section 3 develops a $\sqrt{n}-$ consistent semiparametric estimator. Section 4 presents Monte Carlo evidence of the finite sample performance of the estimator. Section 5 provides empirical illustrations. Section 6 concludes the paper. The proofs are in the appendix.

2 Identification

This section considers the general case of misclassification error. Suppose x, x^*, and z have the same support which contains a finite number of points $\{1, 2, ..., k\}$. We use the following

notation:

$$\mathbf{F}_{yx|zw} = \begin{pmatrix} f_{yx|zw}(y,1|1,w) & ... & f_{yx|zw}(y,k|1,w) \\ ... & ... & ... \\ f_{yx|zw}(y,1|k,w) & ... & f_{yx|zw}(y,k|k,w) \end{pmatrix},$$

$$\mathbf{F}_{x^*|zw} = \begin{pmatrix} f_{x^*|zw}(1|1,w) & ... & f_{x^*|zw}(k|1,w) \\ ... & ... & ... \\ f_{x^*|zw}(1|k,w) & ... & f_{x^*|zw}(k|k,w) \end{pmatrix},$$

$$\mathbf{F}_{x|x^*w} = \begin{pmatrix} f_{x|x^*w}(1|1,w) & ... & f_{x|x^*w}(k|1,w) \\ ... & ... & ... \\ f_{x|x^*w}(1|k,w) & ... & f_{x|x^*w}(k|k,w) \end{pmatrix},$$

$$\mathbf{F}_{y|x^*w} = \begin{pmatrix} f_{y|x^*w}(y|1,w) & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & f_{y|x^*w}(y|k,w) \end{pmatrix},$$

$$\mathbf{F}_{y|zw} = \begin{pmatrix} f_{y|zw}(y|1,w) & ... & f_{y|zw}(y|k,w) \end{pmatrix}^T.$$

By assumptions 1 and 2 and the law of total probability, we have

Lemma 1 *Suppose that assumptions 1 and 2 are satisfied. Then*

$$f_{yx|zw}(y,x|z,w) = \sum_{x^*} f_{y|x^*w}(y|x^*,w) f_{x|x^*w}(x|x^*,w) f_{x^*|zw}(x^*|z,w), \tag{2}$$

$$f_{x|zw}(x|z,w) = \sum_{x^*} f_{x|x^*w}(x|x^*,w) f_{x^*|zw}(x^*|z,w), \tag{3}$$

and

$$\mathbf{F}_{yx|zw} = \mathbf{F}_{x^*|zw} \times \mathbf{F}_{y|x^*w} \times \mathbf{F}_{x|x^*w}, \tag{4}$$

$$\mathbf{F}_{x|zw} = \mathbf{F}_{x^*|zw} \times \mathbf{F}_{x|x^*w}. \tag{5}$$

Proof. See the appendix. ■

Similarly, from equations

$$f_{y|zw}(y|z,w) = \sum_{x^*} f_{y|x^*w}(y|x^*,w) f_{x^*|zw}(x^*|z,w) \tag{6}$$

we have

$$\mathbf{F}_{y|zw} = \mathbf{F}_{x^*|zw} \times \mathbf{F}_{y|x^*w} \times \mathbf{i} \tag{7}$$

where $\mathbf{i} = \begin{pmatrix} 1 & ... & 1 \end{pmatrix}^T$. In order to identify $\mathbf{F}_{y|x^*w}$, we need the following assumption:

Assumption 2.1: Rank $\left(\mathbf{F}_{x^*|zw}\right) = k$.

The assumption 2.1 has been used in Darolles, Florens, and Renault (2000) and Newey and Powell (2003). Note that this assumption does not necessarily require the instrument to be correlated with x^*. Since the proposed approach makes use of the full joint distribution of z and x^*, it can exploit the presence of any form of statistical dependence between z and x^* to achieve identification. Here we briefly discuss the relationship between the invertibility of $\mathbf{F}_{x^*|zw}$ and the correlation between x^* and z. Let's ignore the variables w for the time being. Given $\Pr(z = j) \neq 0$ for all j, $\mathbf{F}_{x^*|z}$ is invertible if and only if $\mathbf{F}_{x^*z}$ is invertible, where the i-th row and j-th column entry of matrix $\mathbf{F}_{x^*z}$ is the joint probability $\Pr(x^* = x_j^*, z = z_i)$. Without loss of generality, we assume $E(z) = 0$, then $\rho^2_{x^*z} = \frac{[E(x^*z)]^2}{var(z)var(x^*)}$. First, the singularity of $\mathbf{F}_{x^*z}$ does not imply $\rho^2_{x^*z} = 0$. For example, suppose x^* and z have three possible values -1, 0, $1, i.e.$, $x_1^* = z_1 = -1$, $x_2^* = z_2 = 0$, and $x_3^* = z_3 = 1$. Let the joint probability matrix $\mathbf{F}_{x^*z}$ of x^* and z be as follows:

$$\mathbf{F}_{x^*z} = \begin{pmatrix} 1/6 & 1/6 & 0 \\ 1/6 & 1/6 & 0 \\ 0 & 0 & 1/3 \end{pmatrix}.$$

The matrix $\mathbf{F}_{x^*z}$ is singular. But $E(x^*z) = 1/2$, and $E(z) = 0$ so that $\rho_{x^*z} = 0.75$. Although the correlation coefficient is large, the variable z is actually not a good instrument if we only consider that x^* equals -1 or 0. In fact, conditional on $x^* \neq 1$ and $z \neq 1$, x^* is independent

of z. Secondly, $\mathbf{F}_{x^*z}$ may be invertible even if $\rho^2_{x^*z} = 0$. For example, suppose we have

$$\mathbf{F}_{x^*z} = \begin{pmatrix} 1/8 & 1/4 & 0 \\ 0 & 0 & 1/4 \\ 1/4 & 0 & 1/8 \end{pmatrix}$$

in the last example. In this case, $E(x^*z) = 0$ and $E(z) = 0$ so that $\rho^2_{x^*z} = 0$. But $\mathbf{F}_{x^*z}$ is invertible. If we interchange the columns of $\mathbf{F}_{x^*z}$, we may have

$$\begin{pmatrix} 1/4 & 0 & 1/8 \\ 0 & 1/4 & 0 \\ 0 & 1/8 & 1/4 \end{pmatrix}.$$

This matrix is strictly diagonally dominant, which implies that z actually is a good instrument. In fact, the invertibility (or the determinant) of $\mathbf{F}_{x^*z}$ (or $\mathbf{F}_{x^*|z}$) is a better measurement of the validity of an instrument than $|\rho_{x^*z}|$. And the magnitude of correlation coefficient may be misleading in identifying a weak instrument. Since the problem of weak instruments is not the major focus of this paper, we will leave it for future research and assume we have a valid instrument.

From equation (7), we then have

$$\mathbf{F}_{y|x^*w} \times \mathbf{i} = \mathbf{F}^{-1}_{x^*|zw} \times \mathbf{F}_{y|zw}. \tag{8}$$

The next step is to find $\mathbf{F}_{x^*|zw}$. We make the next assumption:

Assumption 2.2: $\mathbf{F}_{x|x^*w}$ *is invertible.*

One sufficient condition to assumption 2.2 is that the matrix $\mathbf{F}_{x|x^*w}$ is strictly diagonally dominant, i.e,.$f_{x|x^*w}(i|i,w) > \sum_{j\neq i} f_{x|x^*w}(j|i,w)$. or $\Pr(x = i|x^* = i,w) > \Pr(x \neq i|x^* = i,w)$. This condition implies that x still contains enough correct information on x^*.

We then find $\mathbf{F}_{x^*|zw}$ through equation (5) as follows:

$$\mathbf{F}_{x^*|zw} = \mathbf{F}_{x|zw} \times \mathbf{F}^{-1}_{x|x^*w}. \tag{9}$$

Plug in the expression of $\mathbf{F}_{x^*|zw}$ into equation (8), and we have

$$\mathbf{F}_{y|x^*w} \times \mathbf{i} = \mathbf{F}_{x|x^*w} \times \mathbf{F}_{x|zw}^{-1} \times \mathbf{F}_{y|zw}. \tag{10}$$

Equation (10) implies that $\mathbf{F}_{y|x^*w}$ is linear in $\mathbf{F}_{x|x^*w}$. Plug in the expression of $\mathbf{F}_{x^*|zw}$ into equation (4), and we have

$$\mathbf{F}_{x|x^*w} \times \mathbf{F}_{x|zw}^{-1} \times \mathbf{F}_{yx|zw} = \mathbf{F}_{y|x^*w} \times \mathbf{F}_{x|x^*w}. \tag{11}$$

This equation implies n^2 restrictions with $n(n+1)$ unknowns, i.e., $\mathbf{F}_{y|x^*w}$ and $\mathbf{F}_{x|x^*w}$. Since $\mathbf{F}_{y|x^*w}$ is linear in $\mathbf{F}_{x|x^*w}$, we have a system of quadratic equations in the misclassification probabilities $\mathbf{F}_{x|x^*w}$. Furthermore, we have additional n restrictions:

$$\mathbf{F}_{x|x^*w} \times \mathbf{i} = \mathbf{i}. \tag{12}$$

Therefore, the unknowns $\mathbf{F}_{y|x^*w}$ and $\mathbf{F}_{x|x^*w}$ are determined by the equations (11) and (12). Solving the system of equations is certainly not an easy task. Moreover, such a system may have multiple solutions.

Fortunately, there is a much easier way to obtain the nonparametric identification and estimation. Moreover, there is no need to directly solve this complicated system of quadratic equations. We define

$$\mathbf{A} := \mathbf{F}_{x|zw}^{-1} \times \mathbf{F}_{yx|zw}. \tag{13}$$

Note the matrix A is observed in the sample. From equation (11), we have

$$\mathbf{F}_{y|x^*w} = \mathbf{F}_{x|x^*w} \times \mathbf{A} \times \mathbf{F}_{x|x^*w}^{-1}. \tag{14}$$

This equation means that the matrix $\mathbf{F}_{y|x^*w}$ is *similar* to the matrix $\mathbf{A}$.[2] In other words, the latent model $\mathbf{F}_{y|x^*w}$ is similar to the observed model described in $\mathbf{A}$. And the misclassification

[2] A k-by-k matrix B is said to be *similar* to a k-by-k matrix A if there exists a nonsingular k-by-k matrix S such that $B = SAS^{-1}$. If A and B are similar, then they have the same eigenvalues. If $B = SAS^{-1}$ and B is a diagonal matrix, then A has a set of k linearly independent eigenvectors and the ith row of S is a left eigenvector of A associated with the ith diagonal entry of B.

probabilities simply consist of the eigenvectors. By the similarity, the two matrixes should have the same eigenvalues. Since $\mathbf{F}_{y|x^*w}$ is diagonal with diagonal elements $f_{y|x^*w}(y|j,w)$, it should be equal to an eigenvalue of the matrix $\mathbf{A}$. And the eigenvector matrix $\mathbf{F}_{x|x^*w}$ is identified up to permutations of the rows.

All that remains is to determine which eigenvalue of $\mathbf{A}$ corresponds to $f_{y|x^*w}(y|j,w)$ for each j. This issue of identification can be summarized in the expression

$$Q\mathbf{F}_{y|x^*w}Q^{-1} = Q\mathbf{F}_{x|x^*w} \times \mathbf{A} \times \left(Q\mathbf{F}_{x|x^*w}\right)^{-1} \tag{15}$$

where Q is an elementary matrix generated by interchanging rows of the identity matrix. The pair $\left(Q\mathbf{F}_{y|x^*w}Q^{-1}, Q\mathbf{F}_{x|x^*w}\right)$ is observationally equivalent to $\left(\mathbf{F}_{y|x^*w}, \mathbf{F}_{x|x^*w}\right)$. And, therefore, further restrictions are needed to identify the model. First, if there exist duplicate eigenvalues, the identification of $\mathbf{F}_{x|x^*w}$ may fail. A sufficient condition to avoid duplicate eigenvalues is that $f_{y|x^*w}(y|i,w) \neq f_{y|x^*w}(y|j,w)$ *for* $i \neq j$. However, it is actually enough to assume:

Assumption 2.3: there exists a function $\omega(y)$ *such that* $E\left[\omega(y)|x^* = i, w\right] \neq E\left[\omega(y)|x^* = j, w\right]$ *for all* $i \neq j$.

The reasoning behind this assumption is as follows: by equation 2, we have

$$\int \omega(y) f_{yx|zw}(y, x|z, w)dy = \sum_{x^*} \left\{E\left[\omega(y)|x^*, w\right]\right\} f_{x|x^*w}(x|x^*, w) f_{x^*|zw}(x^*|z, w);$$

Thus, if we replace $f_{y|x^*w}(y|j,w)$ with $E\left[\omega(y)|x^* = j, w\right]$ in $\mathbf{F}_{y|x^*w}$, and $f_{yx|zw}(y, i|j, w)$ with $\int \omega(y) f_{yx|zw}(y, x|z, w)dy$ in $\mathbf{F}_{yx|zw}$, the eigenvalue-eigenvector decomposition above still holds with $E\left[\omega(y)|x^* = j, w\right]$ as eigenvalues, but the eigenvectors $Q\mathbf{F}_{x|x^*w}$ do not change. Therefore, assumption 2.3 rules out the cases of duplicate eigenvalues, i.e. $E\left[\omega(y)|x^* = i, w\right] = E\left[\omega(y)|x^* = j, w\right]$ for $i \neq j$.

The next step is to find different conditions which lead to identification of the ordering of the eigenvalues or the eigenvectors. First, if we can find a sequence of the eigenvalues, the model is then identified. For example, suppose that $f_{y|x^*w}(y|j,w)$ satisfies

Assumption 2.4: given y and w, the conditional density $f_{y|x^*w}(y|x^*,w)$ is increasing or decreasing in x^*.

This assumption is not as strong as it looks. For example, we consider a binary choice model with a linear index. The conditional density $f_{y|x^*w}(y|x^*,w)$ equals $F(\beta x^* + w\gamma)$ when $y=1$, and $1-F(\beta x^* + w\gamma)$ when $y=0$, where $F(.)$ is a cdf. Assumption 2.4 holds if and only if the sign of β is known. Furthermore, the economic theory may suggest the sign of the coefficient in such an application as a study on the impact of education on labor supply. This fact suggests that the key identification restrictions on the misclassification probabilities in Mahajan (2003) can be simply replaced with $\beta > 0$. Under assumption 2.4, the matrix $\mathbf{F}_{y|x^*w}$ is then not observationally equivalent to $Q\mathbf{F}_{y|x^*w}Q^{-1}$ for any $Q \neq I$. This is because we obtain $\mathbf{F}_{y|x^*w}$ up to permutations of the diagonal entries, and if we know the sequence of the diagonal entries, then $\mathbf{F}_{y|x^*w}$ is totally identified. We define $\lambda_j(\mathbf{A})$ for $j=1,2,...,k$ as the eigenvalues of the matrix $\mathbf{A}$ with $\lambda_1(\mathbf{A}) > \lambda_2(\mathbf{A}) > ... > \lambda_k(\mathbf{A})$. Then the model is nonparametrically identified as follows:

$$f_{y|x^*w}(y|j,w) = \lambda_j(\mathbf{A}). \tag{16}$$

Moreover, it is actually enough to find the sequence of the conditional expectation of some function of y rather than the sequence of the conditional density $f_{y|x^*w}$ itself. We make the following assumption.

Assumption 2.5: there exists a function $\omega(y)$ such that $E[\omega(y)|x^*=j,w] > E[\omega(y)|x^*=j+1,w]$ for all $j=1,2,...,k-1$.

As we mentioned before, $E[\omega(y)|x^*=j,w]$ can also play a role of an eigenvalue. Therefore, if we know a sequence of $E[\omega(y)|x^*=j,w]$, the sequence of the eigenvalues and eigenvectors is fixed and the model is then identified. A straightforward pick of the function $\omega(y)$ is that $\omega(y)=y$. That means the conditional mean of y is monotonic in x^*, which is a reasonable assumption at least in a linear regression model. Other choices of the user-specified function $\omega(y)$ are $\omega(y)=(y-Ey)^2$, $\omega(y)=1(y\leq y_0)$ or $\omega(y)=\delta(y-y_0)$ for some given y_0. What is worth mentioning is that there are no restrictions on the misclassifi-

cation probabilities besides the invertibility of $\mathbf{F}_{x|x^*w}$ in this case. Assumption 2.5 suggests that certain monotonicity restrictions on the latent model may leads to its identification out of the observed model without imposing restrictive assumptions on the misclassification error.

As an alternative, we can impose restrictions on the misclassification matrix $\mathbf{F}_{x|x^*w}$. Note that if we know a sequence of either one of the columns of the matrix $\mathbf{F}_{x|x^*w}$, it is then not observationally equivalent to $Q\mathbf{F}_{x|x^*w}$ for any $Q \neq I$. Without loss of generality, we impose the restriction on the first column as follows:

Assumption 2.6: $\Pr(x = 1|x^* = j, w) > \Pr(x = 1|x^* = j+1, w)$ *for all* $j = 1, 2, ..., k-1$.

Under this assumption, $\mathbf{F}_{x|x^*w}$, and therefore the model $\mathbf{F}_{y|x^*w}$, are identified. The exact expression of $\mathbf{F}_{y|x^*w}$ can be found by diagonalizing $\mathbf{A} = S^{-1}\Lambda S$ with $S \times \mathbf{i} = \mathbf{i}$, then using a matrix Q to find the right QS satisfying assumption 2.6 in

$$\mathbf{A} = (QS)^{-1}\left[Q\Lambda Q^{-1}\right]QS. \tag{17}$$

Then $\mathbf{F}_{y|x^*w}$ just equals the diagonal matrix on the right-hand side, i.e.,

$$\mathbf{F}_{y|x^*w} = Q\Lambda Q^{-1}. \tag{18}$$

The model is therefore identified.

A fourth alternative assumption is that

Assumption 2.7: $\Pr(x = i|x^* = i, w) > \Pr(x = j|x^* = i, w)$ *for* $j \neq i$.

The intuition of assumption 2.7 is that the probability of reporting the correct case is higher than that of reporting either of other cases. The identification procedure is as follows: after an eigenvector, i.e. a row of $Q\mathbf{F}_{x|x^*w}$, is found, we find the largest entry in the row; suppose it is the j-th entry, we then put the eigenvector in the j-th row of the matrix $\mathbf{F}_{x|x^*w}$. Note that assumption 2.7 is obvisouly weaker than the assumption that that $\mathbf{F}_{x|x^*w}$ is strictly diagonally dominant, i.e. $\Pr(x = i|x^* = i, w) > \Pr(x \neq i|x^* = i, w)$ for all $i = 1, 2, ..., k$, which is used in the early version of this paper. Therefore, we do not require the diagonal

entries of $\mathbf{F}_{x|x^*w}$ to be larger than 0.5, i.e. $\Pr(x = i|x^* = i, w) > 0.5$. Assumption 2.7 suggests that if the diagonal entries of matrix $\mathbf{F}_{x|x^*w}$ are the largest in each row, the matrix $\mathbf{F}_{x|x^*w}$ is not observationally equivalent to $Q\mathbf{F}_{x|x^*w}$ for any Q, and, therefore, the model is identified.

The results are summarized as follows:

Theorem 1 *(Nonparametric Identification)*

Suppose that assumptions 1, 2, 2.1, 2.2, 2.3, and one of 2.4-2.7 are satisfied. Then the model $f_{y|x^*w}$ together with $f_{x|x^*w}$ and $f_{x^*|zw}$ is nonparametrically identifiable and directly estimable.

This theorem uses an instrumental variable to shows that these density functions are point-identified and directly estimable. We will show that the point identification leads to a simple "plug-in" semiparametric estimator.

2.1 The dichotomous case

We illustrate the key idea of this paper using the 0-1 dichotomous case with misclassification error. The key is to express the conditional density of y on x, w and z, $f_{yx|wz}(y, x|w, z)$, as a function of $f_{y|x^*w}(y|x^*, w)$ and directly estimable densities. By law of total probability, we have

$$f_{yx|wz}(y, x|w, z) = \sum_{x^*=0,1} f_{y|x^*w}(y|x^*, w) f_{x|x^*w}(x|x^*, w) f_{x^*|wz}(x^*|w, z). \tag{19}$$

First, we can solve for $f(x^*|w, z)$ through

$$f_{y|wz}(y|w, z) = \sum_{x^*=0,1} f_{y|x^*w}(y|x^*, w) f_{x^*|wz}(x^*|w, z). \tag{20}$$

Define $f_{x^*|wz}(1|w, z) := f_{x^*|wz}(x^* = 1|w, z)$. We then have

$$f_{x^*|wz}(1|w, z) = \frac{f_{y|wz}(y|w, z) - f_{y|x^*w}(y|0, w)}{f_{y|x^*w}(y|1, w) - f_{y|x^*w}(y|0, w)}, \tag{21}$$

$$f_{x^*|wz}(0|w, z) = \frac{f_{y|x^*w}(y|1, w) - f_{y|wz}(y|w, z)}{f_{y|x^*w}(y|1, w) - f_{y|x^*w}(y|0, w)}. \tag{22}$$

Second, we solve for the misclassification probability $f_{x|x^*w}(x|x^*, w)$ using the instrumental variable z through

$$f_{x|wz}(x|w,z) = \sum_{x^*=0,1} f_{x|x^*w}(x|x^*,w) f_{x^*|wz}(x^*|w,z). \tag{23}$$

We then have

$$f_{x|x^*w}(x|1,w) = \frac{f_{x|wz}(x|w,1) f_{x^*|wz}(0|w,0) - f_{x|wz}(x|w,0) f_{x^*|wz}(0|w,1)}{f_{x^*|wz}(0|w,0) - f_{x^*|wz}(0|w,1)}, \tag{24}$$

$$f_{x|x^*w}(x|0,w) = \frac{f_{x|wz}(x|w,0) f_{x^*|wz}(1|w,1) - f_{x|wz}(x|w,1) f_{x^*|wz}(1|w,0)}{f_{x^*|wz}(0|w,0) - f_{x^*|wz}(0|w,1)}. \tag{25}$$

Assumption 2.1 does not specify whether the instrument is weak or not. If the instrument is weak, one would expect the difference between $f_{x^*|wz}(x^*|w,0)$ and $f_{x^*|wz}(x^*|w,1)$ is small. Since the difference is in the denominator in equations (24), and (25), the weakness of the instrument may still be a problem in the estimation. Because the major purpose here is to introduce a new estimator, we assume the instrument z is valid. Combining equations (21), (22), (24), and (25), we obtain

$$f_{x|x^*w}(x|1,w) = \frac{1}{A}[f_{y|x^*w}(y|1,w) - B], \tag{26}$$

$$f_{x|x^*w}(x|0,w) = \frac{1}{A}[f_{y|x^*w}(y|0,w) - B], \tag{27}$$

where

$$A = \frac{f_{y|wz}(y|w,1) - f_{y|wz}(y|w,0)}{f_{x|wz}(x|w,1) - f_{x|wz}(x|w,0)},$$

$$B = \frac{f_{y|wz}(y|w,0) f_{x|wz}(x|w,1) - f_{y|wz}(y|w,1) f_{x|wz}(x|w,0)}{f_{x|wz}(x|w,1) - f_{x|wz}(x|w,0)}.$$

Again, the denominator should not be equal to zero by assumption 2.1. Finally, we obtain the expression of $f_{yx|wz}(y,x|w,z)$:

$$f_{yx|wz}(y,x|w,z) = \frac{f_{y|wz}(y|w,z)}{A}\{f_{y|x^*w}(y|1,w) + f_{y|x^*w}(y|0,w) \tag{28}$$

$$-\frac{1}{f_{y|wz}(y|w,z)} f_{y|x^*w}(y|1,w) f_{y|x^*w}(y|0,w) - B\}.$$

This equation holds for $z = 1$ and $z = 0$ so that we can solve for $f_{y|x^*w}$ as follows:

$$f_{y|x^*w}(y|1,w) + f_{y|x^*w}(y|0,w) = C + B, \tag{29}$$

$$f_{y|x^*w}(y|1,w) f_{y|x^*w}(y|0,w) = D, \tag{30}$$

where

$$C = \frac{f_{yx|wz}(y,x|w,1) - f_{yx|wz}(y,x|w,0)}{f_{x|wz}(x|w,1) - f_{x|wz}(x|w,0)},$$

$$D = \frac{f_{y|wz}(y|w,0) f_{yx|wz}(y,x|w,1) - f_{y|wz}(y|w,1) f_{yx|wz}(y,x|w,0)}{f_{x|wz}(x|w,1) - f_{x|wz}(x|w,0)}.$$

Therefore, the density $f_{y|x^*w}$ can be solved as follows:

$$f_{y|x^*w}(y|x^*,w)|_{x^*=0,1} = \frac{1}{2}[(C+B) \pm \sqrt{(C+B)^2 - 4D}]. \tag{31}$$

Obviously, $f_{y|x^*w}(y|1,w)$ and $f_{y|x^*w}(y|0,w)$ are symmetric because the values 0 or 1 are just symbols for two different statuses. The exact solution of $f_{y|x^*w}(y|1,w)$ and $f_{y|x^*w}(y|0,w)$ can also be obtained if we know the sign of $f_{y|x^*w}(y|1,w) - f_{y|x^*w}(y|0,w)$. The sign of $f_{y|x^*w}(y|1,w) - f_{y|x^*w}(y|0,w)$ can also be determined through equation (21) as follows:

$$f_{x^*|wz}(1|w,1) - f_{x^*|wz}(1|w,0) = \frac{f_{y|wz}(y|w,1) - f_{y|wz}(y|w,0)}{f_{y|x^*w}(y|1,w) - f_{y|x^*w}(y|0,w)}. \tag{32}$$

Alternatively, if we assume the misclassification error is not very severe so that x^* and x are still positively correlated conditional on w, then we have $f_{x|x^*w}(1|1,w) > f_{x|x^*w}(1|0,w)$. The solution of $f_{y|x^*w}(y|1,w)$ and $f_{y|x^*w}(y|0,w)$ can be determined from (26-27) as follows:

$$f_{x|x^*w}(1|1,w) - f_{x|x^*w}(1|0,w) = \frac{1}{A}\Big|_{x=1} [f_{y|x^*w}(y|1,w) - f_{y|x^*w}(y|0,w)]. \tag{33}$$

Therefore, the model is identified.

2.2 Identification when $k = 3$

This section shows the nonparametric identification when $k = 3$ by expressing $\mathbf{F}_{y|x^*w}$, $\mathbf{F}_{x|x^*w}$, and $\mathbf{F}_{x^*|zw}$ as explicit functions of $\mathbf{F}_{yx|zw}$ and $\mathbf{F}_{x|zw}$. By definition, we have

$$\mathbf{F}_{yx|zw} = \begin{pmatrix} f_{yx|zw}(y,1|1,w) & f_{yx|zw}(y,2|1,w) & f_{yx|zw}(y,3|1,w) \\ f_{yx|zw}(y,1|2,w) & f_{yx|zw}(y,2|2,w) & f_{yx|zw}(y,3|2,w) \\ f_{yx|zw}(y,1|3,w) & f_{yx|zw}(y,2|3,w) & f_{yx|zw}(y,3|3,w) \end{pmatrix},$$

and

$$\mathbf{F}_{x|zw} = \begin{pmatrix} f_{x|zw}(1|1,w) & f_{x|zw}(2|1,w) & f_{x|zw}(3|1,w) \\ f_{x|zw}(1|2,w) & f_{x|zw}(2|2,w) & f_{x|zw}(3|2,w) \\ f_{x|zw}(1|3,w) & f_{x|zw}(2|3,w) & f_{x|zw}(3|3,w) \end{pmatrix}.$$

First, we solve for the eigenvalues of the matrix $\mathbf{A}$ defined as $\mathbf{A} := \mathbf{F}_{x|zw}^{-1} \times \mathbf{F}_{yx|zw}$. Note that all the density functions in the matrix A are observed in the sample. The characteristic polynomial of the matrix $\mathbf{A}$ is as follows:

$$p(t) = t^3 - \mathrm{tr}(\mathbf{A})t^2 + \mathrm{m}(\mathbf{A})t - \det(\mathbf{A}) \tag{34}$$

where $\mathrm{m}(A)$ is the sum of all the 2-by-2 principal minors of the matrix $\mathbf{A}$. [3] By equation 14, the matrix $\mathbf{A}$ has three real eigenvalues so that the cubic equation $p(t) = 0$ has three different real roots as follows:

$$\begin{aligned} \lambda_1 &= 2\sqrt{-p}\cos\left(\frac{\theta}{3}\right) + \frac{1}{3}\mathrm{tr}(\mathbf{A}), \\ \lambda_2 &= 2\sqrt{-p}\cos\left(\frac{\theta + 2\pi}{3}\right) + \frac{1}{3}\mathrm{tr}(\mathbf{A}), \\ \lambda_3 &= 2\sqrt{-p}\cos\left(\frac{\theta + 4\pi}{3}\right) + \frac{1}{3}\mathrm{tr}(\mathbf{A}), \end{aligned} \tag{35}$$

where

$$p = \frac{3\,\mathrm{m}(\mathbf{A}) - \mathrm{tr}(\mathbf{A})^2}{9},$$

[3] A k-by-k principal submatrix of A is one lying in the same set of k rows and columns, and a k-by-k principal minor is the determinant of such a principal submatrix.

$$q = \frac{-9\,\mathrm{m}(\mathbf{A})\,\mathrm{tr}(\mathbf{A}) + 27\det(\mathbf{A}) + 2\,\mathrm{tr}(\mathbf{A})^3}{54},$$

$$\theta = \cos^{-1}\left(\frac{q}{\sqrt{-p^3}}\right).$$

As shown before, we have

$$\begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} = \begin{pmatrix} f_{y|x^*w}(y|\clubsuit, w) & 0 & 0 \\ 0 & f_{y|x^*w}(y|\heartsuit, w) & 0 \\ 0 & 0 & f_{y|x^*w}(y|\spadesuit, w) \end{pmatrix}, \tag{36}$$

where the set $\{\clubsuit, \heartsuit, \spadesuit\} = \{1, 2, 3\}$. The one-to-one correspondence between the two sets is unknown. If we replace $f_{yx|zw}(y, i|j, w)$ with $\int \omega(y) f_{yx|zw}(y, x|z, w)dy$ in $\mathbf{F}_{yx|zw}$, we have

$$\begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} = \begin{pmatrix} E\left[\omega(y)|x^* = \clubsuit, w\right] & 0 & 0 \\ 0 & E\left[\omega(y)|x^* = \heartsuit, w\right] & 0 \\ 0 & 0 & E\left[\omega(y)|x^* = \spadesuit, w\right] \end{pmatrix}. \tag{37}$$

Second, we need to find the left eigenvector corresponding to each eigenvalue λ_i ($i = 1, 2, 3.$). Define the eigenvector as $v_i = (v_{i1}, v_{i2}, v_{i3})$ satisfying $v_{i1} + v_{i2} + v_{i3} = 1$. By equation 14, each v_i corresponds to a row of the matrix $\mathbf{F}_{x|x^*w}$. So we have

$$\lambda_i \times v_i = v_i \times \mathbf{A}. \tag{38}$$

We then have

$$E_i \times v_i^T = e \tag{39}$$

where $E_i = \begin{pmatrix} \mathbf{A} - \lambda_i I, & \mathbf{i} \end{pmatrix}^T$ and $e = \begin{pmatrix} 0, & 0, & 0, & 1 \end{pmatrix}^T$. Let E_i^+ Moore-Penrose matrix inverse of E_i. [4] The eigenvector v_i can be found as follows:

$$v_i = \left(E_i^+ \times e\right)^T. \tag{40}$$

[4]The Moore-Penrose matrix inverse of a matrix B is B^+ satisfying $BB^+B = B$, $B^+BB^+ = B^+$, $\left(BB^+\right)^T = BB^+$, and $\left(B^+B\right)^T = B^+B$. If B^TB is invertible, then $B^+ = \left(B^TB\right)^{-1}B^T$.

The matrix of the eigenvectors is as follows:

$$V = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \equiv \begin{pmatrix} f_{x|x^*w}(1|\clubsuit, w) & f_{x|x^*w}(2|\clubsuit, w) & f_{x|x^*w}(3|\clubsuit, w) \\ f_{x|x^*w}(1|\heartsuit, w) & f_{x|x^*w}(2|\heartsuit, w) & f_{x|x^*w}(3|\heartsuit, w) \\ f_{x|x^*w}(1|\spadesuit, w) & f_{x|x^*w}(2|\spadesuit, w) & f_{x|x^*w}(3|\spadesuit, w) \end{pmatrix}. \tag{41}$$

We have the set $\{\clubsuit, \heartsuit, \spadesuit\} = \{1, 2, 3\}$. The model is identified if we can find the one-to-one correspondence between the two sets. Each of the assumptions 2.4-2.7 may lead to such a correspondence. For example, by assumption 2.4 or 2.5, the sequence $\lambda_1 > \lambda_2 > \lambda_3$ suggests that the symbol "$\clubsuit$" stands for 1, "$\heartsuit$" for 2, and "$\spadesuit$" for 3. The advantage of assumption 2.4 or 2.5 is that $\mathbf{F}_{x|x^*w}$ is identified without further restrictions, such as 2.6 or 2.7.

We can also obtain the misclassification probability matrix $\mathbf{F}_{x|x^*w}$ by interchanging v_1, v_2, v_3 so that $\mathbf{F}_{x|x^*w}$ satisfies assumptions 2.6 or 2.7. We have

$$\mathbf{F}_{x|x^*w} = Q \times V \tag{42}$$

where the matrix Q is an elementary matrix generated by interchanging rows of the identity matrix. For example, suppose that assumption 2.6 holds, and that the data show that $f_{x|x^*w}(1|\heartsuit, w) > f_{x|x^*w}(1|\clubsuit, w) > f_{x|x^*w}(1|\spadesuit, w)$. We need to interchange v_1 and v_2. So we have

$$Q = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{43}$$

and that the symbol "$\heartsuit$" stands for 1, "$\clubsuit$" for 2, and "$\spadesuit$" for 3. Suppose that assumption 2.7 holds so that there is an entry larger than any other entry in each row and these entries are in different columns, and that the data shows that these entries are $f_{x|x^*w}(3|\clubsuit, w)$, $f_{x|x^*w}(2|\heartsuit, w)$ $f_{x|x^*w}(1|\spadesuit, w)$. Then we need to interchange rows in V such that these entries are on the diagonal. So we have

$$Q = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}. \tag{44}$$

and that symbol "♠" stands for 1, "♡" for 2, and "♣" for 3. Third, $\mathbf{F}_{y|x^*w}$ is achieved by interchanging the eigenvalues on the diagonal correspondingly as follows:

$$\mathbf{F}_{y|x^*w} = Q \times \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} \times Q^{-1}. \tag{45}$$

The last step is to find $\mathbf{F}_{x^*|zw}$ through equation 9 as follows:

$$\mathbf{F}_{x^*|zw} = \mathbf{F}_{x|zw} \times V^{-1} \times Q^{-1}. \tag{46}$$

In summary, we have shown the explicit expression of $f_{y|x^*w}$ together with $f_{x|x^*w}$ and $f_{x^*|zw}$ as functions of $f_{yx|zw}$, i.e.,

$$f_{y|x^*w}(y|x^*, w) = \phi\left(x^*, f_{yxwz}\right), \tag{47}$$

$$f_{x|x^*w}(x|x^*, w) = \varphi\left(x^*, f_{yxwz}\right), \tag{48}$$

$$f_{x^*|zw}(x^*|z, w) = \psi\left(x^*, f_{yxwz}\right). \tag{49}$$

Moreover, these functions ϕ, φ, and ψ are continuous and differentiable. The explicit expression of these functions for a general k is expected to be very complicated. However, it is not necessary to find the explicit expressions in the estimation. Given the current statistical software, it is quite easy to find eigenvalues and eigenvectors numerically using computers. In that sense, one of the major contributions of this paper is to reveal the similarity relationship between the observed model and the latent model. This relationship makes the identification very clear.

3 Estimation

This section considers the estimation of a parametric model in the form of a conditional density function of y as follows:

$$f_{y|x^*w}(y|x^*, w; \theta_0), \tag{50}$$

where $f_{y|x^*w}$ is known up to the unknown parameter θ_0, and x^* is a 0-1 dichotomous variable subject to misclassification error. We observe x as a proxy of x^* with y and w. Define the nuisance parameters as

$$\gamma_0 = (f_{ywz},\ f_{xwz},\ f_{wz})^T$$

From equation (28), we have

$$\begin{aligned} & \ln f_{y|xwz}(y|x,w,z;\theta_0,\gamma_0) \qquad (51) \\ = \ & \ln\left(f_{y|x^*w}(y|1,w;\theta_0) + f_{y|x^*w}(y|0,w;\theta_0) - \frac{f_{wz}(w,z)}{f_{ywz}(y,w,z)} f_{y|x^*w}(y|1,w;\theta_0) f_{y|x^*w}(y|0,w;\theta_0) - B\right) \qquad (52) \\ & + \ln\left(\frac{f_{y|wz}(y,w,z)}{f_{x|wz}(x,w,z)}\frac{1}{A}\right) \end{aligned}$$

with

$$B = \frac{f_{ywz}(y,w,0) f_{xwz}(x,w,1) - f_{ywz}(y,w,1) f_{xwz}(x,w,0)}{f_{xwz}(x,w,1) f_{wz}(w,0) - f_{xwz}(x,w,0) f_{wz}(w,1)}.$$

Note that the second term in the expression of $\ln f_{y|xwz}$ does not play a role in the estimation because it does not contain the unknown parameters θ_0. The nuisance parameter γ_0 can be estimated nonparametrically as follows:

$$\widehat{\gamma} = \left(\widehat{f}_{ywz},\ \widehat{f}_{xwz},\ \widehat{f}_{wz}\right)^T$$

where

$$\widehat{f}_{ywz}(y,w,z) = \frac{1}{n}\sum_{i=1}^{n} I(z_i = z)\left[\frac{1}{h^{r+1}} K\left((\frac{y-y_i}{h}, \frac{w-w_i}{h})^T\right)\right],$$

$$\widehat{f}_{xwz}(x,w,z) = \frac{1}{n}\sum_{i=1}^{n} I(x_i = x) I(z_i = z)\left[\frac{1}{h^{r}} K(\frac{w-w_i}{h})\right].$$

$$\widehat{f}_{wz}(w,z) = \frac{1}{n}\sum_{i=1}^{n} I(z_i = z)\left[\frac{1}{h^{r}} K(\frac{w-w_i}{h})\right].$$

The constant r is the dimension of w. The function $I(.)$ is an indicator function and the function $K(.)$ is a known kernel function with bandwidth h. Since we will have estimated probability densities in the denominator, they must be bounded away from zero in order for the estimator to be well behaved. In the estimation, we use a fixed trimming technique to

avoid the problem.

We then can semiparametrically estimate the unknown parameter of interest, $\theta_0 \in \Theta$, through the density function $f_{y|xwz}$. The semi-parametric MLE is defined as

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \ln f_{y|xwz}(y_i|x_i, w_i, z_i; \theta, \widehat{\gamma}) \tag{53}$$

with $f_{y|xwz}(y_i|, x_i, w_i, z_i; \theta)$ the conditional density in which we replace γ with its nonparametric estimator.

In order to show the consistency of $\widehat{\theta}$, we need the uniform convergence of $\widehat{\gamma}$. The kernel density estimator has been studied extensively. Let $\omega := (y, x, w, z)$. Define the norm $\|.\|$ as

$$\|\widehat{\gamma} - \gamma_0\| = \sup_{\omega\in\mathcal{W}} \left|\widehat{f}_{ywz} - f_{ywz}\right| + \sup_{\omega\in\mathcal{W}} \left|\widehat{f}_{xwz} - f_{xwz}\right| + \sup_{\omega\in\mathcal{W}} \left|\widehat{f}_{wz} - f_{wz}\right|.$$

We have the following results from Newey (1992):

Lemma 2 *Suppose*

1) $\omega \in \mathcal{W}$ *and* $\mathcal{W}$ *is a compact set.*

2) $\gamma_0(\omega)$ *is continuously differentiable to order* d *with bounded derivatives on an open set containing* $\mathcal{W}$.

2) $K(u)$ *is differentiable of order* d*, and the derivatives of order* d *are bounded.* $K(u)$ *is zero outside a bounded set.* $\int_{-\infty}^{\infty} K(x)dx = 1$ *, and there is a positive integer* m *such that for all* $j < m$, $\int_{-\infty}^{\infty} K(u)u^j du = 0$. *And the characteristic function of* K *is absolutely integrable.*

3) $h \to 0$ *and* $nh^r \to \infty$, *as* $n \to \infty$.

Then

$$\|\widehat{\gamma} - \gamma_0\| = O_p\left[(\ln n)^{1/2}\left(nh^{r+2d}\right)^{-1/2} + h^m\right]. \tag{54}$$

The semiparametic estimator in this paper falls into the framework introduced in section 8.3 of Newey and McFadden (1994). The estimator in this paper can be considered as an application of the general semiparametric estimator in their chapter. We will therefore make similar assumptions and will just give a brief discussion if it has been covered in the chapter.

Let $\omega := (y, x, w, z)$. Define the score function as

$$g(\omega, \theta, \gamma) = \nabla_\theta \ln f_{y|xwz}(\omega, \theta, \gamma)$$

To guarantee the consistency of the estimator, we make the following assumptions:

Assumption 4.1: θ_0 *is identifiable in* $f_{y|x^*w}(y|x^*w;\theta)$, $\theta_0 \in \Theta$, *and* Θ *is compact.*

Assumption 4.2: $f_{y|x^*w}(y|x^*w;\theta)$ *is continuously differentiable in* θ, *and* $E\left[f_{y|xwz}(\omega)^2\right] < \infty$.

Assumption 4.3: There are constants $0 < m_0 < m_1 < \infty$ *such that for all* $x^* \in \{0,1\}$, $\omega \in \mathcal{W}$, *and* $\theta \in \Theta$

$$m_0 \leq f_{y|x^*w}(y|x^*, w;\theta) \leq m_1,$$

$$\left|\nabla_\theta f_{y|x^*w}(y|x^*, w;\theta)\right| \leq m_1,$$

$$m_0 \leq f_{wz}(w, z).$$

Assumption 4.4: $\ln n/(nh^{r+2d}) \to 0$ *as* $n \to \infty$.

Theorem 2 *(Consistency) Supposed that Assumptions 4.1-4.4 and the assumptions of lemma 2 and theorem 1 are satisfied. Then*

$$\hat{\theta} \xrightarrow{p} \theta_0.$$

Proof. See the Appendix. ■

To obtain the asymptotic normality of $\hat{\theta}$, we first show the estimated score converges to the true score as follows:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(\omega_i, \theta, \widehat{\gamma}) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[g(\omega_i, \theta, \gamma_0) + \delta(\omega_i)\right] = o_p(1), \tag{55}$$

where

$$\delta(\omega) = v(\omega) - Ev(\omega), \tag{56}$$

$$v(\omega) = v_1(\omega) + v_2(\omega) + v_3(\omega). \tag{57}$$

The correction term $\delta(\omega_i)$ is due to the nonparametric estimation of γ_0. The formula of $v(\omega)$ is from the linearization of $g(\omega, \theta, \gamma)$ with respect to γ. In order to find the explicit expression of the correction term, we define

$$g(\omega, \theta, \gamma_0) = \frac{\nabla_\theta H}{H} \tag{58}$$

with

$$\begin{aligned} H &= f_{y|x^*w}(y|1,w;\theta) + f_{y|x^*w}(y|0,w;\theta) - \frac{f_{wz}(w,z)}{f_{ywz}(y,w,z)} f_{y|x^*w}(y|1,w;\theta) f_{y|x^*w}(y|0,w;\theta) - B \\ \nabla_\theta H &= \nabla_\theta f_{y|x^*w}(y|1,w;\theta) + \nabla_\theta f_{y|x^*w}(y|0,w;\theta) \\ &\quad - \frac{f_{wz}(w,z)}{f_{ywz}(y,w,z)} \left[\nabla_\theta f_{y|x^*w}(y|1,w;\theta) f_{y|x^*w}(y|0,w;\theta) + f_{y|x^*w}(y|1,w;\theta) \nabla_\theta f_{y|x^*w}(y|0,w;\theta) \right] \end{aligned}$$

We use γ_0 rather than γ in $g(\omega, \theta, \gamma_0)$ in order to use $(f_{ywz}, f_{xwz}, f_{wz})$ in the expression, which is easier to understand than the general expression $g(\omega, \theta, \gamma)$ with $\gamma \equiv (\gamma_1, \gamma_2, \gamma_3)^T$. We have the Frechet derivative of g with respect to γ as follows: for $i = 1$, 2, and 3

$$g_{\gamma_i}(\omega, \theta_0, \gamma_0) = \frac{-1}{H^2} \left(\nabla_\theta H \right) \left(H_{\gamma_i} \right) + \frac{1}{H} \left(\nabla_\theta H \right)_{\gamma_i} \tag{59}$$

with

$$\begin{aligned} H_{\gamma_1} &= \frac{f_{wz}(w,z)}{f_{ywz}^2(y,w,z)} f_{y|x^*w}(y|1,w;\theta) f_{y|x^*w}(y|0,w;\theta) \\ &\quad - \frac{(-1)^z f_{xwz}(x,w,1-z)}{f_{xwz}(x,w,1) f_{wz}(w,0) - f_{xwz}(x,w,0) f_{wz}(w,1)}, \end{aligned}$$

$$\left(\nabla_\theta H \right)_{\gamma_1} = \frac{f_{wz}(w,z)}{f_{ywz}^2(y,w,z)} \left[\nabla_\theta f_{y|x^*w}(y|1,w;\theta) f_{y|x^*w}(y|0,w;\theta) + f_{y|x^*w}(y|1,w;\theta) \nabla_\theta f_{y|x^*w}(y|0,w;\theta) \right].$$

$$\begin{aligned} H_{\gamma_2} &= \frac{(-1)^z f_{ywz}(y,w,1-z)}{f_{xwz}(x,w,1) f_{wz}(w,0) - f_{xwz}(x,w,0) f_{wz}(w,1)} \\ &\quad + B \frac{(-1)^z f_{wz}(w,1-z)}{f_{xwz}(x,w,1) f_{wz}(w,0) - f_{xwz}(x,w,0) f_{wz}(w,1)}, \end{aligned}$$

$$\left(\nabla_\theta H \right)_{\gamma_2} = 0,$$

$$
\begin{aligned}
H_{\gamma_3} &= -\frac{1}{f_{ywz}(y,w,z)} f_{y|x^*w}(y|1,w;\theta) f_{y|x^*w}(y|0,w;\theta) \\
&\quad + B \frac{(-1)^z f_{xwz}(x,w,1-z)}{f_{xwz}(x,w,1) f_{wz}(w,0) - f_{xwz}(x,w,0) f_{wz}(w,1)},
\end{aligned}
$$

$$
\left(\nabla_\theta H\right)_{\gamma_3} = -\frac{1}{f_{ywz}(y,w,z)} \left[\nabla_\theta f_{y|x^*w}(y|1,w;\theta) f_{y|x^*w}(y|0,w;\theta) + f_{y|x^*w}(y|1,w;\theta) \nabla_\theta f_{y|x^*w}(y|0,w;\theta)\right].
$$

Therefore, the correction terms are

$$
\begin{aligned}
v_1(\omega) &= E\left[g_{\gamma_1}(\omega,\theta_0,\gamma_0)\middle| y,w,z\right] \\
v_2(\omega) &= E\left[g_{\gamma_2}(\omega,\theta_0,\gamma_0)\middle| x,w,z\right] \\
v_3(\omega) &= E\left[g_{\gamma_3}(\omega,\theta_0,\gamma_0)\middle| w,z\right]
\end{aligned} \tag{60}
$$

This correction term is consistent with the results in Newey (1994). The second term in equation (55) should converge to a normal distribution by the central limit theorem. Therefore, we have the following lemma:

Lemma 3 *Suppose the assumptions of theorem 2 are satisfied,*

1) $f_{yxwz}(\omega)$ *is continuously differentiable of order 5.*

2) $\sqrt{n}h^{2m} \to 0$, *and* $\sqrt{n}\ln n/(nh^{r+2d}) \to 0$ *as* $n \to \infty$.

Then,

$$
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(\omega_i,\theta,\widehat{\gamma}) \xrightarrow{d} N(0,\Omega) \tag{61}
$$

where

$$
\Omega = Var\left[g(\omega,\theta_0,\gamma_0) + \delta(\omega)\right].
$$

Proof. See the appendix. ■

The last step is to show the asymptotic normality of the estimator $\hat{\theta}$ as follows:

Theorem 3 *(Asymptotic Normality)*

Suppose that the assumptions of lemma 3 are satisfied, and

1) $\theta_0 \in interior(\Theta)$.

2) $E\left[\nabla_{\theta\theta} \ln f_{y|xwz}\right]$ *exists and is nonsingular.*

Then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, G_\theta^{-1} \Omega G_\theta^{-1\prime}), \tag{62}$$

where

$$G_\theta = E\left[\nabla_\theta g(\omega, \theta_0, \gamma_0)\right].$$

Proof. See the appendix. ■

In the general discrete case, the expression of $f_{y|xwz}(y|x, w, z; \theta_0, \gamma_0)$ is obviously more complicated than in the dichotomous case. Therefore, we propose a semiparametric GMM, where the nonparametric part is $f_{x^*|zw}$, i.e., the eigenvectors in the eigenvalue-eigenvector decomposition. The nonparametric identification shown above can lead to a semiparametric estimator because the density functions $f_{y|x^*w}$, $f_{x|x^*w}$, and $f_{x^*|zw}$ can be explicitly expressed as functions of the observed density f_{yxwz}. Consider a parametric GMM model as follows:

$$E(y|x^*, w) = m^*(x^*, w; \theta_0) \tag{63}$$

where m^* is a known moment function and θ_0 is the true value of the unknown parameter of interest. In the sample, we observe y, x, and w. Since x^* is not observed in the sample, we estimate the parameter θ_0 through an observed moment

$$E(y|z, w) = m(z, w; \theta_0) \tag{64}$$

with

$$m(z, w; \theta_0) = \sum_{x^*} m^*(x^*, w; \theta_0) f_{x^*|zw}(x^*|z, w). \tag{65}$$

From the results in nonparametric identification, we can express the unknown densities $f_{x^*|zw}$ as follows:

$$f_{x^*|zw}(x^*|z, w) = \psi\left(x^*, f_{yxwz}\right) \tag{66}$$

with a known function ψ. In general, the function ψ is a complicated but specific algebraic function.

First, we show that the parameter θ_0 is identifiable in the observed moment condition $E(y|z, w) = m(z, w; \theta_0)$, if and only if the parameter θ_0 is identifiable in the model

$E(y|x^*, w) = m^*(x^*, w; \theta_0)$. We define

$$\begin{aligned}
\mathbf{M}(\theta) &= \left(\, m(z,w;\theta)|_{z=1} \,, \quad m(z,w;\theta)|_{z=2} \,, \quad ..., \quad m(z,w;\theta)|_{z=k} \, \right)^T, \\
\mathbf{M}^*(\theta) &= \left(\, m^*(x^*,w;\theta)|_{x^*=1} \,, \quad m^*(x^*,w;\theta)|_{x^*=2} \,, \quad ..., \quad m^*(x^*,w;\theta)|_{x^*=k} \, \right)^T.
\end{aligned}$$

Therefore, we have

$$\mathbf{M}(\theta) = \mathbf{F}_{x^*|zw} \times \mathbf{M}^*(\theta) \tag{67}$$

Suppose θ_0 is not identifiable so that there exists θ_1, which is observationally equivalent to θ_0, i.e., $\mathbf{M}(\theta_1) - \mathbf{M}(\theta_0) = 0$. Since $\mathbf{F}_{x^*|zw}$ is identified and has rank k, we must have $\mathbf{M}^*(\theta_1) - \mathbf{M}^*(\theta_0) = 0$. Thus, the parameter θ_0 is not identified in the latent model if it is not identified in the observed model. In other words, the parameter θ_0 is identified in the observed model if it is identified in the latent model. It is obvious that the parameter θ_0 is not identified in the observed model if it is not identified in the latent model. The parametric identification is summarized as follows:

Theorem 4 *(Parametric Identification)*

Suppose that $\mathbf{F}_{x^*|zw}$ is identified and has rank k. The parameter θ_0 is identifiable in the observed model (64) if and only if it is identifiable in the latent model (63).

Second, we estimate the parameter of interest through a semiparametric model. By equations 65 and 66, the observed model can be written as follows:

$$E(y|z, w) = m(z, w; \theta_0, f_0) \tag{68}$$

with the nuisance function $f_0 = f_{yxwz}(y, x, w, z)$. Since the function ψ in equation 66 is known, the moment function $m(z, w; \theta_0, f_0)$ is known up to the parameter θ_0 and the nuisance function f_0. Although the function f_0 is unknown, we observe the joint distribution of $\{y, x, w, z\}$ in the sample. Therefore, the nuisance parameter f_0 can be estimated nonparametrically as follows:

$$\widehat{f}(y, x, w, z) = \frac{1}{n} \sum_{i=1}^{n} I(x_i = x) I(z_i = z) \left[\frac{1}{h^r} K \left((\frac{y - y_i}{h}, \frac{w - w_i}{h})^T \right) \right]. \tag{69}$$

The constant r is the dimension of w and y. The function $I(.)$ is an indicator function and the function $K(.)$ is a known kernel function with bandwidth h. If the dependent variable y is discrete, we should apply the indicator function to y rather than the kernel function, and r is just the dimension of w. We then can estimate the unknown parameter of interest, $\theta_0 \in \Theta$, through a "plug-in" semiparametric estimator in which we replace f_{yxwz} with its nonparametric estimators. The semiparametric estimator $\widehat{\theta}$ is defined as follows:

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \left[y_i - m(z_i, w_i; \theta, \widehat{f}) \right]^2. \tag{70}$$

The consistency and asymptotic normality of the GMM estimator can be shown in the same way as the semiparametric MLE in the dichotomous case. The major difference between the two cases lies in the expression of the function $\psi\left(x^*, f_{yxwz}\right)$. As shown in the identification, the computation of the function ψ is as easy as finding eigenvectors. The exact proof of the consistency and asymptotic normality requires analyses of the explicit expression of the function ψ, which is tedious but still feasible for a given k. Another issue in the estimation is that the estimated eigenvalues and eigenvectors may not always be real. This problem is not unique in this study. For example, a kernel density estimator with a higher order kernel may not always be positive, and therefore, an estimation problem involving the square root of the estimated density may have the problem of dealing with complex numbers. Another example may be a density estimator involving empirical characteristic functions. A popular solution to this problem is to take the real part of the estimator. Since the identification shows that all the latent densities are real and positive, the probability of encountering a complex value should go to zero as the sample size goes to infinity.

4 Simulation

This section first applies the method developed above to a probit model with a mismeasured 0-1 dichotomous explanatory variable. The conditional density function of the probit model

is

$$\begin{aligned} f^*(y|x^*, w; \theta) &= P(y, x^*, w; \theta)^y (1 - P(y, x^*, w; \theta))^{1-y} \\ P(y, x^*, w; \theta) &= \Phi(\beta_0 + \beta_1 x^* + \beta_2 w) \end{aligned} \tag{71}$$

where $\theta = (\beta_0, \beta_1, \beta_2)'$ and Φ is the standard normal cdf. Three estimators are considered: The first is the ML probit estimator that uses mismeasured covariate x in the primary sample as if it were accurate, i.e., it ignores the measurement error. The MLE is not consistent. The second estimator is the infeasible ML probit estimator that uses the latent true x^* as covariate. This estimator is consistent and has the smallest asymptotic variance of all estimators that we consider. The conditional density function is $f^*(y|x^*, w; \theta)$. The third estimator is the semi-parametric MLE developed above that uses the instrumental variable. For each estimator, we report Root Mean Squared Error (RMSE), the average bias of estimates, and the standard deviation of the estimates over the replications. We should expect that the second estimator has the smallest mean squared error (MSE), that the first one has the largest MSE, and that the MSE of the semiparametric IV estimator is between those of the other two estimators. Since the first estimator is biased due to the misclassification error, the bias of the first estimator should dominate its MSE. The semiparametric IV estimator fixes the bias due to the misclassification error, but the semiparametric estimation should cause a larger variance than the variance of the second estimator. Since the last two estimators are consistent, their variance should dominate their MSE.

Table 1 shows that the MLE which ignores the misclassification error is significantly biased as expected. The bias of the coefficient of the mismeasured independent variable is larger than the bias of the coefficient of the other covariate or the constant. The small-sample biases in the new semi-parametric MLE are similar to those of the other consistent estimator. In all cases the MSE of the infeasible MLE is much smaller than that of the other consistent estimators. The loss of precision is associated with the fact that x^* is not observed, but that we must integrate with respect to its distribution given x, w and z. As the misclassification probability changes, the semi-parametric MLE performs well in each case.

In the general discrete case, we consider a nonlinear regression model as follows:

$$y = e^{-(\beta_0+\beta_1 x^*+\beta_2 w)} + u. \tag{72}$$

The variables w and u have a standard normal distribution. The true value of the parameters are $\beta_0 = -2$, $\beta_1 = 1$, and $\beta_2 = 1$. The latent discrete variable x^*, the misclassified variable x, and the instrumental variable z may have four values, 1, 2, 3, and 4. The distribution of x^* and z are $P_{x^*} = (.2, .3, .3, .2)$, $P_z = (.3, .2, .3, .2)$, where $P_v := (P(v = 1), P(v = 2), P(v = 3), P(v = 4))$. The variable x^* is generated by assigning different discrete values to different ranges of the support of a uniformly distributed η_{x^*} on $[0, 1]$ as follows:

$$x^* = P_{x^*}(\eta_{x^*}) \equiv \begin{cases} 1 & if\ \eta_{x^*} \leq P(x^* = 1) \\ 2 & if\ P(x^* = 1) < \eta_{x^*} \leq P(x^* \leq 2) \\ 3 & if\ P(x^* \leq 2) < \eta_{x^*} \leq P(x^* \leq 3) \\ 4 & if\ P(x^* \leq 3) < \eta_{x^*} \leq P(x^* \leq 4) \end{cases}. \tag{73}$$

We abuse the notation of P_{x^*} to write x^* as a function of the random variable η_{x^*} as follows: $x^* = P_{x^*}(\eta_{x^*})$. Similarly, we define $z = P_z(.6\eta_{x^*} + .4\eta_z)$, where η_z is another independent random variable with a uniform distribution on $[0, 1]$. The correlation between x^* and z is generated by the common random variable η_{x^*}.

We consider three cases of $\mathbf{F}_{x|x^*w}$. The first case presents the constant misclassification probabilities as follows:

$$\mathbf{F}_{x|x^*w} = \mathbf{F}_{x|x^*} = \begin{pmatrix} .6 & .2 & .1 & .1 \\ .2 & .6 & .1 & .1 \\ .1 & .1 & .7 & .1 \\ .1 & .1 & .1 & .7 \end{pmatrix}. \tag{74}$$

The matrix $\mathbf{F}_{x|x^*}$ is defined as the matrix $\mathbf{F}_{x|x^*w}$ with $f_{x|x^*w} = P(x = 1|x^*)$ in each entry. The matrix $\mathbf{F}_{x|x^*w}$ is strictly diagonally dominant so that the model is identified according to Theorem (1). For a given x^*, the value of x is determined by the corresponding row in $\mathbf{F}_{x|x^*}$ and another independent random variable, η_x, with a uniform distribution on $[0, 1]$ in

the same way as in the generation of x^*.

$$x = \mathbf{F}_{x|x^*}(\eta_x) \equiv \begin{cases} 1 & if\ \eta_x \leq P(x=1|x^*) \\ 2 & if\ P(x=1|x^*) < \eta_x \leq P(x \leq 2|x^*) \\ 3 & if\ P(x \leq 2|x^*) < \eta_x \leq P(x \leq 3|x^*) \\ 4 & if\ P(x \leq 3|x^*) < \eta_x \leq P(x \leq 4|x^*) \end{cases} . \tag{75}$$

We abuse the notation again to write x as a function of η_x as follows: $x = \mathbf{F}_{x|x^*}(\eta_x)$. In the other two cases, we consider the correlation between the misclassification error and the other explanatory variable w as follows: $x = \mathbf{F}_{x|x^*}(.9\eta_x + .1\Phi(w))$ and $x = \mathbf{F}_{x|x^*}(.9\eta_x + .1(1-\Phi(w)))$, where Φ is the cumulative distribution function of w.

With each drawn sample, the simulation results (Table 2) contain three estimators similar to those in Table 1. The first estimator is a nonlinear least squared estimator using the misclassified variable x as if it were x^*. That means that we ignore the misclassification error in the first estimation. The second one uses the accurate data without misclassification errors. The last estimator is the semiparametric IV estimate developed in this paper. As expected, the simulation results (Table 2) show that the first estimator ignoring misclassification errors has the larger mean squared error (MSE) than the second estimator using accurate data, because the misclassification errors cause significant biases. The third estimator has a smaller MSE then the first estimator. Moreover, the developed estimator effectively reduces the bias. An interesting fact is that the first estimator with misclassification error ignored has a smaller MSE when the misclassification error is correlated with other explanatory variables w. This may be because the variation of the covariate helps identify part of the misclassification error. The semiparametric IV estimator also performs better when the misclassification error is correlated with other explanatory variables. This is because the semiparametric IV estimator treats the misclassification probabilities as fully nonparametric functions of other explanatory variables. In the situation where the misclassification probabilities are just constants, one should expect certain efficiency loss due to the nonparametric estimation of the misclassification probabilities. Another fact is that the misclassification error not only causes a large bias in the estimated coefficient $\widehat{\beta}_1$ of the latent variable x^*, but also leads to a significant bias in the estimated constant term $\widehat{\beta}_0$ in the first estimation. The simulation

result shows that the semiparametric IV estimator also reduces the bias in $\widehat{\beta}_0$.

In summary, the semiparametric IV estimator proposed in this paper performs well in the finite sample. The new estimator successfully fixes the bias problem due to the misclassification error. And the simulation results are also consistent with the asymptotic properties of the estimator.

5 Empirical illustrations

5.1 The dichotomous case

In the empirical illustration, we apply the estimator to a probit model of labor supply, which investigate the impact of education on women's labor supply. The population considered are all the women at the age from 18 to 65 who still lives with their parents and have left the school. The parents' education level is used as the instrumental variable in the model. The education level is treated as a dichotomous variable, which equals zero if an individual finished high school education or less. If a sample contains the information on parents' education or other instrumental variables, we can consider a lager population. The dependent variable is a dichotomous indicator of employment status. The independent variable contains education, age, work experience, and race. The marital status is not contained in the model because only less than 2 percent of 1457 women are married in the sample. The data are from the March supplement of the 2002 Current Population Survey (CPS). The joint distribution of women and their parents' education level is shown in Table 3. Table 4 contains the descriptive statistics of other variables.

The developed estimator uses parents' education level as an instrument, and allows the misclassification error to be correlated with other explanatory variables, such as age, work experience, and race. The estimation results in Table 5 contain two estimates. The second and the third columns contain the estimate ignoring misclassification error. And the estimates of the developed method is shown in the last two columns. The asymptotic standard error of the new estimator is estimated through equation 8.18 in Newey and McFadden (1994). The results show that the impact of education on women's labor supply is much

larger than in the case that the misclassification error is ignored.

5.2 The general discrete case

This section applies the estimator to a count data model to investigate the impact of education on women's fertility. Since the dependent variable, i.e., number of children, takes on the value zero for a nontrivial fraction of the population, it is better to directly model the expectation of the dependent variable y conditional on the explanatory variables x^* and w. A detailed discussion on a count data model can be found in Wooldridge (2002, page 645). We use a popular functional form, i.e., the exponential function, as follows:

$$m^*(x^*, w) = e^{\beta_0 + \beta_1 x^* + \beta_2 w}. \tag{76}$$

The coefficient β_1 or β_2 is related to the semielasticity of $E(y|x^*, w)$ with respect to x^* or w. For small changes Δx^*, the percentage change in the conditional mean $E(y|x^*, w)$ is roughly $100\beta_1 \Delta x^*$. Since the true education level for each individual is not observed, we use the parents' education level as the instrumental variable to estimate the parameter of interest $\beta' s$. As introduced above, we estimate the parameters through the moment condition

$$m(z, w) = \sum_{x^*} e^{\beta_0 + \beta_1 x^* + \beta_2 w} f_{x^*|zw}(x^*|z, w). \tag{77}$$

One can certainly apply a Nonlinear Least Squares (NLS) estimator to this count data model. The NLS estimator is consistent but inefficient because the discrete distribution of the count data implies heteroskedasticity. We will instead use the Poisson regression model, which is the most popular model for count data. When the distribution of the dependent variable conditional on the independent variables $f_{y|x^*w}$ is Poisson, the estimator using the latent model $f_{y|x^*w}$ with conditional mean $m^*(x^*, w)$ is just a maximum likelihood estimator with the log likelihood for observation i as follows:

$$l_i^*(y_i, x_i^*, w_i; \beta) = y_i \ln m^*(x_i^*, w_i; \beta) - m^*(x_i^*, w_i; \beta). \tag{78}$$

In fact, the Poisson assumption is not necessary for consistent estimation of the parameters. When the distribution $f_{y|x^*w}$ is not Poisson, the same estimator is called the Poisson quasi-maximum likelihood estimator (QMLE), which is fully robust to distributional misspecification. Here we have two possible estimators. First, we ignore the measurement error and use x_i as x_i^* in the likelihood $l_i(y_i, x_i^*, w_i; \beta)$. Second, we use the $m(z, w)$ as a conditional mean of y on z and w, and use a QMLE with a likelihood function as follows:

$$l_i(y_i, z_i, w_i; \beta) = y_i \ln m(z_i, w_i; \beta) - m(z_i, w_i; \beta). \tag{79}$$

Under the regularity conditions, this QMLE estimator is consistent and asymptotically normal.

The population considered is composed of women who have left school and who still live with their parents, which is a little larger than the population in the dichotomous case. We still use the parents' education level as the instrumental variable in the model. If a sample contains information on their parents' education or other instrumental variables, we can consider a larger population. The dependent variable is the number of children that a woman has. The independent variable consists of education, age, employment status, and race. The sample is from the March supplement to the 2002 Current Population Survey (CPS). In this estimation, we have three categories of education: high school education or lower, some college education, and college education or higher. The number of years of education assigned to each category is 9, 14, and 16 respectively. The joint distribution of women and their parents' education level is shown in Table 6. More than half of the women in the sample do not have any college education. And 26.5% of the individuals in the sample entered college but did not finish. The correlation coefficient between the education levels of the women and their parents is 0.256. Table 7 contains the descriptive statistics of other variables. There are 53% of the women having no children, 27% having one child, and 20% having two or more. About 80 percent of the women in the sample are employed. About 20 percent are black. The median age is about 22, and the first and the third quartiles are 19 and 25. Marital status is not considered in the model because less than 1.6 percent of the 1688 women in the sample are married.

We assume the misclassification error in the education level of an individual is independent of the education level of this individual's parents and the number of children of this individual conditional on this individual's education level, employment status, age, and race. And the misclassification probability is assumed to satisfy assumptions 2.2 and 2.7. This assumption means that people are more willing to tell the truth than to lie conditional on their education, employment status, age, and race. The advantage of the estimator developed in this paper is that it allows the measurement error in education to be correlated with all the explanatory variables: the true education level, age, employment status, and race. For example, individuals at different ages may have different probabilities of misreporting their level of education in reality. Suppose the error is independent of other explanatory variables except age and education. At each age level, the misclassification probability contains six unknown parameters. If age is considered to be continuous, there are six unknown density functions in the misclassification probability matrix. If we include all the other explanatory variables, the six unknown functions will have multiple arguments. Without imposing further restrictions, it is not clear how to use the existing methods to identify and estimate these functions. Using the method in this paper, we can nonparametrically identify these unknown functions and parameters of interest, and use a simple "plug-in" semiparametric estimator to estimate these unknown densities and parameters.

Table 8 contains two NLS estimates, and Table 9 shows the two QMLE estimates. The second and third columns of the tables contain the estimates of ignoring misclassification error. The estimates of the developed method are shown in the last two columns. When the measurement error is ignored, both the NLS estimator and the QMLE estimator are inconsistent, and the NLS estimator also has the problem of heteroskedasticity. When an instrumental variable is used in the current method, the NLS and QMLE estimators both are consistent, and the QMLE should have the correct estimate of asymptotic standard deviation. The most interesting parameter in the model is the coefficient on education. When the measurement error is ignored, the estimate is biased toward zero compared with the estimate using the instrumental variable. The results show that the impact of education on women's fertility is more significant than commonly thought. If we ignore the measurement error, the QMLE estimate suggests that one more year of education will lead to a 2.6%

decrease in the number of children born. But the new estimator shows that this percentage change is underestimated: There is a 5.4% decrease in the number of children born for one more year of education. And the effect is more significant than in the case where the measurement error is ignored. Employment status and race do not have a significant impact on women's fertility in any of the four estimates. The impact of age on women's fertility is very significant, as expected.

6 Conclusion

This paper provides a general solution to the problem of identification and estimation of nonlinear models with misclassification error when instrumental variables are available. The misclassification error can be correlated with all the explanatory variables. The results show that certain monotonicity restrictions on the latent model may lead to its identification with virtually no restrictions on the misclassification probabilities. An alternative identification condition implies that the nonparametric identification may rely on the belief that people always have a higher probability of telling the truth than of misreporting. The results show that the latent model is nonparametrically identified and directly estimable. The nonparametric identification in this paper directly leads to a nonparametric or semiparametric estimator, which does not require estimation of the misclassification probability. The results also imply that one should focus on the latent model rather than on the misclassification probability when dealing with misclassification error models.

7 Appendix

Remark 1

The key difference between Mahajan (2003) and this paper lies in Assumption 2. The comparable part, i.e., section 4, of the former paper relies on the assumption as follows:

Assumption 3: $f_{x|x^*wz}(x|x^*,w,z) = f_{x|x^*}(x|x^*)$.

Assumption 3 means the misclassification error is independent of the IV and other explana-

tory variables conditional on the latent variable. Obviously, Assumption 2 considered in this paper is much weaker than Assumption 3. Besides this difference in assumptions there are important differences in the identification strategy. Under Assumption 3, the misclassification probability $f_{x|x^*}$ is treated as additional unknown parameters (or constants) in Mahajan (2003). The identification of the parameter of interest together with $f_{x|x^*}$ is proved by contradiction in his Lemma 4 and hence is not constructive in the sense that it leads directly to an estimator. The parameters are estimated together as a "plug-in" semiparametric MLE (Newey and McFadden, 1994). [5] In this paper, the identification approach is to solve a nonlinear inversion problem. The model $f_{y|x^*w}$ is expressed as a known function of densities of observed distributions, and, therefore, is nonparametrically identified. The likelihood function is much more complicated than that in Mahajan (2003, 2004) but the advantage is that it can allow for a more general form of the misclassification error model. Moreover, the estimator is still a simple "plug-in" semiparametric MLE.

Under Assumption 2, Mahajan (2003) considers the case where repeated measurements are available. Although there is some similarity between IV and repeated measurements, the likelihood based approach in his paper exploits the specific structure imposed by repeated measurements and makes it hard to compare his section 5 with this paper. However, there are still two points worthy mentioning. First, his identification in Lemma 9 still relies on Assumption 3, which is stronger than the assumption used in this paper. Second, if one treats the secondary measurement as an IV, this paper suggests that the likelihood can be written in such a way that there exists a simple "plug-in" semiparametric MLE. Therefore, a sieve estimator of nuisance functions in his section 5 is redundant.[6]

Proof. Lemma 1

[5] Assumption 3 can be relaxed to $f_{x|x^*wz}(x|x^*, w_1, w_2, z) = f_{x|x^*w_1}(x|x^*, w_1)$ with $w = (w_1, w_2)$. This assumption means that the misclassification probability has to be independent of part of the explanatory variable w_2 and the IV z conditional on the rest of the explanatory variables x^* and w_2. If one wants to generalize the estimator under Assumption 3 to this case, the misclassification probability $f_{x|x^*w_1}$ has to be estimated as a nuisance unknown function, or an infinitely dimensional unknown parameter. The estimator can not be as easy as a "plug-in" semiparametric MLE anymore. A sieve estimator of $f_{x|x^*w_1}$ has to be used in the estimation of the parameter of interest. The identification condition still needs to be found in this case.

[6] I am thankful to Geert Ridder for his suggestions on comparison between Mahajan (2003) and this paper.

First, we prove equations 2 and 4. We have

$$f_{yx|zw}(y,x|z,w)=\sum_{x^*}f_{yxx^*|zw}(y,x,x^*|z,w) \tag{80}$$

where

$$f_{yxx^*|zw}(y,x,x^*|z,w)=f_{y|xx^*zw}(y|x,x^*,z,w)f_{x|x^*zw}(x|x^*,z,w)f_{x^*|zw}(x^*|z,w). \tag{81}$$

By assumptions 1 and 2, we have

$$f_{yxx^*|zw}(y,x,x^*|z,w)=f_{y|x^*w}(y|x^*,w)f_{x|x^*w}(x|x^*,w)f_{x^*|zw}(x^*|z,w). \tag{82}$$

Therefore, we have

$$f_{yx|zw}(y,x|z,w)=\sum_{x^*}f_{y|x^*w}(y|x^*,w)f_{x|x^*w}(x|x^*,w)f_{x^*|zw}(x^*|z,w). \tag{83}$$

The expression of $f_{yx|zw}(y,x|z,w)$ in matrix terms in

$$\mathbf{F}_{yx|zw}=\mathbf{F}_{x^*|zw}\times\mathbf{F}_{y|x^*w}\times\mathbf{F}_{x|x^*w} \tag{84}$$

can be shown by directly checking the equation. For example, let $k=2$. The right-hand side of equation 4 is

$$\mathbf{F}_{x^*|zw}\times\mathbf{F}_{y|x^*w}\times\mathbf{F}_{x|x^*w} \tag{85}$$

$$\begin{aligned}
&= \begin{pmatrix} f_{x^*|zw}(1|1,w) & f_{x^*|zw}(2|1,w) \\ f_{x^*|zw}(1|2,w) & f_{x^*|zw}(2|2,w) \end{pmatrix} \times \begin{pmatrix} f_{y|x^*w}(y|1,w) & 0 \\ 0 & f_{y|x^*w}(y|2,w) \end{pmatrix} \\
&\quad \times \begin{pmatrix} f_{x|x^*w}(1|1,w) & f_{x|x^*w}(2|1,w) \\ f_{x|x^*w}(1|2,w) & f_{x|x^*w}(2|2,w) \end{pmatrix}
\end{aligned} \tag{86}$$

$$= \begin{pmatrix} f_{x^*|zw}(1|1,w) & f_{x^*|zw}(2|1,w) \\ f_{x^*|zw}(1|2,w) & f_{x^*|zw}(2|2,w) \end{pmatrix} \times \begin{pmatrix} f_{y|x^*w}(y|1,w)f_{x|x^*w}(1|1,w) & f_{y|x^*w}(y|1,w)f_{x|x^*w}(2|1,w) \\ f_{y|x^*w}(y|2,w)f_{x|x^*w}(1|2,w) & f_{y|x^*w}(y|2,w)f_{x|x^*w}(2|2,w) \end{pmatrix}.$$

By assumptions 1 and 2, we have

$$f_{yx|x^*w}(y,x|x^*,w) = f_{y|x^*w}(y|x^*,w) f_{x|x^*w}(x|x^*,w) \tag{87}$$

and therefore,

$$\begin{aligned} & \mathbf{F}_{x^*|zw} \times \mathbf{F}_{y|x^*w} \times \mathbf{F}_{x|x^*w} \qquad (88) \\ = & \begin{pmatrix} f_{x^*|zw}(1|1,w) & f_{x^*|zw}(2|1,w) \\ f_{x^*|zw}(1|2,w) & f_{x^*|zw}(2|2,w) \end{pmatrix} \times \begin{pmatrix} f_{yx|x^*w}(y,1|1,w) & f_{yx|x^*w}(y,2|1,w) \\ f_{yx|x^*w}(y,1|2,w) & f_{yx|x^*w}(y,2|2,w) \end{pmatrix} \\ = & \begin{pmatrix} \begin{pmatrix} f_{yx|x^*w}(y,1|1,w) f_{x^*|zw}(1|1,w) \\ +f_{yx|x^*w}(y,1|2,w) f_{x^*|zw}(2|1,w) \end{pmatrix} & \begin{pmatrix} f_{yx|x^*w}(y,2|1,w) f_{x^*|zw}(1|1,w) \\ +f_{yx|x^*w}(y,2|2,w) f_{x^*|zw}(2|1,w) \end{pmatrix} \\ \begin{pmatrix} f_{yx|x^*w}(y,1|1,w) f_{x^*|zw}(1|2,w) \\ +f_{yx|x^*w}(y,1|2,w) f_{x^*|zw}(2|2,w) \end{pmatrix} & \begin{pmatrix} f_{yx|x^*w}(y,2|1,w) f_{x^*|zw}(1|2,w) \\ +f_{yx|x^*w}(y,2|2,w) f_{x^*|zw}(2|2,w) \end{pmatrix} \end{pmatrix}. \end{aligned}$$

Again by assumption 1 and 2, we have

$$f_{yxx^*|zw}(y,x,x^*|z,w) = f_{yx|x^*w}(y,x|x^*,w) f_{x^*|zw}(x^*|z,w)$$

and then

$$\begin{aligned} & \mathbf{F}_{x^*|zw} \times \mathbf{F}_{y|x^*w} \times \mathbf{F}_{x|x^*w} \qquad (89) \\ = & \begin{pmatrix} f_{yxx^*|zw}(y,1,1|1,w) + f_{yxx^*|zw}(y,1,2|1,w) & f_{yxx^*|zw}(y,2,1|1,w) + f_{yxx^*|zw}(y,2,2|1,w) \\ f_{yxx^*|zw}(y,1,1|2,w) + f_{yxx^*|zw}(y,1,2|2,w) & f_{yxx^*|zw}(y,2,1|2,w) + f_{yxx^*|zw}(y,2,2|2,w) \end{pmatrix}. \end{aligned}$$

Since

$$f_{yx|zw}(y,x|z,w) = f_{yxx^*|zw}(y,x,1|z,w) + f_{yxx^*|zw}(y,x,2|z,w), \tag{90}$$

we have

$$\begin{aligned} & \mathbf{F}_{x^*|zw} \times \mathbf{F}_{y|x^*w} \times \mathbf{F}_{x|x^*w} \qquad (91) \\ = & \begin{pmatrix} f_{yx|zw}(y,1|1,w) & f_{yx|zw}(y,2|1,w) \\ f_{yx|zw}(y,1|2,w) & f_{yx|zw}(y,2|2,w) \end{pmatrix} = \mathbf{F}_{yx|zw}. \end{aligned}$$

Secondly, we show equations 3 and 5. We integrate y out in equation 83 to get

$$f_{x|zw}(x|z,w) = \sum_{x^*} f_{x|x^*w}(x|x^*,w) f_{x^*|zw|}(x^*|z,w). \tag{92}$$

We want to show that equation

$$\mathbf{F}_{x|zw} = \mathbf{F}_{x^*|zw} \times \mathbf{F}_{x|x^*w}, \tag{93}$$

93 is the same as equation 92 with different values of x and z. The right-hand side of equation 93 is

$$\begin{aligned}
& \mathbf{F}_{x^*|zw} \times \mathbf{F}_{x|x^*w} && (94)\\
= & \begin{pmatrix} f_{x^*|zw}(1|1,w) & f_{x^*|zw}(2|1,w) \\ f_{x^*|zw}(1|2,w) & f_{x^*|zw}(2|2,w) \end{pmatrix} \times \begin{pmatrix} f_{x|x^*w}(1|1,w) & f_{x|x^*w}(2|1,w) \\ f_{x|x^*w}(1|2,w) & f_{x|x^*w}(2|2,w) \end{pmatrix} \\
= & \begin{pmatrix} \begin{pmatrix} f_{x|x^*w}(1|1,w)f_{x^*|zw}(1|1,w) \\ +f_{x|x^*w}(1|2,w)f_{x^*|zw}(2|1,w) \end{pmatrix} & \begin{pmatrix} f_{x|x^*w}(2|1,w)f_{x^*|zw}(1|1,w) \\ +f_{x|x^*w}(2|2,w)f_{x^*|zw}(2|1,w) \end{pmatrix} \\ \begin{pmatrix} f_{x|x^*w}(1|1,w)f_{x^*|zw}(1|2,w) \\ +f_{x|x^*w}(1|2,w)f_{x^*|zw}(2|2,w) \end{pmatrix} & \begin{pmatrix} f_{x|x^*w}(2|1,w)f_{x^*|zw}(1|2,w) \\ +f_{x|x^*w}(2|2,w)f_{x^*|zw}(2|2,w) \end{pmatrix} \end{pmatrix} && (95)
\end{aligned}$$

By assumptions 1 and 2, we have

$$f_{xx^*|zw}(x,x^*|z,w) = f_{x|x^*w}(x|x^*,w) f_{x^*|zw}(x^*|z,w)$$

and then

$$\begin{aligned}
& \mathbf{F}_{x^*|zw} \times \mathbf{F}_{x|x^*w} && (96)\\
= & \begin{pmatrix} f_{xx^*|zw}(1,1|1,w) + f_{xx^*|zw}(1,2|1,w) & f_{xx^*|zw}(2,1|1,w) + f_{xx^*|zw}(2,2|1,w) \\ f_{xx^*|zw}(1,1|2,w) + f_{xx^*|zw}(1,2|2,w) & f_{xx^*|zw}(2,1|2,w) + f_{xx^*|zw}(2,2|2,w) \end{pmatrix}.
\end{aligned}$$

Since

$$f_{x|zw}(x|z,w) = f_{xx^*|zw}(x,1|z,w) + f_{xx^*|zw}(x,2|z,w), \tag{97}$$

we have

$$\begin{aligned} & \mathbf{F}_{x^*|zw} \times \mathbf{F}_{x|x^*w} \\ = & \begin{pmatrix} f_{x|zw}(1|1,w) & f_{x|zw}(2|1,w) \\ f_{x|zw}(1|2,w) & f_{x|zw}(2|2,w) \end{pmatrix} = \mathbf{F}_{x|zw}. \end{aligned} \tag{98}$$

■

Proof. Theorem 2 (consistency)

The score function in the semiparametric MLE is

$$g(\omega, \theta, \gamma) = \nabla_\theta \ln f_{y|xwz}(\omega, \theta, \gamma), \tag{99}$$

with

$$\begin{aligned} & \ln f_{y|xwz}(y|x,w,z;\theta,\gamma) \\ = & \ln\left(f_{y|x^*w}(y|1,w;\theta) + f_{y|x^*w}(y|0,w;\theta) - \frac{f_{wz}(w,z)}{f_{ywz}(y,w,z)} f_{y|x^*w}(y|1,w;\theta) f_{y|x^*w}(y|0,w;\theta) - B \right) \\ & + \ln\left(\frac{f_{y|wz}(y|w,z)}{f_{x|wz}(x|w,z)} \frac{1}{A} \right) \end{aligned} \tag{100}$$

and

$$B = \frac{f_{ywz}(y,w,0) f_{xwz}(x,w,1) - f_{ywz}(y,w,1) f_{xwz}(x,w,0)}{f_{xwz}(x,w,1) f_{wz}(w,0) - f_{xwz}(x,w,0) f_{wz}(w,1)}.$$

Note that the second term in the expression of $\ln f_{y|xwz}$ does not play a role in the estimation because it does not contain any unknown parameters. In fact,

$$g(\omega, \theta, \gamma) = \frac{\nabla_\theta H}{H}$$

with

$$\begin{aligned} H &= f_{y|x^*w}(y|1,w;\theta) + f_{y|x^*w}(y|0,w;\theta) - \frac{f_{wz}(w,z)}{f_{ywz}(y,w,z)} f_{y|x^*w}(y|1,w;\theta) f_{y|x^*w}(y|0,w;\theta) - B \\ \nabla_\theta H &= \nabla_\theta f_{y|x^*w}(y|1,w;\theta) + \nabla_\theta f_{y|x^*w}(y|0,w;\theta) \\ & \quad - \frac{f_{wz}(w,z)}{f_{ywz}(y,w,z)} \left[\nabla_\theta f_{y|x^*w}(y|1,w;\theta) f_{y|x^*w}(y|0,w;\theta) + f_{y|x^*w}(y|1,w;\theta) \nabla_\theta f_{y|x^*w}(y|0,w;\theta) \right] \end{aligned}$$

Let

$$\gamma \equiv [\gamma_1, \gamma_2, \gamma_3]^T$$

and we have the Frechet derivative of g with respect to γ as follows:

$$g_\gamma(\omega, \theta, \gamma) = \left[g_{\gamma_1}, g_{\gamma_2}, g_{\gamma_3}\right]^T$$

Therefore, we have, for $i = 1$, 2, and 3

$$g_{\gamma_i} = \frac{-1}{H^2} \left(\nabla_\theta H\right) \left(H_{\gamma_i}\right) + \frac{1}{H} \left(\nabla_\theta H\right)_{\gamma_i}$$

with

$$\begin{aligned} H_{\gamma_1} &= \frac{f_{wz}(w,z)}{f^2_{ywz}(y,w,z)} f_{y|x^*w}(y|1,w;\theta) f_{y|x^*w}(y|0,w;\theta) \\ &\quad - \frac{(-1)^z f_{xwz}(x,w,1-z)}{f_{xwz}(x,w,1) f_{wz}(w,0) - f_{xwz}(x,w,0) f_{wz}(w,1)}, \end{aligned}$$

$$\left(\nabla_\theta H\right)_{\gamma_1} = \frac{f_{wz}(w,z)}{f^2_{ywz}(y,w,z)} \left[\nabla_\theta f_{y|x^*w}(y|1,w;\theta) f_{y|x^*w}(y|0,w;\theta) + f_{y|x^*w}(y|1,w;\theta) \nabla_\theta f_{y|x^*w}(y|0,w;\theta)\right].$$

$$\begin{aligned} H_{\gamma_2} &= \frac{(-1)^z f_{ywz}(y,w,1-z)}{f_{xwz}(x,w,1) f_{wz}(w,0) - f_{xwz}(x,w,0) f_{wz}(w,1)} \\ &\quad + B \frac{(-1)^z f_{wz}(w,1-z)}{f_{xwz}(x,w,1) f_{wz}(w,0) - f_{xwz}(x,w,0) f_{wz}(w,1)}, \end{aligned}$$

$$\left(\nabla_\theta H\right)_{\gamma_2} = 0,$$

$$\begin{aligned} H_{\gamma_3} &= -\frac{1}{f_{ywz}(y,w,z)} f_{y|x^*w}(y|1,w;\theta) f_{y|x^*w}(y|0,w;\theta) \\ &\quad + B \frac{(-1)^z f_{xwz}(x,w,1-z)}{f_{xwz}(x,w,1) f_{wz}(w,0) - f_{xwz}(x,w,0) f_{wz}(w,1)}, \end{aligned}$$

$$\left(\nabla_\theta H\right)_{\gamma_3} = -\frac{1}{f_{ywz}(y,w,z)} \left[\nabla_\theta f_{y|x^*w}(y|1,w;\theta) f_{y|x^*w}(y|0,w;\theta) + f_{y|x^*w}(y|1,w;\theta) \nabla_\theta f_{y|x^*w}(y|0,w;\theta)\right].$$

We define

$$Q_n(\theta, \gamma) = \frac{1}{n} \sum_{i=1}^{n} g(\omega_i, \theta, \gamma), \tag{101}$$

$$Q_0(\theta, \gamma) = Eg(\omega, \theta, \gamma). \tag{102}$$

Then

$$\sup_{\theta \in \Theta} |Q_n(\theta, \widehat{\gamma}) - Q_0(\theta, \gamma_0)| \leq \sup_{\theta \in \Theta} |Q_n(\theta, \widehat{\gamma}) - Q_n(\theta, \gamma_0)| + \sup_{\theta \in \Theta} |Q_n(\theta, \gamma_0) - Q_0(\theta, \gamma_0)|. \tag{103}$$

From the explicit expression of $f_{y|xwz}$ in equation 100, we know $Q_0(\theta, \gamma)$ is differentiable in γ. Therefore, the first term on the right-hand side can be bounded as follows:

$$\sup_{\theta \in \Theta} |Q_n(\theta, \widehat{\gamma}) - Q_n(\theta, \gamma_0)| \leq \sup_{\omega \in \mathcal{W}} |\widehat{\gamma}(\omega) - \gamma_0(\omega)| \times \sup_{\theta \in \Theta} \sup_{\|\gamma - \gamma_0\|_\infty \leq \varepsilon} \|\nabla_\gamma Q_n(\theta, \gamma)\| \tag{104}$$

for some $\varepsilon \to 0$ as $n \to \infty$. The norm in the last term is the Sobolev norm. By lemma 2, the assumption that $h \to 0$, and $\ln n/(nh^{r+2d}) \to 0$ *as* $n \to \infty$, guarantee the uniform convergence of $\widehat{\gamma}(\omega)$, $\sup_{\omega \in \mathcal{W}} |\widehat{\gamma}(\omega) - \gamma_0(\omega)| = o_p(1)$. Since $f_{y|x^*w}$, $\nabla_\theta f_{y|x^*w}$, and $\gamma(.)$ are bounded, the observed density $f_{y|xwz}$ and $\nabla_\theta f_{y|xwz}$ are also bounded. The denominator in B is bounded away from zero because the determinant of the matrix $\mathbf{F}_{x|zw}$ is bounded away from zero by assumptions 2.1 and 2.2. Therefore $\sup_{\theta \in \Theta} \sup_{\|\gamma - \gamma_0\|_\infty \leq \varepsilon} \|\nabla_\gamma Q_n(\theta, \gamma)\|$ is finite. Thus, we have

$$\sup_{\theta \in \Theta} |Q_n(\theta, \widehat{\gamma}) - Q_n(\theta, \gamma_0)| = o_p(1). \tag{105}$$

The uniform convergence of $Q_n(\theta, \gamma_0)$ follows the boundedness of $f_{y|xwz}$ and $\nabla_\theta f_{y|xwz}$ and the continuity of $g(\omega, \theta, \gamma)$. Therefore, we have

$$\sup_{\theta \in \Theta} |Q_n(\theta, \gamma_0) - Q_0(\theta, \gamma_0)| = o_p(1) \tag{106}$$

By theorem 4.1.1 in Amemiya (1985b, page 106), we have

$$\hat{\theta} \xrightarrow{p} \theta_0. \tag{107}$$

■

Proof. Lemma 3

The major step is to show

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(\omega_i,\theta,\widehat{\gamma}) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[g(\omega_i,\theta,\gamma_0)+\delta(\omega_i)\right] = o_p(1). \tag{108}$$

Therefore $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(\omega_i,\theta,\widehat{\gamma})$ has the same distribution as $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[g(\omega_i,\theta,\gamma_0)+\delta(\omega_i)\right]$, which converges to a normal distribution. By the explicit expression of $f_{y|xwz}$ in equation 100, the function $g(\omega,\theta,\gamma)$ is differentiable in γ. The linearization of $g(\omega,\theta_0,\gamma)$ with respect to γ gives

$$\begin{aligned} g(\omega,\theta_0,\gamma) &= g(\omega,\theta_0,\gamma_0) + g_\gamma^T(\omega,\theta_0,\gamma_0)\left[\gamma(\omega)-\gamma_0(\omega)\right] \\ &\quad + \left[\gamma(\omega)-\gamma_0(\omega)\right]^T g_{\gamma\gamma}(\omega,\theta_0,\gamma_0)\left[\gamma(\omega)-\gamma_0(\omega)\right] + o(\|\gamma-\gamma_0\|^2). \end{aligned} \tag{109}$$

The expression of $g_{\gamma\gamma}(\omega,\theta_0,\gamma_0)$ is complicated and less intuitive. Therefore, we omit it in this proof. But it is obvious that the explicit expression of $g_{\gamma\gamma}(\omega,\theta_0,\gamma_0)$ is can be found out as that of $g_\gamma(\omega,\theta_0,\gamma_0)$. And the boundedness of $g_{\gamma\gamma}(\omega,\theta_0,\gamma_0)$ is due to that of $f_{y|x^*w}$, $\nabla_\theta f_{y|x^*w}$, and $\gamma(.)$. Therefore, for all γ with $\|\gamma-\gamma_0\|$ small enough, there exists a function $b(\omega)$ with $E\left[b(\omega)\right]<\infty$ and

$$\left\|g(\omega,\theta_0,\gamma) - g(\omega,\theta_0,\gamma_0) - g_\gamma^T(\omega,\theta_0,\gamma_0)\left[\gamma(\omega)-\gamma_0(\omega)\right]\right\| \le b(\omega)\left\|\gamma-\gamma_0\right\|^2. \tag{110}$$

The boundedness of $E\left[b(\omega)\right]$ is because $f_{y|x^*w}$, $\nabla_\theta f_{y|x^*w}$, and $\gamma(.)$ are bounded from above, and γ is also bounded from zero. The assumption that $\sqrt{n}h^{2m}\to 0$, and $\sqrt{n}\ln n/(nh^{r+2d})\to 0$ *as* $n\to\infty$, guarantees $\sqrt{n}\left\|\gamma-\gamma_0\right\|^2 = o_p(1)$. Define

$$G(\omega,\gamma-\gamma_0) = g_\gamma^T(\omega,\theta_0,\gamma_0)\left[\gamma(\omega)-\gamma_0(\omega)\right]. \tag{111}$$

We then have

$$\left\|g(\omega,\theta_0,\gamma) - g(\omega,\theta_0,\gamma_0) - G(\omega,\gamma-\gamma_0)\right\| \le b(\omega)\left\|\gamma-\gamma_0\right\|^2, \tag{112}$$

which is the assumption (i) in theorem 8.11 in Newey and McFadden (1994, page 2209). The assumption (ii) in theorem 8.11 requires $E\left[\left\|g_\gamma(\omega,\theta_0,\gamma_0)\right\|^2\right] < \infty$. This is obvious by the expression of $f_{y|xwz}$ and the boundedness of $f_{y|x^*w}$, $\nabla_\theta f_{y|x^*w}$, and $\gamma(.)$. The function $G(\omega,\gamma)$ is a linear function of γ so that we may have

$$\int G(\omega,\gamma)dF_0(\omega) = \int v(\omega)\gamma(\omega)d\omega, \tag{113}$$

where

$$v(\widetilde{\omega}) = E\left[\left. g_\gamma(\omega,\theta_0,\gamma)|_{\gamma=\gamma_0(\widetilde{\omega})}\right| \widetilde{\omega}\right]. \tag{114}$$

Therefore, assumption (iii) in theorem 8.11 is satisfied. Finally, $v(\widetilde{\omega})$ is bounded and continuous almost everywhere $\int \|v(\omega)\|\, d\omega < \infty$ by the properties of the function $f_{y|xwz}$. Since $f_{yxwz}(\omega)$ is continuously differentiable of order 5, there is $\varepsilon > 0$ such that $E\left[\sup_{\|\eta\|<\varepsilon}\|v(\omega+\eta)\|^4\right] < \infty$. Thus, the last assumption in theorem 8.11 is satisfied. Let

$$\delta(\omega) = v(\omega) - E\left[v(\omega)\right]. \tag{115}$$

Here we abuse the notation to use $v(\omega)$ to stand for its column sum. The explicit expression of $v(\omega)$ is in equation 60 in the maintext. By the central limit theorem, $\frac{1}{\sqrt{n}}\sum_{i=1}^n\left[g(\omega_i,\theta,\gamma_0)+\delta(\omega_i)\right]$ converges to a normal distribution with mean zero and variance $\Omega = Var\left\{g(\omega,\theta_0,\gamma_0)+\delta(\omega)\right\}$. Finally, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n g(\omega_i,\theta_0,\widehat{\gamma}) \xrightarrow{d} N(0,\Omega). \tag{116}$$

■

Proof. Theorem 3 (asymptotic normality)

We show the asymptotic normality of the estimator using the delta method. First, we have

$$\sqrt{n}(\widehat{\theta}-\theta_0) = -\left[\frac{1}{n}\sum_{i=1}^n \nabla_\theta g(\omega_i,\widetilde{\theta},\widehat{\gamma})\right]^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n g(\omega_i,\theta_0,\widehat{\gamma})\right) \tag{117}$$

where $\widetilde{\theta}$ is between $\widehat{\theta}$ and θ_0. The distribution of the second term on the right-hand side is

shown to be normal in lemma 3 as follows:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(\omega_i, \theta_0, \widehat{\gamma}) \xrightarrow{d} N(0, Var\left\{g(\omega, \theta_0, \gamma_0) + \delta(\omega)\right\}). \tag{118}$$

Because the density $f_{x^*|zw}$, $\nabla_\theta f_{y|x^*w}$ and f_{yxwz} are both bounded from infinity and zero on the support, $\nabla_\theta g(\omega, \theta_0, \gamma)$, $g_\gamma(\omega, \theta_0, \gamma)$, $\nabla_{\theta\theta} g(\omega, \theta_0, \gamma)$, $g_{\gamma\gamma}(\omega, \theta_0, \gamma)$, and $(\nabla_\theta g)_\gamma$ are, therefore, bounded. Therefore, there are $b(\omega)$, $\varepsilon > 0$ with $E\left[b(\omega)\right] < \infty$ such that

$$\|\nabla_\theta g(\omega, \theta, \gamma) - \nabla_\theta g(\omega, \theta_0, \gamma_0)\| \leq b(\omega)\left[\|\theta - \theta_0\|^\varepsilon + \|\gamma - \gamma_0\|^\varepsilon\right]. \tag{119}$$

As shown in theorem 8.12 in Newey and McFadden (1994), we have

$$\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta g(\omega_i, \overline{\theta}, \widehat{\gamma}) \xrightarrow{p} E\left[\nabla_\theta g(\omega, \theta_0, \gamma_0)\right]. \tag{120}$$

Because $E\left[\nabla_{\theta\theta} \ln f_{y|xwz}\right]$ exists and is nonsingular, the Slutsky theorem then gives

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, G_\theta^{-1} \Omega G_\theta^{-1\prime}) \tag{121}$$

where

$$G_\theta = E\left[\nabla_\theta g(\omega, \theta_0, \gamma_0)\right] \tag{122}$$

$$\Omega = Var\left[g(\omega, \theta_0, \gamma_0) + \delta(\omega)\right]. \tag{123}$$

■

References

[1] Aigner, D., 1973, "Regression with a binary independent variable subject to errors of observations," Journal of Econometrics, 1, pp. 49-60.

[2] Amemiya, Y., 1985a, "Instrumental variable estimator for the nonlinear errors-in-variables model," Journal of Econometrics, 28, pp. 273-289.

[3] Amemiya, Y., 1985b, Advanced Econometrics, Cambridge, MA: Harvard University Press.

[4] Amemiya, Y., and Fuller, W. A., 1988, "Estimation for the nonlinear functional relationship," Annals of Statistics, 16, pp. 147-160.

[5] Black, D., M. C. Berger, and F. A. Scott, 2000, "Bounding parameter estimates with nonclassical measurement error," Journal of the American Statistical Association, 95, pp. 739-748.

[6] Bollinger, C., 1996, "Bounding mean regressions when a binary regressor is mismeasured," Journal of Econometrics, 73 (1996), pp. 387-399.

[7] Bound, J., C. Brown, and N. Mathiowetz, 2001, "Measurement error in survey data," in Handbook of Econometrics, J. J. Heckman and E. Leamer, eds., vol 5.

[8] Buzas, J. S., 1997, "Instrumental variable estimation in nonlinear measurement error models," Communications in Statistics, Part A – Theory and Methods, 26, pp. 2861-2877.

[9] Carroll, R. J., D. Ruppert, and L. A. Stefanski, 1995, Measurement Error in Nonlinear Models, Chapman & Hall, New York.

[10] 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.

[11] Darolles, S., J.-P. Florens, and E. Renault, 2000, "Nonparametric instrumental regression," Manuscript, GREMAQ, University of Toulouse.

[12] Freeman, R., 1984, "Longitudinal analyses of the effects of trade unions," Journal of Labor Economics, 2, pp. 1-26.

[13] Hausman, J., J. Abreveya, and F. Scott-Morton, 1998, "Misclassification of the dependent variable in a discrete response setting," Journal of Econometrics, 87, pp. 239-269.

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

[15] Hausman, J., W. K. Newey, and J. L. Powell, 1995, "Nonlinear errors in variables: estimation of some Engle curves," Journal of Econometrics, 65, pp. 205-233.

[16] Horn, R., and C. Johnson, 1985, Matrix Analysis, Cambridge University Press.

[17] Horowitz, J., and C. Manski, 1995, "Identification and robustness with contaminated and corrupt data," Econometrica, 63, pp. 281-302.

[18] Kane, T. J., C. E. Rouse, and D. Staiger, 1999, "Estimating returns to schooling when schooling is misreported," NBER working paper #7235.

[19] Levine, P., 1993, "CPS contemporaneous and retrospective unemployment compared," Monthly Labor Review 116, pp. 33-39.

[20] Lewbel, A., 1998, "Semiparametric latent variable model estimation with endogenous or mismeasured regressors," Econometrica, vol. 66, pp. 105-121.

[21] Lewbel, A., 2000, "Identification of the binary choice model with misclassification," Econometric Theory, 16, pp. 603-660.

[22] Lewbel, A., 2003, "Estimation of average treatment effects with misclassification," memo.

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

[24] Mahajan, A., 2003, "Misclassified regressors in binary choice models," memo.

[25] Mahajan, A. 2004, "Identification and estimation of single index models with misclassified regressors," memo.

[26] Molinari, F., 2004, "Partial identification of probability distributions with misclassified data," memo.

[27] Newey, W., 1992, "Partial means, kernel estimation, and a general asymptotic variance estimator," memo, MIT.

[28] Newey, W., 1994, "The asymptotic variance of semiparametric estimators," Econometrica, vol 62, pp. 1349-1382.

[29] Newey, W., 2001, "Flexible simulated moment estimation of nonlinear errors-in-variables models," Review of Economics and Statistics, 83(4), pp. 616-627.

[30] Newey, W., and D. McFadden, 1994, "Large sample estimation and hypothesis testing," in Handbook of Econometrics, R. Engle and D. McFadden, eds., vol. 4, pp. 2111-2245.

[31] Newey, W., and J. Powell, 2003, "Instrumental variable estimation of nonparametric models," Econometrica, vol. 71, no. 5, pp. 1565-1578.

[32] Ramalho, E., 2002, "Regression models for choice-based samples with misclassification in the response variable," Journal of Econometrics, 106, pp. 171-201.

[33] Schennach, S., 2004a, "Estimation of nonlinear models with measurement error," Econometrica, vol. 72, no 1, pp. 33-76.

[34] Schennach, S., 2004b, "Instrumental variable estimation of nonlinear errors-in-variables models," memo.

[35] Wang, L., and C. Hsiao, 1995, "A simulation-based semiparametric estimation of nonlinear errors-in-variables models," Working paper, University of Southern California.

[36] Wooldridge, J., 2002, Econometric Analysis of Cross Section and Panel Data, MIT press.

Table 1: Simulation results of Probit model: sample size 500; number of repetitions 200.

	β_1			β_2			β_0		
$p = .3$ $q = .2$	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.
Ignoring meas. error	.541	-.523	.139	.181	-.103	.149	.293	.277	.095
True x^*	.160	.015	.159	.166	-.013	.165	.104	.000	.104
I.V.	.421	-.095	.410	.307	-.087	.295	.263	.045	.259
$p = .3 - .1w$ $q = .2 + .1w$	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.
Ignoring meas. error	.538	-.520	.136	.210	-.150	.147	.290	.275	.094
True x^*	.157	.012	.157	.165	-.011	.164	.104	.000	.104
I.V.	.409	-.124	.389	.332	-.138	.302	.238	.061	.230
$p = .3 + .1w$ $q = .2 + .1w$	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.
Ignoring meas. error	.509	-.491	.137	.176	-.094	.149	.279	.263	.093
True x^*	.160	.015	.159	.165	-.014	.165	.104	-.001	.104
I.V.	.318	-.108	.299	.307	-.071	.298	.205	.052	.198

Note:
1) $\beta_1 = 1$,$\beta_2 = 1$,$\beta_0 = .5$, $x^* = I(\epsilon < .6)$; $z = I(\epsilon + \delta < .6)$, $\epsilon \sim Uniform(0, 1)$, $\delta \sim N(0, .04)$, $(\rho_{x^*z} \approx .67)$, $w \sim N(0, .25)$.
2) $\Pr(x = 0|x^* = 1, w) = \min(1, \max(0, p))$, $\Pr(x = 1|x^* = 0, w) = \min(1, \max(0, q))$.
3) $K(x) = .5(3 - x^2)\phi(x)$ and $h = .2$, where $\phi(x)$ is the standard normal density.

Table 2: Simulation results of Probit model: sample size 500; number of repetitions 200.

$x = F_{x\|x^*w}(\eta)$ $\eta = \eta_x$	β_1			β_2			β_0		
	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.
Ignoring meas. error	.5662	-.5435	.1587	.2090	.0371	.2057	1.2081	1.1032	0.4924
True x^*	.0285	.0003	.0285	.0180	-.0023	.0178	.0463	-.0063	.0458
I.V.	.3504	-.0119	.3502	.1561	-.0722	.1384	.5860	-.0539	.5835
$x = F_{x\|x^*w}(\eta)$ $\eta = .9\eta_x + .1\Phi(w)$	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.
Ignoring meas. error	.4690	-.4495	.1336	.2103	.0102	.2101	1.0057	.9095	.4291
True x^*	.0285	.0003	.0285	.0180	-.0023	.0178	.0463	-.0063	.0458
I.V.	.3164	-.0268	.3152	.1527	-.0789	.1307	.5162	-.0242	.5157
$x = F_{x\|x^*w}(\eta)$ $\eta = .9\eta_x + .1\left[1 - \Phi(w)\right]$	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.
Ignoring meas. error	.4189	-.3907	.1511	.1975	.0582	.1888	.8712	.7436	.4539
True x^*	.0285	.0003	.0285	.0180	-.0023	.0178	.0463	-.0063	.0458
I.V.	.3077	-.0824	.2965	.1583	-.0728	.1405	.5171	.0612	.5171

Table 3: Joint distribution of education (1457 observations)

	parents' education		
education	high school or lower	college or higher	total
high school or lower	.337	.163	.500
college or higher	.204	.296	.500
total	.541	.459	1

Table 4: Summary statistics of variables (1457 observations)

	mean	std.dev	min	max
employment	.822	.382	0	1
number of children	.577	.906	0	8
weeks worked in last year	41.704	12.720	26	52
age	24.173	6.133	18	56
race (white=1)	.781	.414	0	1

Table 5: Estimation results

	Ignoring meas. error		I.V.	
	estimate	std.dev	estimate	std.dev
education	.3342	.0859	.9578	.1992
work experience	.0296	.0032	.0235	.0057
age	.0155	.0080	.0100	.0146
number of kids	-.0311	.0447	-.0222	.0996
race	.2599	.0955	.1839	.1838
constant	-1.2664	.2388	-1.1367	.6700

Table 6: Joint distribution of education (1688 observations)

	parents' education			
education	high school or lower	some college	college or higher	total
high school or lower	.361	.134	.072	.568
some college	.111	.092	.050	.254
college or higher	.065	.039	.075	.179
total	.537	.265	.198	1

Table 7: Summary statistics of variables (1688 observations)

	mean	std.dev	min	max
number of children	.790	1.092	0	9
employment (yes=1)	.799	.401	0	1
age	22.982	6.443	15	56
race (white=1)	.798	.401	0	1

Table 8: NLS Estimation results

	Ignoring meas. error		I.V.	
	estimate	std.dev	estimate	std.dev
education	-.0188	.0177	-.0539	.0223
employment	-.0145	.0621	-.0077	.0638
age	-.3494	.0284	-.3632	.0254
$age^2/100$	.4482	.0536	.4660	.0508
race	-.0222	.0808	-.0092	.0823
constant	5.2249	.3067	5.8884	.4076

Table 9: QMLE Estimation results

	Ignoring meas. error		I.V.	
	estimate	std.dev	estimate	std.dev
education	-.0264	.0143	-.0541	.0244
employment	-.0220	.0636	-.0065	.0638
age	-.3665	.0240	-.3900	.0220
$age^2/100$	.4979	.0454	.5272	.0438
race	.0710	.0811	.0658	.0815
constant	5.3654	.2809	6.0178	.4226