Available online at www.sciencedirect.com

ScienceDirect

Journal of Econometrics 144 (2008) 27–61

JOURNAL OF Econometrics

www.elsevier.com/locate/jeconom

Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution

Yingyao Hu*

Department of Economics, Johns Hopkins University, 440 Mergenthaler Hall, 3400 N. Charles Street, Baltimore, MD 21218, USA

Available online 17 December 2007

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 the unknowns. Such a claim requires solving a complicated nonlinear system of equations. This paper introduces a matrix diagonalization technique which allows one to easily find the unique solution of the system. The solution shows that the latent model can be expressed as an explicit function of directly observed distribution functions. Therefore, the latent model is 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; Matrix diagonalization

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 (relevant studies using repeated measurements or instrumental variables include Hausman et al., 1991; Wang and Hsiao, 1995; Newey, 2001; Li, 2002; Schennach, 2004, 2007). Measurement error in a continuous variable, such as wage or income, is considered to be continuous, while the error in a discrete variable, such as education, marital status,

*Tel.: +1 410 516 7610.
E-mail address: yhu@jhu.edu

0304-4076/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2007.12.001

or union status, is believed to be discrete. Discrete measurement 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 et al., 2001 for a survey). Using an instrumental variable, this paper achieves the nonparametric identification of a nonlinear model with a general discrete explanatory variable that is subject to misclassification error. The error is allowed to be correlated with all the explanatory variables in the model. The identification procedure may lead to a $\sqrt{n}$-consistent 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 restrictions that the misclassification error is independent of the dependent variable conditional on the true value and that the misclassification error is not very large so that the misreported variable may still be positively correlated with the true value. More restrictive assumptions include the restriction that the misclassification probabilities are independent of other explanatory variables and, therefore, are constants. These assumptions are widely used in relevant studies, such as Aigner (1973), Bollinger (1996), Kane et al. (1999), and Mahajan (2006). 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 regression models with 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 et al. (2000) estimate the slope coefficient in a regression model when a secondary measurement is available. Kane et al. (1999) and Lewbel (2007) also use instruments to solve misclassification in treatment effect models. Two close studies on misclassification in a dichotomous variable are Hui and Walter (1980) and Mahajan (2006), which use a secondary measurement or an instrument to identify a nonlinear model that includes a mismeasured binary regressor.

However, it is not clear how to extend the existing identification 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 true value. Without that independence assumption, $k \times (k-1)$ unknown density functions will need to be identified and estimated. A simple generalization of the identification results in the binary case is to claim that the number of restrictions is no less than that of unknowns. Such a claim is not enough to identify the latent model because it requires solving a complicated nonlinear system of equations. The matrix diagonalization technique introduced in this paper allows one to easily find the unique solution of the system once sufficient assumptions for point identification are imposed.

I will show the identification with misclassification in a general discrete variable under assumptions similar to those in the existing literature. I compare the assumptions in the dichotomous case of Mahajan (2006) with those in this paper. Certain useful results in this paper are also new in the dichotomous case. In the general discrete case, Molinari (2003, 2005) also formalizes misclassification problems in matrix notation to greatly facilitate identification analysis and provides the interval identification of parameters of interest. However, it is not clear when and how the partial identification becomes the point identification. This paper shows that the latent model and the misclassification error distribution 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 may also be treated as a secondary measurement of the latent variable as in Li (2002) and Schennach (2004). Amemiya (1985a) shows that IV estimators are generally biased in the estimation of nonlinear models. Under the assumption that the measurement error vanishes when the sample size increases, Amemiya and Fuller (1988) and Carroll and Stefanski (1990) obtain a consistent IV estimator for nonlinear models. The IV estimator for a polynomial regression model is discussed in Hausman et al. (1991, 1995). Buzas (1997) derives an instrumental variable estimator that is approximately consistent for general nonlinear models. Lewbel (1998) describes a consistent estimator for a particularly specified latent variable model with instrumental variables and an exclusion restriction. Newey (2001) and Schennach (2007) consider a nonlinear regression model using a prediction equation in which an instrumental variable is independent of the prediction error. Most studies on IV estimators focus on the continuous measurement error, on which certain independence restrictions are imposed. In this study, the measurement error is discrete so that the error cannot be independent of the latent true value. Instead, I assume 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 (2004, 2007).

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 a 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 distributions, and, therefore, directly leads to a nonparametric or semiparametric ''plug-in'' estimator.

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 misclassification 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^* and w is

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

The variables $\{y,x,w,z\}$ are observed in an i.i.d. sample, 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. $f_{x|x^*zw}(x|x^*,z,w) = f_{x|x^*w}(x|x^*,w)$.

These two assumptions are widely adopted in the relevant literature, such as Bollinger (1996), Kane et al. (1999), and Mahajan (2006). Assumption 1 implies that the misclassified variable x and the instrumental variable z do not contain any useful information on the dependent variable y beyond the true value x^* and covariates w. It also implies that the misclassification error in x is independent of the dependent variable y conditional on x^* and w. This type of measurement error is called nondifferential error. As discussed in Bound et al. (2001, p. 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, Assumption 1 is weaker than the one discussed in Bound et al. (2001).

Assumption 2 implies that the misclassification error in x is independent of the instrumental variable z conditional on x^* and w. This assumption is also discussed in Bound et al. (2001, p. 3732) without considering other explanatory variables w. Kane et al. (1999) use a similar assumption to obtain a consistent estimator in a GMM setting. Assumption 2 is weaker than those in previous studies because I do not impose any parametric specification on the misclassification probabilities, and these probabilities may depend on other covariates. This paper suggests that such a parametric specification is not necessary because the misclassification probabilities are nonparametrically identified. An important advantage of Assumption 2 is that it allows correlation between the misclassification error and all the explanatory variables. This extension is important because previous studies have shown the significance of such a correlation 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 individual demographic characteristics.

This paper shows that the conditional distribution function of the dependent variable $f_{y|x^*w}$ may be expressed as a known function of directly observed distribution functions and, therefore, is nonparametrically identified. To be specific, the latent density $f_{y|x^*w}$ plays the role of an eigenvalue of a matrix induced by the observed density $f_{yx|zw}$. And the corresponding eigenvector may be the misclassification probability $f_{x|x^*w}$ or the conditional density of the true value $f_{x^*|zw}$. With the density $f_{x^*|zw}$ identified, I can estimate a parameter of interest θ_0 in a latent moment condition $\mathrm{E}(y|x^*, w) = m(x^*, w; \theta_0)$ through an observed moment condition.

This paper is organized as follows. Section 2 shows the nonparametric identification of the latent model and the misclassification error distribution. 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 an empirical illustration. Section 6 concludes the paper. The proofs are in the appendix.

2. Identification

This section considers the nonparametric identification of the latent model and the misclassification error distribution. Suppose that x, x^*, and z share the same support $\{1, 2, \ldots, k\}$. I define the following notations:

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

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

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

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

$$\mathbf{F}_{y|zw} = (f_{y|zw}(y|1,w), \ldots, f_{y|zw}(y|k,w))^{\mathrm{T}}.$$

By Assumptions 1 and 2 and the law of total probability, the relationship between the observed and unobserved densities results as follows:

Lemma 1. *Suppose that Assumptions* 1 *and* 2 *are satisfied. Then,*

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

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

and

$$\mathbf{F}_{yx|zw} = \mathbf{F}_{x^*|zw} \times \mathbf{F}_{y|x^*w} \times \mathbf{F}_{x|x^*w}, \tag{3}$$

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

Proof. See the appendix.

Similarly, one may obtain

$$f_{y|zw}(y|z,w) = \sum_{x^*} f_{y|x^*w}(y|x^*,w) f_{x^*|zw}(x^*|z,w), \tag{5}$$

which is equivalent to

$$\mathbf{F}_{y|zw} = \mathbf{F}_{x^*|zw} \times \mathbf{F}_{y|x^*w} \times \mathbf{i} \tag{6}$$

with $\mathbf{i} = (1,\ldots,1)^{\mathrm{T}}$. In order to identify the matrix $\mathbf{F}_{y|x^*w}$, one needs the following assumption:

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

Assumption 2.1 has been used in Darolles et al. (2000) and Newey and Powell (2003) for a continuous variable, as well as in Mahajan (2006) for a dichotomous variable. Since it uses the whole conditional distribution of x^* in matrix $\mathbf{F}_{x^*|zw}$, my proposed approach can exploit any form of statistical dependence between z and x^* to achieve identification. However, Assumption 2.1 does not necessarily require the instrument z to be correlated with x^*. Here I briefly discuss the relationship between the invertibility of $\mathbf{F}_{x^*|zw}$ and the correlation between x^* and z. I ignore the covariate w for the time being. Given that $\Pr(z=z_i)\neq 0$ for all z_i in the support of z, $\mathbf{F}_{x^*|z}$ is invertible if and only if $\mathbf{F}_{x^*z}$ is invertible, where the ith row and jth column entry of the matrix $\mathbf{F}_{x^*z}$ is the joint probability $\Pr(x^*=x_j^*, z=z_i)$. Without loss of generality, I assume $\mathrm{E}(z)=0$. Then $\rho^2_{x^*z} = \frac{[\mathrm{E}(x^*z)]^2}{var(z)var(x^*)}$. I first show that the singularity of $\mathbf{F}_{x^*z}$ does not imply $\rho^2_{x^*z}=0$. For example, suppose x^* and z share the same support $\{-1,0,1\}$ with $(x_1,x_2,x_3)=(-1,0,1)$ and $(z_1,z_2,z_3)=(-1,0,1)$. Let the joint probability matrix $\mathbf{F}_{x^*z}$ of x^* and z be as follows:

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

Obviously, the matrix $\mathbf{F}_{x^*z}$ is singular. But $\mathrm{E}(x^*z)=\frac{1}{2}$, and $\mathrm{E}(z)=0$ so that $\rho_{x^*z}=0.75$. Although the correlation coefficient is large, the variable z is actually not a good instrument if one only considers the case where x^* equals -1 or 0. In fact, x^* is independent of z conditional on $x^*\neq 1$ and $z\neq 1$. The next step is to show that $\mathbf{F}_{x^*z}$ may be invertible even if $\rho^2_{x^*z}=0$. For example, suppose that

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

in the last example. In this case, $\mathrm{E}(x^*z)=0$ and $\mathrm{E}(z)=0$ so that $\rho^2_{x^*z}=0$, while $\mathbf{F}_{x^*z}$ is invertible. Interchanging the columns of $\mathbf{F}_{x^*z}$ results in

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

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 valid instrument. Since the validity of instruments is not the major focus of this paper, I leave it for future research and assume a valid instrument.

In the case where the instrument takes fewer values than the misclassified regressor, Assumption 2.1 may not hold anymore. However, it is reasonable to believe that the number of possible values of the instrument is the same as that of the misclassified regressor when the instrument is a repeated measurement of the regressor. For example, if respondents report education levels twice in a survey, one may expect the two reported education levels to have the same support. If the instrument takes more values than the misclassified regressor, Assumption 2.1 implies that one can always generate a new instrument taking the same possible values as the regressor such that this assumption still holds for the new instrument. Since such a new instrument is not unique, this case is also related to the case of more than one instrument. On the one hand, I may relax Assumptions 1 or 2 in this case to allow some of these variables to be in the latent model or to be correlated with the misclassification error. On the other hand, the inverse of the matrices $\mathbf{F}_{x^*|zw}$ and $\mathbf{F}_{x|zw}$, to be used later, will have to be replaced with the Moore–Penrose matrix inverse.[1]

By Assumption 2.1, Eq. (6) implies that

$$\mathbf{F}_{y|x^*w} \times \mathbf{i} = \mathbf{F}^{-1}_{x^*|zw} \times \mathbf{F}_{y|zw}. \tag{7}$$

Therefore, the latent density $\mathbf{F}_{y|x^*w}$ is identified if and only if the matrix $\mathbf{F}_{x^*|zw}$ is identified. In order to obtain $\mathbf{F}_{x^*|zw}$ from Eq. (4), I make the following assumption:

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

One sufficient condition for 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 $\Pr(x=i|x^*=i,w)>0.5$. This condition implies that x contains enough correct information on x^*.

I then obtain $\mathbf{F}_{x^*|zw}$ from Eq. (4) as follows:

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

Plugging this expression of $\mathbf{F}_{x^*|zw}$ into Eq. (7) gives

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

Eq. (9) implies that $\mathbf{F}_{y|x^*w}$ is linear in $\mathbf{F}_{x|x^*w}$. Substituting the expression of $\mathbf{F}_{x^*|zw}$ into Eq. (3) results in

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

This equation implies k^2 restrictions on $k(k+1)$ unknowns in matrices $\mathbf{F}_{y|x^*w}$ and $\mathbf{F}_{x|x^*w}$. Since $\mathbf{F}_{y|x^*w}$ is linear in $\mathbf{F}_{x|x^*w}$, the result is a system of nonlinear equations containing the misclassification probabilities in the matrix $\mathbf{F}_{x|x^*w}$. Furthermore, there are additional k restrictions as follows:

$$\mathbf{F}_{x|x^*w} \times \mathbf{i} = \mathbf{i}. \tag{11}$$

Therefore, the unknowns $\mathbf{F}_{y|x^*w}$ and $\mathbf{F}_{x|x^*w}$ are determined by Eqs. (10) and (11). Solving the system of equations is not an easy task. A simple generalization of the identification in the binary case, such as in Mahajan (2006), 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 the complicated nonlinear system of equations. Finding the unique solution of this nonlinear system becomes manageable only when noticing that the problem can be phrased in terms of the matrix diagonalization introduced below.

I define

$$\mathbf{A} := \mathbf{F}^{-1}_{x|zw} \times \mathbf{F}_{yx|zw}. \tag{12}$$

Note that the matrix $\mathbf{A}$ is observed in the sample. From Eq. (10), I obtain

$$\mathbf{F}_{y|x^*w} = \mathbf{F}_{x|x^*w} \times \mathbf{A} \times \mathbf{F}^{-1}_{x|x^*w}. \tag{13}$$

Eq. (13) implies 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 probabilities simply consist of the eigenvectors. Eq. (13) also implies that the identification of the misclassification matrix rests on assumptions ruling out eigenvalues with multiplicity greater than 1. By the similarity property, the two matrices should

[1] One may also use other types of generalized matrix inverses in the identification and estimation. In that case, the identification strategy would be the same except the expression of the observed matrices. Estimators using different inverses would have the asymptotic properties similar to those presented in this paper.

[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.

have the same eigenvalues. Since $\mathbf{F}_{y|x^*w}$ is diagonal, its diagonal element $f_{y|x^*w}(y|j,w)$ 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 its 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 step of the identification can be summarized in the following equation:

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

where Q is an elementary matrix generated by interchanging rows of the identity matrix. The pair $(Q\mathbf{F}_{y|x^*w}Q^{-1}, Q\mathbf{F}_{x|x^*w})$ is observationally equivalent to $(\mathbf{F}_{y|x^*w}, \mathbf{F}_{x|x^*w})$, thus signifying the need for further restrictions to identify the model in $\mathbf{F}_{y|x^*w}$. As mentioned before, 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 sufficient to assume the following:

Assumption 2.3. There exists a function $\varpi(\cdot)$ such that $\mathrm{E}[\varpi(y)|x^* = i, w] \neq \mathrm{E}[\varpi(y)|x^* = j, w]$ for all $i \neq j$.

This assumption generalizes Assumption 5 in Mahajan (2006) with choices of a user-specified function $\varpi(\cdot)$. The reasoning behind Assumption 2.3 is as follows. Eq. (1) implies

$$\int \varpi(y) f_{yx|zw}(y, x|z, w)\,\mathrm{d}y = \sum_{x^*} \mathrm{E}[\varpi(y)|x^*, w] f_{x|x^*w}(x|x^*, w) f_{x^*|zw}(x^*|z, w).$$

Thus, if I replace $f_{y|x^*w}(y|j,w)$ with $\mathrm{E}[\varpi(y)|x^* = j, w]$ in $\mathbf{F}_{y|x^*w}$, and $f_{yx|zw}(y, i|j, w)$ with $\int \varpi(y) f_{yx|zw}(y, i|j, w)\,\mathrm{d}y$ in $\mathbf{F}_{yx|zw}$, the eigenvalue–eigenvector decomposition above still holds with $\mathrm{E}[\varpi(y)|x^* = j, w]$ as eigenvalues. But the eigenvectors in $\mathbf{F}_{x|x^*w}$ do not change. Therefore, Assumption 2.3 rules out duplicate eigenvalues, i.e., $\mathrm{E}[\varpi(y)|x^* = i, w] = \mathrm{E}[\varpi(y)|x^* = j, w]$ for $i \neq j$.

The next step is to find a condition that implies the ordering of the eigenvalues or of the eigenvectors. Such a condition is not unique, so I discuss various possibilities. First, if one can directly identify the ordering of the eigenvalues, the model is then identified. For example, suppose that $f_{y|x^*w}$ satisfies:

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

This assumption is not as strong as it looks. For example, I 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 $F(\cdot)$ is a c.d.f. Assumption 2.4 holds if and only if the sign of β is known. Furthermore, economic theory may suggest the sign of the coefficient in such an application as a study on the impact of education on labor supply. 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 the matrix diagonalization provides the matrix $\mathbf{F}_{y|x^*w}$ up to permutations of its diagonal entries. If the ordering of the diagonal entries is given, then $\mathbf{F}_{y|x^*w}$ is identified. I define $\lambda_j(\mathbf{A})$ for $j = 1, 2, \ldots, k$ as the eigenvalues of the matrix $\mathbf{A}$ with $\lambda_1(\mathbf{A}) < \lambda_2(\mathbf{A}) < \cdots < \lambda_k(\mathbf{A})$. Then the model is nonparametrically identified as follows:

$$f_{y|x^*w}(y|j, w) = \lambda_j(\mathbf{A}), \tag{15}$$

if $f_{y|x^*w}(y|x^*, w)$ is increasing in x^*. Assumption 2.4 allows the consideration of very general misclassification errors because the only restriction imposed on the misclassification probability matrix is its invertibility in Assumption 2.2.

Moreover, it is sufficient to find the ordering of the conditional expectation $\mathrm{E}[\varpi(y)|x^*, w]$ rather than that of the conditional density $f_{y|x^*w}$ itself. I make the following assumption:

Assumption 2.5. There exists a function $\varpi(\cdot)$ such that $\mathrm{E}[\varpi(y)|x^*, w]$ is strictly increasing in x^*.

As mentioned before, $\mathrm{E}[\varpi(y)|x^*, w]$ can also play the role of an eigenvalue. Therefore, if the ordering of $\mathrm{E}[\varpi(y)|x^*, w]$ in x^* is given, the ordering of the eigenvalues and eigenvectors is fixed and the model is then identified. A straightforward choice of the function $\varpi(\cdot)$ is $\varpi(y) = y$. This signifies that the conditional mean of y is monotonic in x^*, which is a reasonable assumption at least in a linear regression model. For example, let $\varpi(y) = y$ when considering a linear regression model $y = \beta x^* + \gamma w + \eta$. Assumption 2.5 holds if $\beta > 0$. Other choices of the user-specified function $\varpi(\cdot)$ are $\varpi(y) = (y - \mathrm{E}y)^2$, $\varpi(y) = 1(y \leqslant y_0)$, or $\varpi(y) = \delta(y - y_0)$ for some given y_0. As discussed above, one may use $\varpi(y) = \delta(y - y_0)$ when considering a probit model. Similarly, one may use $\varpi(y) = 1(y \leqslant y_0)$ for a Tobit model.

It is worth mentioning that the invertibility of $\mathbf{F}_{x|x^*w}$ is the only restriction imposed on the misclassification probabilities under Assumptions 2.4 or 2.5. Assumption 2.5 suggests that certain monotonicity properties of the latent model may lead to its identification without imposing restrictive assumptions on the misclassification error. This result is also new in the dichotomous case. For example, suppose one is interested in a probit model describing the impact of smoking on health with age as a covariate. Researchers are concerned about the misreporting error in the dichotomous reports of smoking behavior. If one imposes the restriction on the misclassification probabilities, one may require that the probability of reporting true smoking behavior is larger than one half (as in Assumption 2.7). Since age is a covariate, this restriction has to hold for people at all ages. However, such a restriction is believed to be very strong for teenagers. The probability of misreporting smoking behavior might be larger than one half because smoking is not legal at that age. Given that no studies show that smoking is healthy, the restriction on the latent model is more desirable.

There are also cases where the restrictions on the misclassification probability are more reasonable. For example, one might be interested in the impact of education on voting behavior. It is not immediately clear whether more educated people are more likely to vote or not. However, validation studies have shown that the misclassification error in self-reported education levels satisfies Assumption 2.7, which will be introduced later. Therefore, the restrictions on the misclassification probabilities are more desirable in this case.

Instead of Assumptions 2.4 and 2.5, one may impose different restrictions on the misclassification matrix $\mathbf{F}_{x|x^*w}$ to identify the model. For example, if the entries in one of its columns are strictly monotonic, the matrix $\mathbf{F}_{x|x^*w}$ is not observationally equivalent to $Q\mathbf{F}_{x|x^*w}$ for any $Q \neq I$. Without loss of generality, the following restriction can be imposed on the first column:

Assumption 2.6. $\Pr(x = 1|x^*, w)$ is strictly decreasing in x^* for $x^* \in \{1, 2, \ldots, k\}$.

In the 0–1 dichotomous case, this assumption implies that $\Pr(x = 0|x^* = 0, w) > \Pr(x = 0|x^* = 1, w)$, which is the same as

$$\Pr(x = 1|x^* = 0, w) + \Pr(x = 0|x^* = 1, w) < 1.$$

Therefore, Assumption 2.6 generalizes Assumption 2 in Mahajan (2006). For example, suppose there are three possible values of education levels, i.e., high school, college, and graduate school. Assumption 2.6 implies that people with a high school education are more likely to report the high school level than those with a college education, while people with a college education are more likely to report a high school level than those with a graduate school education. Under this assumption, the misclassification probability matrix $\mathbf{F}_{x|x^*w}$ and the model $\mathbf{F}_{y|x^*w}$ are identified. The exact expression of $\mathbf{F}_{y|x^*w}$ can be found by diagonalizing $\mathbf{A}$ to $S^{-1}\Lambda S$ with $S \times \mathbf{i} = \mathbf{i}$ and then using an elementary matrix Q to find the right QS satisfying Assumption 2.6 in

$$\mathbf{A} = (QS)^{-1}[Q\Lambda Q^{-1}]QS. \tag{16}$$

The matrix $\mathbf{F}_{y|x^*w}$ equals the diagonal matrix on the right-hand side

$$\mathbf{F}_{y|x^*w} = Q\Lambda Q^{-1}. \tag{17}$$

The latent model $f_{y|x^*w}$ is therefore identified.

A fourth alternative assumption for achieving identification is the following:

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

The intuition of this assumption is that the probability of reporting the true value is higher than that of reporting other values. This assumption is consistent with most validation studies. For example, the misclassification probability matrix of education with three possible values found in a validation study by Kane et al. (1999, Table 1) satisfies this assumption. As summarized in Bound et al. (2001, Table 5, pp. 3794–3797), at least four validation studies show the misclassification probability matrix of employment status with three possible values. All of these misclassification probability matrices satisfy Assumption 2.7. Notice that this assumption is different from Assumption 2.6 in this paper and Assumption 2 in Mahajan

(2006). Assumption 2.7 imposes restrictions on the conditional density $f_{x|x^*w}(\cdot|i,w)$ for each $i \in \{1,2,\ldots,k\}$, while Assumption 2.6 imposes restrictions on k conditional probabilities $f_{x|x^*w}(1|\cdot,w)$ simultaneously. Assumption 2.7 implies Assumption 2.6 only in the dichotomous case. In the 0–1 dichotomous case, Assumption 2.7 implies that $\Pr(x=1|x^*=1,w)>0.5$ and $\Pr(x=0|x^*=0,w)>0.5$, which is different from Assumption 2 in Mahajan (2006).

The identification procedure under Assumption 2.7 is as follows. After an eigenvector (i.e., a row of $Q\mathbf{F}_{x|x^*w}$) is obtained from the diagonalization, one may identify the largest entry in that row. If it is the jth entry, that eigenvector is equal to the jth row of the matrix $\mathbf{F}_{x|x^*w}$. Note that Assumption 2.7 is weaker than the assumption that $\mathbf{F}_{x|x^*w}$ is strictly diagonally dominant (i.e., $\Pr(x=i|x^*=i,w)>0.5$ for all $i=1,2,\ldots,k$). 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 $Q\neq I$, and the model is identified.

The identification results are summarized as follows:

Theorem 1 (*Nonparametric identification*). *Suppose that Assumptions* 1, 2, 2.1–2.3, *and one of Assumptions* 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 show that the latent model $f_{y|x^*w}$ and the misclassification error distribution $f_{x|x^*w}$ are point-identified and directly estimable. I will show that the point identification leads to a "plug-in" semiparametric estimator.

2.1. *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,

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

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, I 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 $\mathbf{A}$ are observed in the sample. The characteristic polynomial of the matrix $\mathbf{A}$ is as follows:

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

where $\mathrm{m}(\mathbf{A})$ is the sum of all the two-by-two principal minors of the matrix $\mathbf{A}$.[3] By Eq. (13), the matrix $\mathbf{A}$ has three real eigenvalues so that the cubic equation $p(t)=0$ has three different real roots as follows:

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

[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.

where

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

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

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

As shown before, the eigenvalues satisfy

$$\begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} = \begin{pmatrix} f_{y|x^*w}(y|\widetilde{i},w) & 0 & 0 \\ 0 & f_{y|x^*w}(y|\widetilde{j},w) & 0 \\ 0 & 0 & f_{y|x^*w}(y|\widetilde{k},w) \end{pmatrix},$$

where the set $\{\widetilde{i},\widetilde{j},\widetilde{k}\} = \{1,2,3\}$. If one can identify the values of the indices $\widetilde{i},\widetilde{j}$, and $\widetilde{k}$, the latent density $f_{y|x^*w}$ is identified. In the case where $f_{yx|zw}(y,i|j,w)$ is replaced with $\int \varpi(y) f_{yx|zw}(y,i|j,w)\,\mathrm{d}y$ in the matrix $\mathbf{F}_{yx|zw}$, the following equality results:

$$\begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} = \begin{pmatrix} \mathrm{E}[\varpi(y)|x^* = \widetilde{i},w] & 0 & 0 \\ 0 & \mathrm{E}[\varpi(y)|x^* = \widetilde{j},w] & 0 \\ 0 & 0 & \mathrm{E}[\varpi(y)|x^* = \widetilde{k},w] \end{pmatrix}. \tag{18}$$

The second step is to obtain the left eigenvector corresponding to each eigenvalue λ_i $(i = 1,2,3)$. I define the eigenvectors as $v_i = (v_{i1}, v_{i2}, v_{i3})$ satisfying $v_{i1} + v_{i2} + v_{i3} = 1$. By Eq. (13), each v_i corresponds to a row of the matrix $\mathbf{F}_{x|x^*w}$. The result is

$$\lambda_i \times v_i = v_i \times \mathbf{A},$$

which implies

$$E_i \times v_i^{\mathrm{T}} = e$$

with $E_i = (\mathbf{A} - \lambda_i I,\ \mathbf{i})^{\mathrm{T}}$ and $e = (0,\ 0,\ 0,\ 1)^{\mathrm{T}}$. Let E_i^+ be the Moore–Penrose matrix inverse of E_i.[4] The eigenvector v_i can be found as follows:

$$v_i = (E_i^+ \times e)^{\mathrm{T}}.$$

The matrix of the eigenvectors is as follows:

$$V = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \equiv \begin{pmatrix} f_{x|x^*w}(1|\widetilde{i},w) & f_{x|x^*w}(2|\widetilde{i},w) & f_{x|x^*w}(3|\widetilde{i},w) \\ f_{x|x^*w}(1|\widetilde{j},w) & f_{x|x^*w}(2|\widetilde{j},w) & f_{x|x^*w}(3|\widetilde{j},w) \\ f_{x|x^*w}(1|\widetilde{k},w) & f_{x|x^*w}(2|\widetilde{k},w) & f_{x|x^*w}(3|\widetilde{k},w) \end{pmatrix}.$$

The set of indices $\{\widetilde{i},\widetilde{j},\widetilde{k}\}$ is equal to $\{1,2,3\}$. The matrix $\mathbf{F}_{x|x^*w}$ is identified if the values of $\widetilde{i},\widetilde{j}$, and $\widetilde{k}$ can be identified. In fact, each of Assumptions 2.4–2.7 may lead to such identification. I discuss each of these assumptions in what follows.

Assumptions 2.4 or 2.5 imply the ordering of the eigenvalues directly. Suppose that the ordering is $\lambda_2 > \lambda_1 > \lambda_3$. Assumption 2.5 then implies that $\widetilde{i} = 2$, $\widetilde{j} = 3$, and $\widetilde{k} = 1$ in Eq. (18). The advantage of Assumptions 2.4 or 2.5 is that $\mathbf{F}_{x|x^*w}$ is identified without further restrictions.

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

Assumption 2.6 implies that the misclassification probability matrix $\mathbf{F}_{x|x^*w}$ can be identified by interchanging the rows v_1, v_2, v_3 in V. In other words, Assumption 2.6 implies that

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

where Q is an elementary matrix. For example, suppose the value of three entries $f_{x|x^*w}(1|\widetilde{j}, w)$, $f_{x|x^*w}(1|\widetilde{i}, w)$, and $f_{x|x^*w}(1|\widetilde{k}, w)$ satisfying $f_{x|x^*w}(1|\widetilde{j}, w) > f_{x|x^*w}(1|\widetilde{i}, w) > f_{x|x^*w}(1|\widetilde{k}, w)$. Since Assumption 2.6 implies $f_{x|x^*w}(1|1, w) > f_{x|x^*w}(1|2, w) > f_{x|x^*w}(1|3, w)$, the result is $\widetilde{i} = 2$, $\widetilde{j} = 1$, and $\widetilde{k} = 3$. That means v_1 and v_2 need to be interchanged in the matrix V to obtain $\mathbf{F}_{x|x^*w}$ with

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

Assumption 2.7 may also lead to such an elementary matrix Q. This assumption implies that all the rows of V have a unique largest entry in different columns. Suppose that the data show that these entries are $f_{x|x^*w}(3|\widetilde{i}, w)$, $f_{x|x^*w}(2|\widetilde{j}, w)$ $f_{x|x^*w}(1|\widetilde{k}, w)$. Then the rows in V need to be interchanged such that these entries are on the diagonal with

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

The result is that $\widetilde{i} = 3$, $\widetilde{j} = 2$, and $\widetilde{k} = 1$.

With the matrix Q identified, the latent model $\mathbf{F}_{y|x^*w}$ is achieved by interchanging the eigenvalues on the diagonal correspondingly as follows:

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

The last step is to find $\mathbf{F}_{x^*|zw}$ through Eq. (8) as follows:

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

In summary, I have shown the explicit expressions of $f_{y|x^*w}$, $f_{x|x^*w}$, and $f_{x^*|zw}$ as functions of $f_{yx|zw}$, such that

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

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

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

Although the explicit expression of these functions for a general k is expected to be complicated, a general result in Andrew et al. (1993) shows that these functions are in fact analytic. Moreover, it is not necessary to use these explicit expressions in the computation. Most statistical software packages can easily compute eigenvalues and eigenvectors of a given matrix. 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. Such a relationship makes the identification very clear.

3. Estimation

This section focuses on the following parametric conditional moment model:

$$\mathrm{E}(y|x^*, w) = m^*(x^*, w; \theta_0), \tag{19}$$

where m^* is a known moment function and θ_0 is an unknown parameter of interest. The variables y, x, z, and w are observed in an i.i.d. sample. Since x^* is not observed, one can estimate the parameter θ_0 through an observed moment

$$\mathrm{E}(y|z,w) = m(z,w;\theta_0), \tag{20}$$

with

$$m(z,w;\theta_0) = \sum_{x^*} m^*(x^*,w;\theta_0) f_{x^*|zw}(x^*|z,w). \tag{21}$$

As discussed in Assumptions 2.3 and 2.5, the identification results still hold with $f_{yx|wz}(y,x|w,z)$ and $f_{y|x^*w}(y|x^*,w)$ replaced by $\int \varpi(y) f_{yxwz}(y,x|w,z)\,\mathrm{d}y$ and $\mathrm{E}[\varpi(y)|x^*,w]$ for a known function $\varpi(\cdot)$. In this section, let $\varpi(y)=y$ in Assumptions 2.3 and 2.5. The identification results imply that one can express the unknown density $f_{x^*|zw}$ as follows:

$$f_{x^*|zw}(x^*|z,w) = \psi(x^*,\gamma_0(z,w)), \tag{22}$$

with a known function ψ, $\gamma_0(z,w) = [\gamma_{10}(z,w), \gamma_{20}(z,w), \gamma_{30}(z,w)]^{\mathrm{T}}$, and

$$\gamma_{10}(z,w) = \left(\int \varpi(y) f_{yxwz}(y,1,w,z)\,\mathrm{d}y, \ldots, \int \varpi(y) f_{yxwz}(y,k,w,z)\,\mathrm{d}y\right),$$
$$\gamma_{20}(z,w) = (f_{xwz}(1,w,z), \ldots, f_{xwz}(k,w,z)),$$
$$\gamma_{30}(z,w) = f_{wz}(w,z).$$

Although the expression of the function ψ is complicated, a general result in Andrew et al. (1993) shows that the function ψ is a well-behaved analytic function around γ_0.

The next step is to show that the parameter θ_0 is identifiable in the observed model $\mathrm{E}(y|z,w) = m(z,w;\theta_0)$ if and only if the parameter θ_0 is identifiable in the latent model $\mathrm{E}(y|x^*,w) = m^*(x^*,w;\theta_0)$. Define two vectors:

$$\mathbf{M}(\theta) = (m(z,w;\theta)|_{z=1}, m(z,w;\theta)|_{z=2}, \ldots, m(z,w;\theta)|_{z=k})^{\mathrm{T}},$$
$$\mathbf{M}^*(\theta) = (m^*(x^*,w;\theta)|_{x^*=1}, m^*(x^*,w;\theta)|_{x^*=2}, \ldots, m^*(x^*,w;\theta)|_{x^*=k})^{\mathrm{T}}.$$

Therefore,

$$\mathbf{M}(\theta) = \mathbf{F}_{x^*|zw} \times \mathbf{M}^*(\theta). \tag{23}$$

Suppose θ_0 is not identifiable so that there exists θ_1 which is observationally equivalent to θ_0 in the sense that $\mathbf{M}(\theta_1) - \mathbf{M}(\theta_0) = 0$. Since $\mathbf{F}_{x^*|zw}$ is identified and has rank k, one 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 2 (*Parametric identification*). *Suppose that Assumptions* 1, 2, 2.1–2.3 *and one of Assumptions* 2.4–2.7 *are satisfied. The parameter θ_0 is identifiable in the observed model* (20) *if and only if it is identifiable in the latent model* (19).

Given that the density function $f_{x^*|zw}$ can be explicitly expressed as a function of the observed density f_{yxwz}, I propose a "plug-in" semiparametric estimator. Although this estimator may not be the most efficient one, I make the estimator as simple as possible. By Eqs. (21) and (22), the observed model can be written as follows:

$$\mathrm{E}\{[y - m(z,w;\theta_0,\gamma_0(z,w))]|z,w\} = 0, \tag{24}$$

with

$$m(z,w;\theta_0,\gamma_0(z,w)) = \sum_{x^*} m^*(x^*,w;\theta_0)\psi(x^*,\gamma_0(z,w))$$

and the nuisance function $\gamma_0(z,w) = [\gamma_{10}(z,w), \gamma_{20}(z,w), \gamma_{30}(z,w)]^{\mathrm{T}}$. Since the function ψ in Eq. (22) is known, the moment function $m(z,w;\theta_0,\gamma_0)$ is known up to the parameter θ_0 and the nuisance function γ_0. Note that the moment function depends on the value of the nuisance function at observed points through the known function ψ, while a general GMM model may consider a moment function depending on the entire nuisance function.

Since the joint distribution of $\{y, x, w, z\}$ is observed in the sample, the nuisance parameter γ_0 can be estimated nonparametrically as follows:

$$\widehat{\gamma}(z, w) = [\widehat{\gamma}_1(z, w), \widehat{\gamma}_2(z, w), \widehat{\gamma}_3(z, w)]^{\mathrm{T}},$$

where

$$\widehat{\gamma}_1(z, w) = \left(\int \varpi(y)\widehat{f}_{yxwz}(y, 1, w, z)\,\mathrm{d}y, \ldots, \int \varpi(y)\widehat{f}_{yxwz}(y, k, w, z)\,\mathrm{d}y\right),$$

$$\widehat{\gamma}_2(z, w) = (\widehat{f}_{xwz}(1, w, z), \ldots, \widehat{f}_{xwz}(k, w, z)),$$

$$\widehat{\gamma}_3(z, w) = \widehat{f}_{wz}(w, z),$$

and

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

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

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

The constant r is the dimension of w. The function $I(\cdot)$ is an indicator function, and the function $K(\cdot)$ is a known kernel function with bandwidth h. If the dependent variable y is discrete, one should apply the indicator function to y rather than the kernel function.

When I replace γ_0 by $\widehat{\gamma}$ in the known function ψ, I must guarantee that the moment function $m(z, w; \theta_0, \widehat{\gamma}(z, w))$ is well-behaved. For this reason, I consider a weighted moment function

$$\widetilde{m}(y, z, w; \theta_0, \gamma_0(z, w)) = \tau(z, w)[y - m(z, w; \theta_0, \gamma_0(z, w))]$$

with a weight function $\tau(z, w)$. The moment condition then becomes

$$\mathrm{E}[\widetilde{m}(y, z, w; \theta_0, \gamma_0(z, w))|z, w] = 0. \tag{25}$$

This condition holds for any weight function $\tau(z, w)$.

The next step is to choose a desirable $\tau(z, w)$ such that the weighted moment function $\widetilde{m}(y, z, w; \theta_0, \widehat{\gamma}(z, w))$ is well-behaved. Let $\mathscr{S}$ be the compact support of (z, w) and Γ be a closed subset of the range of $\gamma_0(\cdot, \cdot)$. Since $0 \leqslant f_{x^*|zw}(x^*|z, w) \leqslant 1$ in Eq. (22) for any $(z, w) \in \mathscr{S}$, the identification results guarantee that $0 \leqslant \psi(x^*, r) \leqslant 1$ for all $r \in \Gamma$. Moreover, ψ is an analytic function so that there exists an open set $\overline{\Gamma}$ such that (i) $-\infty < \psi(x^*, r) < \infty$ for all $r \in \overline{\Gamma}$ and (ii) Γ is a subset of $\overline{\Gamma}$. The existence of $\overline{\Gamma}$ guarantees that the function $\psi(x^*, \cdot)$ is well-behaved even on the boundary of Γ. Notice that the set Γ does not depend on the function ψ. The problem arises when $\widehat{\gamma}(z, w) \notin \Gamma$ for some (z, w) so that $\psi(x^*, \widehat{\gamma}(z, w))$ might be unbounded. I adopt the fixed trimming technique to solve this problem. By the uniform convergence of $\widehat{\gamma}$ to γ_0, there exists a fixed closed set $\mathscr{S}^* \subset int(\mathscr{S})$ such that $\widehat{\gamma}(z, w) \in \Gamma$ and $\gamma_0(z, w) \in \Gamma$ for all $(z, w) \in \mathscr{S}^*$. Therefore, if $\tau(z, w) = 0$ for all $(z, w) \notin \mathscr{S}^*$, the weighted moment function $\widetilde{m}(y, z, w; \theta_0, \widehat{\gamma}(z, w))$ is well-behaved.

Since the weighted moment condition is used to estimate θ_0, an extra identification assumption is needed for the parameter θ_0 to be identified by the weighted moment condition (25). It will be introduced later as Assumption 4.1(ii). This assumption requires that (z, w) is informative enough on the set $\mathscr{S}^*$ to identify the unknown parameter θ_0. It is true that the latent moment function $m^*(x^*, w)$ may not be nonparametrically identified after using the trimming function $\tau(z, w)$. However, this extra identification assumption implies that the identification of the parameter θ_0 is still feasible after trimming. Although a different weight function $\tau(z, w)$ may affect the efficiency, the estimator for θ_0 based on the weighted moment condition (25) may still be consistent and asymptotically normal.

I then estimate the unknown parameter of interest, $\theta_0 \in \Theta$, through a "plug-in" semiparametric estimator in which I replace γ_0 with its nonparametric estimator. The semiparametric estimator $\widehat{\theta}$ is defined as follows:

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} \tau(z_i, w_i)[y_i - m(z_i, w_i; \theta, \widehat{\gamma}(z_i, w_i))]^2. \tag{26}$$

As discussed in Newey (1994b), the weight function τ is used to focus on regions where $m(z_i, w_i; \theta, \widehat{\gamma})$ provides reasonable estimates. As discussed above, the set $\mathscr{S}^*$ can be used as the fixed trimming set. A simple weight function may be $\tau(z, w) = I_{\mathscr{S}^*}(z, w)$. As Hardle and Marron (1990) discuss, fixed trimming avoids the boundary effects associated with kernel density estimators. Aside from the fixed trimming function, there are other types of weight functions. For example, Fan and Li (1996) propose a data-dependent weight function. This paper chooses to focus on the fixed set $\mathscr{S}^*$ in order to avoid estimating the conditional density $f_{x^*|zw}$ in the area where the estimate of the conditional density is not very accurate. Fixed trimming also makes the theoretical proof relatively convenient.

Another issue in the estimation is that the estimated eigenvalues and eigenvectors may not always be real. A 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. Additional Monte Carlo experiments, which are not included in this paper, imply that the misspecification of the model always causes a significant bias in the real parts of the estimates but does not necessarily cause the imaginary parts to be significantly different from zero. In other words, the fact that the imaginary parts are significantly different from zero implies that the model may be misspecified. However, the fact that the imaginary parts are close to zero does not necessarily mean that the model is correctly specified. The reason behind this phenomenon is as follows. The matrix $\mathbf{A}$ is always real even when the model is misspecified. That means the characteristic polynomial $p(t)$ of $\mathbf{A}$ always has real coefficients. Although the zeros of a real polynomial may not necessarily be real, it is still possible that the roots of $p(t) = 0$, i.e., the eigenvalues of $\mathbf{A}$, are all real under certain misspecification of the model. In that case, the imaginary parts of the estimates would be close to zero while the real parts are biased. In the case where some roots of $p(t) = 0$ are not real due to misspecification, both the real parts and the imaginary parts would be biased.

3.1. Consistency

In order to show the consistency of the estimator $\widehat{\theta}$, I first show the uniform convergence of $\widehat{\gamma}$. The kernel density estimator used in $\widehat{\gamma}$ has been studied extensively. Let $\omega := (y, x, w, z)$ and $\widehat{\gamma} = (\widehat{\gamma}_1, \widehat{\gamma}_2, \widehat{\gamma}_3)^{\mathrm{T}}$. Define the norm $\|\cdot\|_\infty$ as

$$\|\widehat{\gamma} - \gamma_0\|_\infty = \sup_{\omega\in\mathscr{W}} |\widehat{\gamma}_1 - \gamma_{10}| + \sup_{\omega\in\mathscr{W}} |\widehat{\gamma}_2 - \gamma_{20}| + \sup_{\omega\in\mathscr{W}} |\widehat{\gamma}_3 - \gamma_{30}|.$$

The following results come from Newey (1992):

Lemma 2. *Suppose*:

(Assumption 3.1) *$\omega \in \mathscr{W}$ and $\mathscr{W}$ is a compact set.*

(Assumption 3.2) *$\gamma_0(\omega)$ is continuously differentiable to order d with bounded derivatives on an open set containing $\mathscr{W}$.*

(Assumption 3.3) *$K(u)$ is differentiable of order d, and the derivatives of order d are bounded. $K(u)$ is zero outside a bounded set. $\int_{-\infty}^{\infty} K(x)\,\mathrm{d}x = 1$, and there is a positive integer m such that for all $j<m$, $\int_{-\infty}^{\infty} K(u)u^j\,\mathrm{d}u = 0$. And the characteristic function of K is absolutely integrable.*

(Assumption 3.4) *$h \to 0$ and $nh^r \to \infty$, as $n \to \infty$.*

Then

$$\|\widehat{\gamma} - \gamma_0\|_\infty = \mathrm{O_p}[(\ln n)^{1/2}(nh^{r+2d})^{-1/2} + h^m]. \tag{27}$$

The next step is to show the uniform convergence of the estimated density $\widehat{f}_{x^*|zw} = \psi(x^*, \widehat{\gamma})$. Consider the function γ_0 as a mapping from $\mathscr{W} \cap \mathscr{S}^*$ to a set Γ. For any given $x^* = j$, the function $\psi(j, \cdot)$ is a known and nonstochastic function which does not depend on ω. From Eqs. (3) and (4), one can show

$$\mathbf{F}_{\varpi(y)x|zw} \times \mathbf{F}_{x|zw}^{-1} = \mathbf{F}_{x^*|zw} \times \mathbf{F}_{\varpi(y)|x^*w} \times \mathbf{F}_{x^*|zw}^{-1},$$

where $\mathbf{F}_{\varpi(y)x|zw} = [\int \varpi(y) f_{yx|zw}(y, j|i, w)\,\mathrm{d}y]_{i,j}$ and $\mathbf{F}_{\varpi(y)|x^*w}$ is a diagonal matrix with the ith diagonal element equal to $\mathrm{E}[\varpi(y)|x^* = i, w]$. I define $\mathbf{B} \equiv \mathbf{F}_{\varpi(y)x|zw} \times \mathbf{F}_{x|zw}^{-1}$. This means a column in $\mathbf{F}_{x^*|zw}$ is an eigenvector of $\mathbf{B}$.

In order to derive the derivative of the eigenvalues and eigenvectors, I define $\gamma_0^v = (vec(\mathbf{F}_{\varpi(y)xzw})^{\mathrm{T}}, vec(\mathbf{F}_{xzw})^{\mathrm{T}}, \mathbf{F}_{zw}^{\mathrm{T}})^{\mathrm{T}}$ where $vec(M)$ denotes the vector formed by collecting the entries of the matrix M in a single vector, and

$$\mathbf{F}_{\varpi(y)xzw} = \begin{pmatrix} \int \varpi(y) f_{yxzw}(y, 1, 1, w)\,\mathrm{d}y & \ldots & \int \varpi(y) f_{yxzw}(y, k, 1, w)\,\mathrm{d}y \\ \vdots & \vdots & \vdots \\ \int \varpi(y) f_{yxzw}(y, 1, k, w)\,\mathrm{d}y & \ldots & \int \varpi(y) f_{yxzw}(y, k, k, w)\,\mathrm{d}y \end{pmatrix},$$

$$\mathbf{F}_{xzw} = \begin{pmatrix} f_{xzw}(y, 1, 1, w) & \ldots & f_{xzw}(y, k, 1, w) \\ \vdots & \vdots & \vdots \\ f_{xzw}(y, 1, k, w) & \ldots & f_{xzw}(y, k, k, w) \end{pmatrix},$$

$$\mathbf{F}_{zw} = (f_{zw}(1, w) \;\ldots\; f_{zw}(k, w))^{\mathrm{T}}.$$

Note that the vector γ_0^v contains the same information as γ_0. Similarly, I define $\widehat{\gamma}^v$ as the vector version of $\widehat{\gamma}$.

Now consider $k \times k$ matrix-valued functions $\mathbf{B}(\gamma^v)$ where γ^v is a vector of arguments. The eigenvalues $\lambda(\gamma^v)$ and eigenvectors $\boldsymbol{\psi}(\gamma^v)$ of $\mathbf{B}$ satisfy

$$[\mathbf{B}(\gamma^v) - \lambda(\gamma^v)\mathbf{I}]\boldsymbol{\psi}(\gamma^v) = 0,$$

where $\mathbf{I}$ is an identity matrix. In this case, an eigenvalue $\lambda(\gamma_0^v)$ equals $f_{y|x^*w}(y|x_j^*, w)$ and its corresponding eigenvector is

$$\begin{aligned} \boldsymbol{\psi}(\gamma_0^v) &= [\psi(x_j^*, \gamma_0^v), \ldots, \psi(x_j^*, \gamma_0^v)]^{\mathrm{T}} \\ &= [f_{x^*|zw}(x_j^*|1, w), \ldots, f_{x^*|zw}(x_j^*|k, w)]^{\mathrm{T}} \end{aligned}$$

for a given index x_j^* of the eigenvalues. The derivatives of eigenvalues and eigenvectors of matrix functions have been studied thoroughly. Here, I use a general result from Andrew et al. (1993). Their Theorem 2.1 shows that there is a neighborhood $\mathscr{N}_0$ of γ_0^v on which there exists an eigenvalue function $\lambda(\gamma^v)$ and an eigenvector function $\boldsymbol{\psi}(\gamma^v)$ that are all analytic functions of γ^v. I define $\widetilde{\Gamma} = \{v : v = \gamma^v(\omega) \text{ and } \gamma^v \in \mathscr{N}_0\}$, which is the union of the range of all γ^v near γ_0^v. Thus, the range of γ_0^v is a subset of $\widetilde{\Gamma}$ and the function $\boldsymbol{\psi}(\cdot)$ is analytic on $\widetilde{\Gamma}$.

Now recall the concept of the pathwise derivative. I assume the existence of a continuous path $\{\gamma^v(t) : t \in [0, 1]\}$ such that $\gamma^v(0) = \gamma_0^v$ and $\gamma^v(1) = \gamma^v$. When γ^v is close enough to γ_0^v, the linear combination $(1-t)\gamma_0^v + t\gamma^v$ is in $\mathscr{N}_0$ and its range is a subset of $\widetilde{\Gamma}$. Therefore, $\boldsymbol{\psi}((1-t)\gamma_0^v + t\gamma^v)$ is continuously differentiable at $t = 0$. The pathwise derivative of $\boldsymbol{\psi}(\cdot)$ evaluated at $\gamma^v - \gamma_0^v$ can be defined as

$$\frac{\mathrm{d}\boldsymbol{\psi}(\gamma_0^v)}{\mathrm{d}\gamma^v}[\gamma^v - \gamma_0^v] \equiv \left.\frac{\mathrm{d}\boldsymbol{\psi}((1-t)\gamma_0^v + t\gamma^v)}{\mathrm{d}t}\right|_{t=0} \tag{28}$$

almost everywhere (under the probability measure of ω). The pathwise derivative is a linear functional that approximates $\boldsymbol{\psi}(\gamma^v)$ in the neighborhood of γ_0^v (i.e., for small values of $\gamma^v - \gamma_0^v$). Notice that the nonstochastic analytic function $\boldsymbol{\psi}$ only depends on the values of the nuisance function γ^v at observed points instead of the entire function. The pathwise derivative can be expressed as the ordinary derivative through

$$\frac{\mathrm{d}\boldsymbol{\psi}((1-t)\gamma_0^v+t\gamma^v)}{\mathrm{d}t}=\frac{\partial\boldsymbol{\psi}((1-t)\gamma_0^v+t\gamma^v)}{\partial(\gamma^v)^{\mathrm{T}}}\times(\gamma^v-\gamma_0^v).$$

I can then use the results in Lemma 2 to show that the function $\boldsymbol{\psi}(\widehat{\gamma}^v)$ converges uniformly with $\widehat{\gamma}$.

Let $\mathbf{1}$ equal a vector of ones $(1,\ldots,1)^{\mathrm{T}}$, $\mathbf{I}$ denote an identity matrix, and M^+ stand for the Moore–Penrose generalized inverse of the matrix M. Let $(\gamma^v)_j$ denote the jth entry of vector γ^v and $\|\cdot\|_1$ denote the L_1 norm or the sum norm. When $\gamma^v=\gamma_0^v$, $\boldsymbol{\psi}(\gamma_0^v)$ is a column of $\mathbf{F}_{x^*|zw}$. Define $\mu(\gamma_0^v)$ as a column of the matrix $(\mathbf{F}_{x^*|zw}^{-1})^{\mathrm{T}}$, which is actually the right eigenvector corresponding to $\lambda(\gamma_0^v)$. The result is summarized in the following lemma:

Lemma 3. *Suppose that assumptions in Lemma 2 hold, and that there exists a constant $c>0$ such that $\det(\mathbf{F}_{x|zw})\geqslant c$ and $f_{zw}\geqslant c$. Then for some $\varepsilon\to 0$ as $n\to\infty$*

$$\sup_{\|\gamma-\gamma_0\|_\infty\leqslant\varepsilon}\|\boldsymbol{\psi}(\gamma^v)-\boldsymbol{\psi}(\gamma_0^v)\|_1=\mathrm{O}(\|\gamma-\gamma_0\|_\infty).$$

Moreover,

$$\sup_{\|\gamma-\gamma_0\|_\infty\leqslant\varepsilon}\left\|\boldsymbol{\psi}(\gamma^v)-\boldsymbol{\psi}(\gamma_0^v)-\frac{\partial\boldsymbol{\psi}(\gamma_0^v)}{\partial(\gamma^v)^{\mathrm{T}}}(\gamma^v-\gamma_0^v)\right\|_1=\mathrm{O}(\|\gamma-\gamma_0\|_\infty^2),$$

where

$$\frac{\partial\boldsymbol{\psi}(\gamma_0^v)}{\partial(\gamma^v)^{\mathrm{T}}}=\left(\frac{\partial\boldsymbol{\psi}(\gamma_0^v)}{\partial(\gamma^v)_1},\frac{\partial\boldsymbol{\psi}(\gamma_0^v)}{\partial(\gamma^v)_2},\ldots,\frac{\partial\boldsymbol{\psi}(\gamma_0^v)}{\partial(\gamma^v)_{2k^2+k}}\right),$$

$$\frac{\partial\boldsymbol{\psi}(\gamma_0^v)}{\partial(\gamma^v)_j}=-(\mathbf{I}-\boldsymbol{\psi}(\gamma_0^v)\mathbf{1}^{\mathrm{T}})[\mathbf{B}(\gamma_0^v)-\lambda(\gamma_0^v)\mathbf{I}]^+\left(\frac{\partial\mathbf{B}(\gamma_0^v)}{\partial(\gamma^v)_j}-\frac{\partial\lambda(\gamma_0^v)}{\partial(\gamma^v)_j}\mathbf{I}\right)\boldsymbol{\psi}(\gamma_0^v),$$

and

$$\frac{\partial\lambda(\gamma_0^v)}{\partial(\gamma^v)_j}=\mu(\gamma_0^v)^{\mathrm{T}}\frac{\partial\mathbf{B}(\gamma_0^v)}{\partial(\gamma^v)_j}\boldsymbol{\psi}(\gamma_0^v).$$

Proof. See the appendix.

The semiparametric estimator in this paper may be considered as an application of the general semiparametric estimator in Section 8.3 of Newey and McFadden (1994). I will therefore make similar assumptions and just provide a brief discussion as they have been covered in that handbook chapter. Define the score function as

$$g(\omega,\theta,\gamma)=\tau(z,w)[y-m(z,w;\theta,\gamma)]\frac{\mathrm{d}}{\mathrm{d}\theta}m(z,w;\theta,\gamma),$$

where $\omega{:=}(y,x,w,z)$. Let $\|\cdot\|_2$ stand for the L_2 norm. In order to guarantee the consistency of the estimator, I make the following assumptions:

Assumption 4.1. (i) Assumptions in Theorem 2, Lemma 2, and Lemma 3 hold and $\theta_0\in\Theta$, where Θ is compact; (ii) θ_0 is identifiable from the weighted moment condition in Eq. (25).

Assumption 4.2. $m^*(x^*,w;\theta)$ is continuously differentiable in θ for all w and is a measurable function of x^* and w for all $\theta\in\Theta$.

Assumption 4.3. There is $d(\omega)$ with $\|g(\omega,\theta,\gamma_0)\|_2\leqslant d(\omega)$, $\|\tau(z,w)m^*(x^*,w;\theta)\|_2\leqslant d(\omega)$, and $\|\tau(z,w)m^*(x^*,w_i;\theta)[y-m(z,w;\theta,\gamma_0)]\|_2\leqslant d(\omega)$ for all $\theta\in\Theta$ and all $x^*\in\{1,2,\ldots,k\}$ such that $\mathrm{E}[d(\omega)]<\infty$.

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

The consistency of this estimator is summarized in the following theorem:

Theorem 3 (*Consistency*). *Suppose that Assumptions* 4.1–4.4 *are satisfied. Then,*

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

Proof. See the appendix.

3.2. Asymptotic normality

I now show the asymptotic distribution of the estimator $\widehat{\theta}$ that solves

$$\frac{1}{n}\sum_{i=1}^{n} g(\omega_i, \widehat{\theta}, \widehat{\gamma}) = 0.$$

The standard delta method leads to

$$-\frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}} g(\omega_i, \widetilde{\theta}, \widehat{\gamma})(\widehat{\theta} - \theta_0) = \frac{1}{n}\sum_{i=1}^{n} g(\omega_i, \theta_0, \widehat{\gamma}),$$

where $\widetilde{\theta}$ is an intermediate value between $\widehat{\theta}$ and θ_0. To obtain the asymptotic normality of $\hat{\theta}$, I first show that the right-hand side equals

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(\omega_i, \theta_0, \widehat{\gamma}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} [g(\omega_i, \theta_0, \gamma_0) + \delta(\omega_i)] + \mathrm{o_p}(1), \tag{29}$$

where

$$\delta(\omega) = v(\omega) - \mathrm{E}v(\omega),$$

$$v(\widetilde{\omega}) = \mathrm{E}\left[\tau(z_i, w_i)\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbf{s}(y_i, z_i, w_i; \theta_0, \gamma)\frac{\partial \boldsymbol{\psi}(\gamma^v)}{\partial (\gamma^v)^{\mathrm{T}}}\mathbf{1}\Big|_{\gamma^v = \gamma_0^v(\widetilde{\omega})} \Bigg| \widetilde{\omega} \right],$$

with the vector $\mathbf{1}$ having the same dimension as γ^v, and

$$\mathbf{s}(y_i, z_i, w_i; \theta, \gamma_0) = \begin{pmatrix} (y_i - m(z_i, w_i; \theta, \gamma_0))m^*(x^*_{j_1}, w_i; \theta) \\ \vdots \\ (y_i - m(z_i, w_i; \theta, \gamma_0))m^*(x^*_{j_k}, w_i; \theta) \end{pmatrix}^{\mathrm{T}},$$

with the ordering of $x^*_{j_1}, \ldots, x^*_{j_k}$ being the same as that of x^* in the vector $\boldsymbol{\psi}(\gamma^v)$.

The correction term $\delta(\omega_i)$ in Eq. (29) is due to the nonparametric estimation of γ_0. The formula of $v(\omega)$ is derived from the linearization of $g(\omega, \theta_0, \gamma)$ with respect to γ. The expression of this correction term is consistent with the results in Newey (1994a). The first term on the right-hand side of Eq. (29) converges to a normal distribution by the standard central limit theorem. This result is summarized in the following lemma:

Lemma 4. *Suppose that assumptions in Theorem* 3 *are satisfied, and*:

(Assumption 5.1) $m^*(x^*, w; \theta)$ *is continuously differentiable of order* 5.
(Assumption 5.2) $\mathrm{E}[\|g(\omega, \theta_0, \gamma_0) + \delta(\omega)\|_2] < \infty$.
(Assumption 5.3) *There is* $d(\omega)$ *with*

$$\left\| \tau(z, w)m^*(x^*_1, w; \theta_0)\frac{\mathrm{d}}{\mathrm{d}\theta}m^*(x^*_2, w; \theta_0) \right\|_1 \leqslant d(\omega),$$

and

$$\left\|\tau(z,w)\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbf{s}(y,z,w;\theta_0,\gamma_0)\right\|_1 \leqslant d(\omega)$$

for all x_1^*, $x_2^* \in \{1,2,\dots,k\}$ *such that* $\mathrm{E}[d^2(\omega)] < \infty$.

(Assumption 5.4) $\sqrt{n}h^{2m} \to 0$, *and* $\sqrt{n}\ln n/(nh^{r+2d}) \to 0$ *as* $n \to \infty$.

Then,

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

where

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

Proof. See the appendix.

The last step is to show the asymptotic normality of the estimator $\hat{\theta}$ through the delta method. The results are summarized as follows:

Theorem 4 (*Asymptotic normality*). *Suppose that assumptions in Theorem* 3 *and Lemma* 4 *hold*, *and*:

(Assumption 5.5) $\theta_0 \in interior(\Theta)$, *and* $\mathrm{E}[\|g(\omega,\theta_0,\gamma_0)\|_2^2] < \infty$.

(Assumption 5.6) *There is* $d(\omega)$ *with*

$$\left\|\tau(z,w)m^*(x^*,w;\theta)\frac{\mathrm{d}^2}{\mathrm{d}\theta\,\mathrm{d}\theta^{\mathrm{T}}}m(z,w;\theta,\gamma_0)\right\|_2 \leqslant d(\omega),$$

$$\left\|\tau(z,w)[y-m(z,w;\theta,\gamma_0)]\frac{\mathrm{d}^2}{\mathrm{d}\theta\,\mathrm{d}\theta^{\mathrm{T}}}m^*(x^*,w;\theta)\right\|_2 \leqslant d(\omega),$$

$$\left\|\tau(z,w)\frac{\mathrm{d}}{\mathrm{d}\theta_j}m^*(x_1^*,w;\widetilde{\theta})\frac{\mathrm{d}^2}{\mathrm{d}\theta\,\mathrm{d}\theta^{\mathrm{T}}}m^*(x_2^*,w;\theta)\right\|_2 \leqslant d(\omega),$$

$$\left\|\tau(z,w)\frac{\mathrm{d}}{\mathrm{d}\theta_j}m^*(x_1^*,w;\theta)\frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}}m(z,w;\theta,\gamma_0)\right\|_2 \leqslant d(\omega),$$

and

$$\left\|\tau(z,w)\frac{\mathrm{d}}{\mathrm{d}\theta_j}m^*(x_1^*,w;\theta)\frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}}m^*(x_2^*,w;\theta)\right\|_2 \leqslant d(\omega)$$

for all $\theta \in \Theta$, *all* x_1^*, $x_2^* \in \{1,2,\dots,k\}$, *and all* $1 \leqslant j \leqslant \dim(\theta)$ *such that* $\mathrm{E}[d(\omega)] < \infty$.

(Assumption 5.7) $\mathrm{E}[\nabla_\theta g(\omega,\theta_0,\gamma_0)]$ *exists and is nonsingular.*

Then,

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

where

$$G_\theta = \mathrm{E}[\nabla_\theta g(\omega,\theta_0,\gamma_0)].$$

Proof. See the appendix.

Since I have the explicit expression of the asymptotic variance, a consistent estimator of the asymptotic variance can be constructed by substituting estimates for true values in $G_\theta^{-1}\Omega G_\theta^{-1\prime}$. However, the estimator of the correction term $\delta(\omega)$ may be very complicated and difficult to compute. As suggested by Newey (1994b, equation 7), an alternative way to compute the correction term is to use the numerical differentiation of the influence function instead of using the explicit expression. The detailed estimation of the asymptotic variance may be found in Section 3 in Newey (1994b). I also implement that variance estimator in the empirical illustration.

4. Simulation

This section applies the estimator developed above to a probit model with a mismeasured 0–1 dichotomous explanatory variable and to a nonlinear regression model containing a discrete regressor with four possible values. The conditional density function of the probit model is

$$\begin{aligned} &f^*(y|x^*, w; \theta) = P(x^*, w; \theta)^y (1 - P(x^*, w; \theta))^{1-y}, \\ &P(x^*, w; \theta) = \Phi(\beta_0 + \beta_1 x^* + \beta_2 w), \end{aligned} \tag{30}$$

where $\theta = (\beta_0, \beta_1, \beta_2)^{\mathrm{T}}$ and Φ is the standard normal c.d.f. I consider three estimators. The first is the ML probit estimator that uses the mismeasured variable x in the sample as if it were accurate. That is, it ignores the misclassification error. This ML estimator is not consistent. The second estimator is the infeasible ML probit estimator that uses the latent true x^*. This estimator is consistent and has the smallest asymptotic variance of all the estimators considered here. The third estimator is the semi-parametric MLE developed above which uses the instrumental variable. For each estimator, I report the Root Mean Squared Error (RMSE), the average bias, and the standard deviation of the estimates over the replications. One 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 should dominate the MSE of that estimator. The semiparametric IV estimator corrects the bias, but its variance should be larger than that of the second estimator. Since the last two estimators are consistent, their variance should dominate their MSE.

Table 1 shows that the MLE that ignores the misclassification error is significantly biased as expected. The bias of the estimated coefficient on the mismeasured independent variable is larger than the biases of other

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

	β_1			β_2			β_0		
	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.
$p = 0.3$									
$q = 0.2$									
Ignoring meas. error	0.541	−0.523	0.139	0.181	−0.103	0.149	0.293	0.277	0.095
True x^*	0.160	0.015	0.159	0.166	−0.013	0.165	0.104	0.000	0.104
I.V.	0.421	−0.095	0.410	0.307	−0.087	0.295	0.263	0.045	0.259
$p = 0.3 - 0.1w$									
$q = 0.2 + 0.1w$									
Ignoring meas. error	0.538	−0.520	0.136	0.210	−0.150	0.147	0.290	0.275	0.094
True x^*	0.157	0.012	0.157	0.165	−0.011	0.164	0.104	0.000	0.104
I.V.	0.409	−0.124	0.389	0.332	−0.138	0.302	0.238	0.061	0.230
$p = 0.3 + 0.1w$									
$q = 0.2 + 0.1w$									
Ignoring meas. error	0.509	−0.491	0.137	0.176	−0.094	0.149	0.279	0.263	0.093
True x^*	0.160	0.015	0.159	0.165	−0.014	0.165	0.104	−0.001	0.104
I.V.	0.318	−0.108	0.299	0.307	−0.071	0.298	0.205	0.052	0.198

Note: (1) $\beta_1 = 1$, $\beta_2 = 1$, $\beta_0 = 0.5$, $x^* = I(\varepsilon < 0.6)$; $z = I(\varepsilon + \delta < 0.6)$, $\varepsilon \sim \mathrm{Uniform}(0, 1)$, $\delta \sim \mathrm{N}(0, 0.04)$, $(\rho_{x^*z} \approx 0.67)$, $w \sim \mathrm{N}(0, 0.25)$.
(2) $\Pr(x = 0|x^* = 1, w) = \min(1, \max(0, p)), \Pr(x = 1|x^* = 0, w) = \min(1, \max(0, q))$.
(3) $K(x) = 0.5(3 - x^2)\phi(x)$ and $h = 0.2$, where $\phi(x)$ is the standard normal density.

estimated coefficients. The biases of the new semiparametric estimator are smaller than those of the estimator with misclassification error ignored. In all cases, the MSE of the infeasible MLE is much smaller than that of the other two estimators. The semiparametric estimator performs well with different specifications of the misclassification error distribution.

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

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

The covariate w and the regression error u have a standard normal distribution. The true values 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 share the same support $\{1, 2, 3, 4\}$. The marginal distributions of x^* and z are $P_{x^*} = (0.2, 0.3, 0.3, 0.2)$, $P_z = (0.3, 0.2, 0.3, 0.2)$, where $P_v{:=}(P(v = 1), P(v = 2), P(v = 3), P(v = 4))$ for a random variable v. The variable x^* is generated as follows:

$$x^* = P_{x^*}(\eta_{x^*}) \equiv \begin{cases} 1 & \text{if } \eta_{x^*} \leqslant P(x^* = 1), \\ 2 & \text{if } P(x^* = 1) < \eta_{x^*} \leqslant P(x^* \leqslant 2), \\ 3 & \text{if } P(x^* \leqslant 2) < \eta_{x^*} \leqslant P(x^* \leqslant 3), \\ 4 & \text{if } P(x^* \leqslant 3) < \eta_{x^*} \leqslant P(x^* \leqslant 4), \end{cases}$$

where η_{x^*} is uniformly distributed on $[0, 1]$ and is independent of all other variables. I abuse the notation P_{x^*} to express x^* as a function of the random variable η_{x^*} in $x^* = P_{x^*}(\eta_{x^*})$. Similarly, I define $z = P_z(0.6\eta_{x^*} + 0.4\eta_z)$, where η_z is another independent random variable with a uniform distribution on $[0, 1]$. The correlation between x^* and z is caused by the common random variable η_{x^*}.

I consider three specifications of the misclassification error distribution in the matrix $\mathbf{F}_{x|x^*w}$. The first specification uses the constant misclassification probabilities as follows:

$$\mathbf{F}_{x|x^*w} = \mathbf{F}_{x|x^*} = \begin{pmatrix} 0.6 & 0.2 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.7 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.7 \end{pmatrix}.$$

The matrix $\mathbf{F}_{x|x^*}$ is defined as the matrix $\mathbf{F}_{x|x^*w}$ with $f_{x|x^*w}(x|x^*, w) = \Pr(x|x^*)$ in each entry. The matrix $\mathbf{F}_{x|x^*w}$ above 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 $[0, 1]$ as follows:

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

I abuse the notation again to express x as a function of η_x in $x = \mathbf{F}_{x|x^*}(\eta_x)$. In the other two specifications, I consider the correlation between the misclassification error and the covariate w as $x = \mathbf{F}_{x|x^*}(0.9\eta_x + 0.1\Phi(w))$ and $x = \mathbf{F}_{x|x^*}(0.9\eta_x + 0.1(1 - \Phi(w)))$, where Φ is the cumulative distribution function of w.

The simulation results in Table 2 contain three estimators similar to those in Table 1. The first estimator is a nonlinear least squares (NLS) estimator using the misclassified variable x as if it were the true value x^*. In other words, the first estimator ignores the misclassification error in x. The second one uses the accurate data without misclassification errors. The last estimator is the semiparametric IV estimator developed in this paper. As expected, the simulation results in Table 2 show that the first estimator has a larger MSE than the second estimator using accurate data, because the misclassification errors cause significant biases. The third estimator has a smaller MSE than the first one. Moreover, the developed estimator effectively reduces the bias. The semiparametric IV estimator performs better when the misclassification error is correlated with other explanatory variables. This is because the semiparametric IV estimator treats the misclassification

Table 2
Simulation results of the nonlinear regression model: sample size 500; number of repetitions 100

	β_1			β_2			β_0		
	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.	Root MSE	Mean bias	Std. dev.
$x = F_{x\|x^*w}(\eta)$									
$\eta = \eta_x$									
Ignoring meas. error	0.5662	−0.5435	0.1587	0.2090	0.0371	0.2057	1.2081	1.1032	0.4924
True x^*	0.0285	0.0003	0.0285	0.0180	−0.0023	0.0178	0.0463	−0.0063	0.0458
I.V.	0.3504	−0.0119	0.3502	0.1561	−0.0722	0.1384	0.5860	−0.0539	0.5835
$x = F_{x\|x^*w}(\eta)$									
$\eta = 0.9\eta_x + 0.1\Phi(w)$									
Ignoring meas. error	0.4690	−0.4495	0.1336	0.2103	0.0102	0.2101	1.0057	0.9095	0.4291
True x^*	0.0285	0.0003	0.0285	0.0180	−0.0023	0.0178	0.0463	−0.0063	0.0458
I.V.	0.3164	−0.0268	0.3152	0.1527	−0.0789	0.1307	0.5162	−0.0242	0.5157
$x = F_{x\|x^*w}(\eta)$									
$\eta = 0.9\eta_x + 0.1[1 - \Phi(w)]$									
Ignoring meas. error	0.4189	−0.3907	0.1511	0.1975	0.0582	0.1888	0.8712	0.7436	0.4539
True x^*	0.0285	0.0003	0.0285	0.0180	−0.0023	0.0178	0.0463	−0.0063	0.0458
I.V.	0.3077	−0.0824	0.2965	0.1583	−0.0728	0.1405	0.5171	0.0612	0.5171

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 $\widehat{\beta}_1$ on the latent variable x^*, but also leads to a significant bias in the estimated constant term $\widehat{\beta}_0$. The simulation results show that the semiparametric IV estimator also reduces the bias in $\widehat{\beta}_0$.

In summary, the semiparametric IV estimator performs well in the finite sample. The new estimator successfully reduces the bias caused by the misclassification error. And the simulation results are also consistent with the asymptotic properties of the estimator.

5. Empirical illustration

This section applies the developed estimator to a count data model to investigate the impact of education on women's fertility. Since the dependent variable (number of children) takes on discrete values for a nontrivial fraction of the population, I directly model the expectation of the dependent variable y conditional on the explanatory variables x^* and w as $\mathrm{E}(y|x^*, w) = m^*(x^*, w)$. A detailed discussion of count data models can be found in Wooldridge (2002, p. 645). I use the exponential function, a popular functional form, as follows:

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

The coefficient β_1 or β_2 is related to the semielasticity of $\mathrm{E}(y|x^*, w)$ with respect to x^* or w. For small changes Δx^*, the percentage change in the conditional mean $\mathrm{E}(y|x^*, w)$ is roughly $100\beta_1\Delta x^*$. Since the true education level of each individual is subject to misreporting error,[5] I use the parents' education level as the instrumental variable to estimate the parameter of interest $\beta = (\beta_0, \beta_1, \beta_2)^{\mathrm{T}}$. Notice that Assumptions 2.2 and 2.7 are consistent with the misclassification probability matrix found in Kane et al. (1999). The father's self-reported education level can be treated as a repeated measurement of the individual's education level. As in other measurement error studies using repeated measurements, such as Li (2002) and Schennach (2004), the repeated measurement is not required to be accurate. As long as it is independent of the individual's self-reported

[5] For example, see Kane et al. (1999, Table 1).

education level and fertility conditional on the true education level, the father's self-reported education level may play the role of the instrumental variable.

As discussed above, the parameters are estimated through the moment condition

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

One can apply a NLS estimator to this count data model. The NLS estimator is consistent but inefficient because the discrete distribution of the count data implies heteroskedasticity. I instead use the Poisson regression model, which is common 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:

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

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. In this situation, there are two possible estimators. First, one can ignore the measurement error and use x_i as x_i^* in the likelihood $l_i(y_i, x_i^*, w_i; \beta)$. Second, one can use the $m(z,w)$ as a conditional mean of y on z and w, and implement a QMLE with a likelihood function as follows:

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

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

The population considered here is composed of women who have left school but still live with their parents. The dependent variable is the number of children for a given woman. 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, education has three categories: high school education or lower, some college education, and college education or higher. Years of education assigned to each category are 9, 14, and 16, respectively. The joint distribution of women's and their parents' education level is shown in Table 3. More than half of the women in the sample did 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 4 contains the descriptive statistics of other variables. There were 53% of the women having no children, 27% having one child, and 20% having two or more. About 80% of the women in the sample were employed and about 20% were black. The median age was 22, and the first and the third quartiles were 19 and 25. Marital status is not considered in the model because less than 1.6% of the 1,688 women in the sample were married.

I assume the misclassification error in a woman's education level is independent of her parents' education level and the number of her children conditional on her true education level, employment status, age, and race. And the misclassification probability is assumed to satisfy Assumptions 2.6 and 2.7. Assumption 2.7 implies that people are more willing to tell the truth than to lie, conditional on their education, employment status,

Table 3
Joint distribution of education (1688 observations)

Education	Parents' education			
	High school or lower	Some college	College or higher	Total
High school or lower	0.361	0.134	0.072	0.568
Some college	0.111	0.092	0.050	0.254
College or higher	0.065	0.039	0.075	0.179
Total	0.537	0.265	0.198	1

Table 4
Summary statistics of variables (1688 observations)

	Mean	Std. dev.	Min	Max
Number of children	0.790	1.092	0	9
Employment (yes = 1)	0.799	0.401	0	1
Age	22.982	6.443	15	56
Race (white = 1)	0.798	0.401	0	1

Table 5
NLS estimation results

	Ignoring meas. error		I.V.	
	Estimate	Std. dev.	Estimate	Std. dev.
Education	−0.0188	0.0177	−0.0539	0.0223
Employment	−0.0145	0.0621	−0.0077	0.0638
Age	−0.3494	0.0284	−0.3632	0.0254
Age^2/100	0.4482	0.0536	0.4660	0.0508
Race	−0.0222	0.0808	−0.0092	0.0823
Constant	5.2249	0.3067	5.8884	0.4076

Table 6
QMLE estimation results

	Ignoring meas. error		I.V.	
	Estimate	Std. dev.	Estimate	Std. dev.
Education	−0.0264	0.0143	−0.0541	0.0244
Employment	−0.0220	0.0636	−0.0065	0.0638
Age	−0.3665	0.0240	−0.3900	0.0220
Age^2/100	0.4979	0.0454	0.5272	0.0438
Race	0.0710	0.0811	0.0658	0.0815
Constant	5.3654	0.2809	6.0178	0.4226

age, and race. The advantage of the estimator developed in this paper is that it allows the misclassification 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 education levels. Suppose the error is independent of other explanatory variables except for 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 all the other explanatory variables are included, 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, I can nonparametrically identify these unknown functions and parameters of interest and use a "plug-in" semiparametric estimator to estimate them. The asymptotic variance is estimated using Theorem 8.13 and equation 8.18 in Newey and McFadden (1994), with more details in Newey (1994b).

Table 5 contains two NLS estimates, and Table 6 shows the two QMLE estimates. The second and third columns of the tables contain the estimates and their estimated standard deviations when the misclassification error is ignored. The estimates of the developed estimator and their standard deviations are shown in the last two columns. When the misclassification error is ignored, both the NLS estimator and the QMLE estimator are inconsistent, and the NLS estimator also has the problem of heteroskedasticity. The two semiparametric IV estimators both are consistent, and the QMLE one should have the correct estimate of asymptotic standard deviations. The most interesting parameter in the model is the coefficient on education. If the misclassification error is ignored, the estimates are biased toward zero when compared with the estimates using the instrumental variable.

The results show that the impact of education on women's fertility is more significant than commonly thought. If one ignores 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. Its interpretation is that 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.

Based on the results in Table 6, one can conduct a test similar to the Hausman test with the null hypothesis that there are no misclassification errors in education levels. The test statistics is $(\widehat{\beta}_{\text{ie}} - \widehat{\beta}_{\text{iv}})^{\text{T}} V^{-1} (\widehat{\beta}_{\text{ie}} - \widehat{\beta}_{\text{iv}}) \sim \mathscr{X}_6^2$, where $\widehat{\beta}_{\text{ie}}$ is the estimator with error ignored, $\widehat{\beta}_{\text{iv}}$ is the IV estimator, and V is the variance–covariance matrix of $(\widehat{\beta}_{\text{ie}} - \widehat{\beta}_{\text{iv}})$. In this empirical illustration, the test statistic equals 15.85 with p-value 0.0146. Therefore, the null hypothesis is rejected at usual significance levels.

In summary, this simple empirical example illustrates that the new estimator performs well with real data.

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. In this case, one may estimate the latent model directly as eigenvalues of an observed matrix without considering the misclassification probability. 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 nonparametric identification in this paper directly leads to a nonparametric or semiparametric estimator.

Acknowledgments

I would like to thank Stephen Donald, Qi Li, Geert Ridder, Maxwell Stinchcombe, seminar audiences at New York University, at Johns Hopkins University, at the Econometric Society 2005 World Congress and at Texas Econometrics Camp, as well as the two anonymous referees for helpful comments and suggestions. I especially thank Xiaohong Chen for her helpful comments in the second revision and Susanne Schennach for her helpful suggestions in the first two revisions of this paper. All errors are mine.

Appendix

Proof of Lemma 1. First, I show Eqs. (1) and (3). The law of total probability implies

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

where

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

By Assumptions 1 and 2, the equation above becomes

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

Therefore, Eq. (1) holds as follows:

$$f_{yx|zw}(y,x|z,w) = \sum_{x^*} f_{y|x^*w}(y|x^*,w) f_{x|x^*w}(x|x^*,w) f_{x^*|zw}(x^*|z,w). \tag{31}$$

The next step is to directly show Eq. (3), i.e.,

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

Here, I only show a detailed proof for $k = 2$, which can be directly generalized to the general discrete case. The right-hand side of Eq. (3) is

$$\begin{aligned}
&\mathbf{F}_{x^*|zw} \times \mathbf{F}_{y|x^*w} \times \mathbf{F}_{x|x^*w} \\
&\quad = \begin{pmatrix} f_{x^*|zw}(1|1,w) & f_{x^*|zw}(2|1,w) \\ f_{x^*|zw}(1|2,w) & f_{x^*|zw}(2|2,w) \end{pmatrix} \times \begin{pmatrix} f_{y|x^*w}(y|1,w) & 0 \\ 0 & f_{y|x^*w}(y|2,w) \end{pmatrix} \\
&\qquad \times \begin{pmatrix} f_{x|x^*w}(1|1,w) & f_{x|x^*w}(2|1,w) \\ f_{x|x^*w}(1|2,w) & f_{x|x^*w}(2|2,w) \end{pmatrix} \\
&\quad = \begin{pmatrix} f_{x^*|zw}(1|1,w) & f_{x^*|zw}(2|1,w) \\ f_{x^*|zw}(1|2,w) & f_{x^*|zw}(2|2,w) \end{pmatrix} \times \begin{pmatrix} f_{y|x^*w}(y|1,w) f_{x|x^*w}(1|1,w) & f_{y|x^*w}(y|1,w) f_{x|x^*w}(2|1,w) \\ f_{y|x^*w}(y|2,w) f_{x|x^*w}(1|2,w) & f_{y|x^*w}(y|2,w) f_{x|x^*w}(2|2,w) \end{pmatrix}.
\end{aligned}$$

By Assumptions 1 and 2,

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

and therefore,

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

Again by Assumptions 1 and 2,

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

and then

$$\begin{aligned}
&\mathbf{F}_{x^*|zw} \times \mathbf{F}_{y|x^*w} \times \mathbf{F}_{x|x^*w} \\
&\quad = \begin{pmatrix} f_{yxx^*|zw}(y,1,1|1,w) + f_{yxx^*|zw}(y,1,2|1,w) & f_{yxx^*|zw}(y,2,1|1,w) + f_{yxx^*|zw}(y,2,2|1,w) \\ f_{yxx^*|zw}(y,1,1|2,w) + f_{yxx^*|zw}(y,1,2|2,w) & f_{yxx^*|zw}(y,2,1|2,w) + f_{yxx^*|zw}(y,2,2|2,w) \end{pmatrix}.
\end{aligned} \tag{32}$$

Since

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

the result is

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

Therefore, Eq. (3) holds.

Second, I show Eqs. (2) and (4). Integrating y out in Eq. (31) results in Eq. (2) as follows:

$$f_{x|zw}(x|z,w) = \sum_{x^*} f_{x|x^*w}(x|x^*,w) f_{x^*|zw|}(x^*|z,w). \tag{33}$$

The next step is to show that Eq. (4), i.e.,

$$\mathbf{F}_{x|zw} = \mathbf{F}_{x^*|zw} \times \mathbf{F}_{x|x^*w}, \tag{34}$$

is equivalent to Eq. (33). The right-hand side of Eq. (34) is

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

By Assumptions 1 and 2,

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

and then

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

Since

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

the result is

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

Therefore, Eq. (4) holds. It is straightforward to show that these results still hold for a general k. □

Proof of Lemma 3. This proof uses a general result in Andrew et al. (1993). Consider $k \times k$ matrix-valued functions $\mathbf{B}(\gamma^v)$ where γ^v is a vector of arguments. The eigenvalues $\lambda(\gamma^v)$ and eigenvectors $\boldsymbol{\psi}(\gamma^v)$ of $\mathbf{B}$ satisfy

$$[\mathbf{B}(\gamma^v) - \lambda(\gamma^v)\mathbf{I}]\boldsymbol{\psi}(\gamma^v) = 0.$$

Theorem 2.1 assumes that the values of γ^v and $\lambda(\gamma^v)$ belong to sets $\mathcal{N}$ and $\mathcal{N}_\lambda$, such that (i) the elements of $\mathbf{B}(\gamma^v)$ are an analytic function of γ^v; (ii) for each value of γ^v in $\mathcal{N}$ there is a value of λ in $\mathcal{N}_\lambda$ such that $\det(\mathbf{B}(\gamma^v) - \lambda\mathbf{I}) \neq 0$. In this paper, the function $\mathbf{B}(\gamma^v)$ is known as $\mathbf{B} \equiv \mathbf{F}_{yx|zw} \times \mathbf{F}^{-1}_{x|zw}$, therefore, such sets $\mathcal{N}$ and $\mathcal{N}_\lambda$ exist. Moreover, condition (i) holds with $\gamma^v = \gamma^v_0$ by the definition of the matrix $\mathbf{B}$ under the assumption that $\det(\mathbf{F}_{x|zw})$ and f_{zw} are bounded away from zero. By their Theorem 3.2, the fact that $\mathbf{B}(\gamma^v)$ has distinctive eigenvalues at $\gamma^v = \gamma^v_0$ implies that $[\mathbf{B}(\gamma^v) - \lambda(\gamma^v)\mathbf{I}]$ has a so-called simple eigenvalue at $\gamma^v = \gamma^v_0$. Therefore, their Theorem 2.1 holds and implies that there is a neighborhood $\mathcal{N}_0$ of γ^v_0 on which there exists an eigenvalue function $\lambda(\gamma^v)$ and eigenvector functions $\boldsymbol{\psi}(\gamma^v)$ that are all analytic functions of

γ^v. I define $\widetilde{\Gamma} = \{v : v = \gamma^v(\omega) \text{ and } \gamma^v \in \mathcal{N}_0\}$, which is the union of the range of all γ^v near γ_0^v. Thus, the function $\boldsymbol{\psi}(\cdot)$ is analytic on $\widetilde{\Gamma}$. Notice that the neighborhood $\mathcal{N}_0$ and the set $\widetilde{\Gamma}$ do not change with the sample size n.

Consider γ^v, γ_0^v and assume the existence of a continuous path $\{\gamma^v(t) : t \in [0,1]\}$ such that $\gamma^v(0) = \gamma_0^v$ and $\gamma^v(1) = \gamma^v$. Given the uniform convergence of $\widehat{\gamma}$ to γ_0, one may only consider γ^v close enough to γ_0^v. When γ^v is close enough to γ_0^v, i.e., $\|\gamma - \gamma_0\|_\infty \leqslant \varepsilon$ for a $\varepsilon \to 0$ as $n \to \infty$, the range of $(1-t)\gamma_0^v + t\gamma^v$ is a subset of $\widetilde{\Gamma}$. Therefore, $\boldsymbol{\psi}((1-t)\gamma_0^v + t\gamma^v)$ is continuously differentiable (actually analytic) at $t = 0$. The pathwise derivative of $\boldsymbol{\psi}(\cdot)$ evaluated at $\gamma^v - \gamma_0^v$ can be defined as

$$\frac{\mathrm{d}\boldsymbol{\psi}(\gamma_0^v)}{\mathrm{d}\gamma^v}[\gamma^v - \gamma_0^v] \equiv \left.\frac{\mathrm{d}\boldsymbol{\psi}((1-t)\gamma_0^v + t\gamma^v)}{\mathrm{d}t}\right|_{t=0} \tag{35}$$

almost everywhere (under the probability measure of ω). Notice that the nonstochastic analytic function $\boldsymbol{\psi}$ only depends on the values of the nuisance function γ^v at observed points instead of the entire function. The pathwise derivative can be expressed as the ordinary derivative through

$$\frac{\mathrm{d}\boldsymbol{\psi}((1-t)\gamma_0^v + t\gamma^v)}{\mathrm{d}t} = \frac{\partial\boldsymbol{\psi}((1-t)\gamma_0^v + t\gamma^v)}{\partial(\gamma^v)^{\mathrm{T}}} \times (\gamma^v - \gamma_0^v).$$

This is a linear functional that approximates $\boldsymbol{\psi}(\gamma^v)$ in the neighborhood of γ_0^v, i.e., for small values of $\gamma^v - \gamma_0^v$. Therefore,

$$\boldsymbol{\psi}(\gamma^v) - \boldsymbol{\psi}(\gamma_0^v) = \frac{\mathrm{d}\boldsymbol{\psi}((1-\widetilde{t})\gamma_0^v + \widetilde{t}\gamma^v)}{\mathrm{d}\gamma^v}[\gamma^v - \gamma_0^v] \tag{36}$$

for some $\widetilde{t} \in [0,1]$.

Let $(\gamma^v)_j$ denote the jth entry of vector γ^v and $\|\cdot\|_1$ denote the L_1 norm or the sum norm. Section 12 in Andrew et al. (1993) shows that

$$\frac{\partial\lambda(\gamma_0^v)}{\partial(\gamma^v)_j} = \mu(\gamma_0^v)^{\mathrm{T}} \frac{\partial\mathbf{B}(\gamma_0^v)}{\partial(\gamma^v)_j}\boldsymbol{\psi}(\gamma_0^v),$$

where $\mu(\gamma^v)$ is the right eigenvector corresponding to $\lambda(\gamma^v)$. When $\gamma^v = \gamma_0^v$, $\boldsymbol{\psi}(\gamma_0^v)$ is a column of $\mathbf{F}_{x^*|zw}$ and $\mu(\gamma_0^v)$ is a column of $(\mathbf{F}_{x^*|zw}^{-1})^{\mathrm{T}}$. The derivative of the eigenvector is

$$\frac{\partial\boldsymbol{\psi}(\gamma_0^v)}{\partial(\gamma^v)_j} = -(\mathbf{I} - \boldsymbol{\psi}(\gamma_0^v)\mathbf{1}^{\mathrm{T}})[\mathbf{B}(\gamma_0^v) - \lambda(\gamma_0^v)\mathbf{I}]^{+}\left(\frac{\partial\mathbf{B}(\gamma_0^v)}{\partial(\gamma^v)_j} - \frac{\partial\lambda(\gamma_0^v)}{\partial(\gamma^v)_j}\mathbf{I}\right)\boldsymbol{\psi}(\gamma_0^v),$$

where $\mathbf{1} = (1,\ldots,1)^{\mathrm{T}}$ and M^{+} stands for the Moore–Penrose generalized inverse of the matrix M. Therefore, Eq. (36) becomes

$$\begin{aligned}\boldsymbol{\psi}(\gamma^v) - \boldsymbol{\psi}(\gamma_0^v) &= \frac{\mathrm{d}\boldsymbol{\psi}((1-\widetilde{t})\gamma_0^v + \widetilde{t}\gamma^v)}{\mathrm{d}\gamma^v}[\gamma^v - \gamma_0^v] \\ &= \sum_j \frac{\partial\boldsymbol{\psi}((1-\widetilde{t})\gamma_0^v + \widetilde{t}\gamma^v)}{\partial(\gamma^v)_j}(\gamma_j^v - \gamma_{0j}^v) \\ &\leqslant \sum_j \left\|\frac{\partial\boldsymbol{\psi}((1-\widetilde{t})\gamma_0^v + \widetilde{t}\gamma^v)}{\partial(\gamma^v)_j}\right\|_1 |\gamma_j^v - \gamma_{0j}^v|.\end{aligned} \tag{37}$$

Since the range of $(1-\widetilde{t})\gamma_0^v + \widetilde{t}\gamma^v$ in Eq. (37) is a subset of $\widetilde{\Gamma}$ and $\boldsymbol{\psi}(\cdot)$ is analytic on $\widetilde{\Gamma}$, the term $\left\|\frac{\partial\boldsymbol{\psi}((1-\widetilde{t})\gamma_0^v + \widetilde{t}\gamma^v)}{\partial(\gamma^v)_j}\right\|_1$ is bounded and the last term in Eq. (37) is $\mathrm{O}(\|\gamma - \gamma_0\|_\infty)$. Therefore, the eigenvectors converge as follows:

$$\sup_{\|\gamma-\gamma_0\|_\infty \leqslant \varepsilon} \|\boldsymbol{\psi}(\gamma^v) - \boldsymbol{\psi}(\gamma_0^v)\|_1 = \mathrm{O}(\|\gamma - \gamma_0\|_\infty).$$

Now consider the linearization of $\boldsymbol{\psi}(\gamma^v)$ with respect to γ^v. The pathwise derivative gives

$$\boldsymbol{\psi}(\gamma^v) - \boldsymbol{\psi}(\gamma_0^v) = \frac{\partial \boldsymbol{\psi}(\gamma_0^v)}{\partial (\gamma^v)^{\mathrm{T}}}(\gamma^v - \gamma_0^v) + \sum_j \sum_l \frac{\partial^2 \boldsymbol{\psi}((1-\widetilde{t})\gamma_0^v + \widetilde{t}\gamma^v)}{\partial(\gamma^v)_j d(\gamma^v)_l}(\gamma_j^v - \gamma_{0j}^v)(\gamma_l^v - \gamma_{0l}^v), \tag{38}$$

where $\widetilde{t} \in [0,1]$. The explicit expression of the second-order derivative $\frac{\partial^2 \boldsymbol{\psi}((1-\widetilde{t})\gamma_0^v + \widetilde{t}\gamma^v)}{\partial(\gamma^v)_j \partial(\gamma^v)_l}$ is complicated but may still be derived. Although I do not show the explicit expression, the same reasoning as in Eq. (37) holds. Since $\boldsymbol{\psi}(\cdot)$ is analytic, its higher-order derivatives are all well defined and bounded. Therefore, the last term on the right-hand side of Eq. (38) is $\mathrm{O}(\|\gamma - \gamma_0\|_\infty^2)$. The result then is

$$\sup_{\|\gamma-\gamma_0\|_\infty \leqslant \varepsilon} \left\| \boldsymbol{\psi}(\gamma^v) - \boldsymbol{\psi}(\gamma_0^v) - \frac{\partial \boldsymbol{\psi}(\gamma_0^v)}{\partial(\gamma^v)^{\mathrm{T}}}(\gamma^v - \gamma_0^v) \right\|_1 = \mathrm{O}(\|\gamma - \gamma_0\|_\infty^2). \qquad \square$$

Proof of Theorem 3 (*Consistency*). Define

$$Q_n(\theta, \widehat{\gamma}) = \frac{1}{n}\sum_{i=1}^n \tau(z_i, w_i)[y_i - m(z_i, w_i; \theta, \widehat{\gamma})]^2$$

and

$$Q_0(\theta, \gamma_0) = \mathrm{E}\{\tau(z_i, w_i)[y_i - m(z_i, w_i; \theta, \gamma_0)]^2\}.$$

I first show that $\sup_{\theta\in\Theta} |Q_n(\theta, \widehat{\gamma}) - Q_0(\theta, \gamma_0)| = \mathrm{o_p}(1)$. The left-hand side is bounded as follows:

$$\sup_{\theta\in\Theta} |Q_n(\theta, \widehat{\gamma}) - Q_0(\theta, \gamma_0)| \leqslant \sup_{\theta\in\Theta} |Q_n(\theta, \widehat{\gamma}) - Q_n(\theta, \gamma_0)| + \sup_{\theta\in\Theta} |Q_n(\theta, \gamma_0) - Q_0(\theta, \gamma_0)|. \tag{39}$$

By Lemma 2.4 in Newey and McFadden (1994) and Assumptions 4.2 and 4.3, the second term on the right-hand side is negligible, i.e.,

$$\sup_{\theta\in\Theta} |Q_n(\theta, \gamma_0) - Q_0(\theta, \gamma_0)| = \mathrm{o_p}(1). \tag{40}$$

I consider the first term on the right-hand side of Eq. (39) as follows:

$$Q_n(\theta, \widehat{\gamma}) - Q_n(\theta, \gamma_0) = \frac{1}{n}\sum_{i=1}^n \tau(z_i, w_i)\{[y_i - m(z_i, w_i; \theta, \widehat{\gamma})]^2 - [y_i - m(z_i, w_i; \theta, \gamma_0)]^2\}.$$

Using the identity

$$\widehat{a}^2 - a^2 = (\widehat{a} - a)^2 + 2a(\widehat{a} - a).$$

I obtain

$$\begin{aligned} Q_n(\theta, \widehat{\gamma}) - Q_n(\theta, \gamma_0) = {} & \frac{1}{n}\sum_{i=1}^n \tau(z_i, w_i)[m(z_i, w_i; \theta, \widehat{\gamma}) - m(z_i, w_i; \theta, \gamma_0)]^2 \\ & - \frac{1}{n}\sum_{i=1}^n \tau(z_i, w_i)2[y_i - m(z_i, w_i; \theta, \gamma_0)][m(z_i, w_i; \theta, \widehat{\gamma}) - m(z_i, w_i; \theta, \gamma_0)] \\ = {} & \frac{1}{n}\sum_{i=1}^n \tau(z_i, w_i)\left(\sum_{x^*} m^*(x^*, w_i; \theta)[\psi(x^*, \widehat{\gamma}) - \psi(x^*, \gamma_0)]\right)^2 \\ & - \frac{1}{n}\sum_{i=1}^n \tau(z_i, w_i)2[y_i - m(z_i, w_i; \theta, \gamma_0)] \\ & \times \left(\sum_{x^*} m^*(x^*, w_i; \theta)[\psi(x^*, \widehat{\gamma}) - \psi(x^*, \gamma_0)]\right). \end{aligned}$$

Given the uniform convergence of $\widehat{\gamma}$, for some $\varepsilon \to 0$ as $n \to \infty$

$$\begin{aligned}\sup_{\theta\in\Theta}|Q_n(\theta,\widehat{\gamma}) - Q_n(\theta,\gamma_0)| \leqslant & \left(\sup_{\|\widehat{\gamma}-\gamma_0\|_\infty\leqslant\varepsilon} \|\boldsymbol{\psi}(\widehat{\gamma}^v) - \boldsymbol{\psi}(\gamma_0^v)\|_1\right)^2 \sum_{x^*}\left(\frac{1}{n}\sum_{i=1}^n \tau(z_i,w_i)|m^*(x^*,w_i;\theta)|\right) \\ & + \left(\sup_{\|\widehat{\gamma}-\gamma_0\|_\infty\leqslant\varepsilon} \|\boldsymbol{\psi}(\widehat{\gamma}^v) - \boldsymbol{\psi}(\gamma_0^v)\|_1\right) \\ & \times \sum_{x^*}\left(\frac{1}{n}\sum_{i=1}^n \tau(z_i,w_i)|m^*(x^*,w_i;\theta)2[y_i - m(z_i,w_i;\theta,\gamma_0)]|\right).\end{aligned}$$

By Assumption 4.3 and Lemma 2.4 in Newey and McFadden (1994),

$$\begin{aligned}\sup_{\theta\in\Theta}|Q_n(\theta,\widehat{\gamma}) - Q_n(\theta,\gamma_0)| & \leqslant \mathrm{O_p}\left(\sup_{\|\widehat{\gamma}-\gamma_0\|_\infty\leqslant\varepsilon} \|\boldsymbol{\psi}(\widehat{\gamma}^v) - \boldsymbol{\psi}(\gamma_0^v)\|_1\right) \\ & = \mathrm{O}(\|\widehat{\gamma}-\gamma_0\|_\infty) \\ & = \mathrm{o_p}(1).\end{aligned}$$

The last two steps are due to Lemmas 2 and 3, and Assumption 4.4. Therefore,

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

Combining Eqs. (39) and (40) results in

$$\sup_{\theta\in\Theta}|Q_n(\theta,\widehat{\gamma}) - Q_0(\theta,\gamma_0)| = \mathrm{o_p}(1). \tag{41}$$

By Theorem 4.1.1 in Amemiya (1985b, p. 106), Assumptions 4.1–4.2 and Eq. (41) imply

$$\hat{\theta} \xrightarrow{\mathrm{p}} \theta_0. \qquad \square$$

Proof of Lemma 4. The major step in this proof is to show

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n g(\omega_i,\theta_0,\widehat{\gamma}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n [g(\omega_i,\theta_0,\gamma_0) + \delta(\omega_i)] + \mathrm{o_p}(1). \tag{42}$$

That means $\frac{1}{\sqrt{n}}\sum_{i=1}^n g(\omega_i,\theta_0,\widehat{\gamma})$ has the same asymptotic distribution as $\frac{1}{\sqrt{n}}\sum_{i=1}^n [g(\omega_i,\theta,\gamma_0) + \delta(\omega_i)]$, which converges to a normal distribution. Consider

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^n g(\omega_i,\theta_0,\widehat{\gamma}) - \frac{1}{\sqrt{n}}\sum_{i=1}^n g(\omega_i,\theta_0,\gamma_0) = & \frac{1}{\sqrt{n}}\sum_{i=1}^n \tau(z_i,w_i)\bigg([y_i - m(z_i,w_i;\theta_0,\widehat{\gamma})]\frac{\mathrm{d}}{\mathrm{d}\theta}m(z_i,w_i;\theta_0,\widehat{\gamma}) \\ & - [y_i - m(z_i,w_i;\theta_0,\gamma_0)]\frac{\mathrm{d}}{\mathrm{d}\theta}m(z_i,w_i;\theta_0,\gamma_0)\bigg).\end{aligned} \tag{43}$$

By the identity

$$\widehat{a}\widehat{b} - ab = (\widehat{a} - a)b + a(\widehat{b} - b) + (\widehat{a} - a)(\widehat{b} - b),$$

the right-hand side of Eq. (43) equals

$$\begin{aligned}= & -\frac{1}{\sqrt{n}}\sum_{i=1}^n \tau(z_i,w_i)[m(z_i,w_i;\theta_0,\widehat{\gamma}) - m(z_i,w_i;\theta_0,\gamma_0)]\frac{\mathrm{d}}{\mathrm{d}\theta}m(z_i,w_i;\theta_0,\gamma_0) \\ & + \frac{1}{\sqrt{n}}\sum_{i=1}^n \tau(z_i,w_i)[y_i - m(z_i,w_i;\theta_0,\gamma_0)]\left(\frac{\mathrm{d}}{\mathrm{d}\theta}m(z_i,w_i;\theta_0,\widehat{\gamma}) - \frac{\mathrm{d}}{\mathrm{d}\theta}m(z_i,w_i;\theta_0,\gamma_0)\right)\end{aligned}$$

$$
\begin{aligned}
&-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tau(z_i,w_i)[m(z_i,w_i;\theta_0,\widehat{\gamma})-m(z_i,w_i;\theta_0,\gamma_0)]\\
&\times\left(\frac{\mathrm{d}}{\mathrm{d}\theta}m(z_i,w_i;\theta_0,\widehat{\gamma})-\frac{\mathrm{d}}{\mathrm{d}\theta}m(z_i,w_i;\theta_0,\gamma_0)\right)\\
=&-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tau(z_i,w_i)\left(\sum_{x^*}m^*(x^*,w_i;\theta_0)[\psi(x^*,\widehat{\gamma})-\psi(x^*,\gamma_0)]\right)\frac{\mathrm{d}}{\mathrm{d}\theta}m(z_i,w_i;\theta_0,\gamma_0)\\
&+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tau(z_i,w_i)[y_i-m(z_i,w_i;\theta_0,\gamma_0)]\left(\sum_{x^*}\frac{\mathrm{d}}{\mathrm{d}\theta}m^*(x^*,w_i;\theta_0)[\psi(x^*,\widehat{\gamma})-\psi(x^*,\gamma_0)]\right)\\
&-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tau(z_i,w_i)\left(\sum_{x^*}m^*(x^*,w_i;\theta_0)[\psi(x^*,\widehat{\gamma})-\psi(x^*,\gamma_0)]\right)\\
&\times\left(\sum_{x^*}\frac{\mathrm{d}}{\mathrm{d}\theta}m^*(x^*,w_i;\theta_0)[\psi(x^*,\widehat{\gamma})-\psi(x^*,\gamma_0)]\right)\\
=&\,A_1+A_2+A_3. \qquad (44)
\end{aligned}
$$

I first show the term A_3 is $\mathrm{o_p}(1)$. The term A_3 is bounded as follows:

$$
\begin{aligned}
|A_3|\leqslant&\left(\sup_{\|\widehat{\gamma}-\gamma_0\|_\infty\leqslant\varepsilon}\|\boldsymbol{\psi}(\widehat{\gamma}^v)-\boldsymbol{\psi}(\gamma_0^v)\|_1\right)^2\\
&\times\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tau(z_i,w_i)\left(\sum_{x^*}|m^*(x^*,w_i;\theta_0)|\right)\left(\sum_{x^*}\left|\frac{\mathrm{d}}{\mathrm{d}\theta}m^*(x^*,w_i;\theta_0)\right|\right)\\
=&\,\mathrm{O_p}\left(\sup_{\|\widehat{\gamma}-\gamma_0\|_\infty\leqslant\varepsilon}\|\boldsymbol{\psi}(\widehat{\gamma}^v)-\boldsymbol{\psi}(\gamma_0^v)\|_1\right)^2\mathrm{O_p}(n^{1/2}).
\end{aligned}
$$

The last step is due to Assumption 5.3. By Lemma 3, $|A_3|$ is equal to $\mathrm{O_p}(n^{1/2}\|\widehat{\gamma}-\gamma_0\|_\infty^2)$. Assumption 5.4 then implies that $|A_3|$ is $\mathrm{o_p}(1)$.

Combining the terms A_1 and A_2 in Eq. (44) results in

$$
\begin{aligned}
A_1+A_2=&\,\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tau(z_i,w_i)\left(\sum_{x^*}\left([y_i-m(z_i,w_i;\theta_0,\gamma_0)]\frac{\mathrm{d}}{\mathrm{d}\theta}m^*(x^*,w_i;\theta_0)\right.\right.\\
&\left.\left.-m^*(x^*,w_i;\theta_0)\frac{\mathrm{d}}{\mathrm{d}\theta}m(z_i,w_i;\theta_0,\gamma_0)\right)[\psi(x^*,\widehat{\gamma})-\psi(x^*,\gamma_0)]\right)\\
=&\,\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tau(z_i,w_i)\left(\sum_{x^*}\frac{\mathrm{d}}{\mathrm{d}\theta}[(y_i-m(z_i,w_i;\theta_0,\gamma_0))m^*(x^*,w_i;\theta_0)][\psi(x^*,\widehat{\gamma})-\psi(x^*,\gamma_0)]\right).
\end{aligned}
$$

Define

$$
\mathbf{s}(y_i,z_i,w_i;\theta,\gamma_0)=\begin{pmatrix}(y_i-m(z_i,w_i;\theta,\gamma_0))m^*(x_{j_1}^*,w_i;\theta)\\ \vdots\\ (y_i-m(z_i,w_i;\theta,\gamma_0))m^*(x_{j_k}^*,w_i;\theta)\end{pmatrix}^{\mathrm{T}},
$$

where the ordering of $x_{j_1}^*,\ldots,x_{j_k}^*$ is the same as that of x^* in the vector $\boldsymbol{\psi}(\gamma^v)$. The result is

$$
A_1+A_2=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tau(z_i,w_i)\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbf{s}(y_i,z_i,w_i;\theta_0,\gamma_0)[\boldsymbol{\psi}(\widehat{\gamma}^v)-\boldsymbol{\psi}(\gamma_0^v)].
$$

The right-hand side equals

$$=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tau(z_i,w_i)\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbf{s}(y_i,z_i,w_i;\theta_0,\gamma_0)\frac{\partial\boldsymbol{\psi}(\gamma_0^v)}{\partial(\gamma^v)^{\mathrm{T}}}(\widehat{\gamma}^v-\gamma_0^v)+R,$$

where the remainder term R is

$$R=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tau(z_i,w_i)\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbf{s}(y_i,z_i,w_i;\theta_0,\gamma_0)\left[\boldsymbol{\psi}(\widehat{\gamma}^v)-\boldsymbol{\psi}(\gamma_0^v)-\frac{\partial\boldsymbol{\psi}(\gamma_0^v)}{\partial(\gamma^v)^{\mathrm{T}}}(\widehat{\gamma}^v-\gamma_0^v)\right].$$

It is bounded as follows:

$$|R|\leqslant\sup_{\|\widehat{\gamma}-\gamma_0\|_\infty\leqslant\varepsilon}\left\|\boldsymbol{\psi}(\widehat{\gamma}^v)-\boldsymbol{\psi}(\gamma_0^v)-\frac{\partial\boldsymbol{\psi}(\gamma_0^v)}{\partial(\gamma^v)^{\mathrm{T}}}(\widehat{\gamma}^v-\gamma_0^v)\right\|_1\times\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\|\tau(z_i,w_i)\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbf{s}(y_i,z_i,w_i;\theta_0,\gamma_0)\right\|_1.$$

The second term on the right-hand side is $\mathrm{O_p}(n^{1/2})$ by Assumption 5.3. By Lemma 3, the first term on the right-hand side is $\mathrm{O_p}(\|\widehat{\gamma}-\gamma_0\|_\infty^2)$. Assumption 5.4 then implies that $|R|=\mathrm{o_p}(1)$. Thus,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g(\omega_i,\theta_0,\widehat{\gamma})=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g(\omega_i,\theta_0,\gamma_0)+\frac{1}{n}\sum_{i=1}^{n}\tau(z_i,w_i)\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbf{s}(y_i,z_i,w_i;\theta_0,\gamma_0)\frac{\partial\boldsymbol{\psi}(\gamma_0^v)}{\partial(\gamma^v)^{\mathrm{T}}}(\widehat{\gamma}^v-\gamma_0^v)+\mathrm{o_p}(1).$$

Define

$$G(\omega,\gamma^v-\gamma_0^v)=\tau(z,w)\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbf{s}(y,z,w;\theta_0,\gamma_0)\frac{\partial\boldsymbol{\psi}(\gamma_0^v)}{\partial(\gamma^v)^{\mathrm{T}}}(\gamma^v-\gamma_0^v).$$

I have actually shown that for $\|\gamma-\gamma_0\|_\infty\leqslant\varepsilon$

$$\sup_{\|\gamma-\gamma_0\|_\infty\leqslant\varepsilon}|g(\omega,\theta_0,\gamma)-g(\omega,\theta_0,\gamma_0)-G(\omega,\gamma^v-\gamma_0^v)|\leqslant b(\omega)\|\gamma-\gamma_0\|_\infty^2,$$

with $\mathrm{E}[b(\omega)]<\infty$. That means condition (i) in Theorem 8.11 in Newey and McFadden (1994) is satisfied. Assumption 5.3 and Lemma 3 guarantee their condition (ii). The function $G(\omega,\gamma)$ is a linear function of γ so that

$$\int G(\omega,\gamma^v)\,\mathrm{d}F_0(\omega)=\int v(\omega)\gamma^v(\omega)\,\mathrm{d}\omega,$$

where

$$v(\widetilde{\omega})=\mathrm{E}\left[\tau(z_i,w_i)\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbf{s}(y_i,z_i,w_i;\theta_0,\gamma)\frac{\partial\boldsymbol{\psi}(\gamma^v)}{\partial(\gamma^v)^{\mathrm{T}}}\mathbf{1}\Big|_{\gamma^v=\gamma_0^v(\widetilde{\omega})}\Bigg|\widetilde{\omega}\right],$$

and the vector $\mathbf{1}$ has the same dimension as γ^v. Therefore, their condition (iii) in Theorem 8.11 is satisfied. Since the function $\boldsymbol{\psi}(\cdot)$ is analytic and ω has a compact support, Assumptions 5.1 and 5.3 implies Assumption (iv) in Theorem 8.11 is satisfied. Let

$$\delta(\omega)=v(\omega)-\mathrm{E}[v(\omega)].$$

Finally, Theorem 8.11 in Newey and McFadden (1994) implies that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g(\omega_i,\theta_0,\widehat{\gamma})=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[g(\omega_i,\theta,\gamma_0)+\delta(\omega_i)]+\mathrm{o_p}(1).$$

By the standard central limit theorem, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[g(\omega_i,\theta,\gamma_0)+\delta(\omega_i)]$ converges to a normal distribution with mean zero and variance $\Omega=Var\{g(\omega,\theta_0,\gamma_0)+\delta(\omega)\}$, i.e.,

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

Proof of Theorem 4 (*Asymptotic normality*). The estimator $\widehat{\theta}$ solves

$$\frac{1}{n}\sum_{i=1}^{n} g(\omega_i, \widehat{\theta}, \widehat{\gamma}) = 0.$$

The left-hand side equals

$$= \frac{1}{n}\sum_{i=1}^{n} g(\omega_i, \theta_0, \widehat{\gamma}) + \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}} g(\omega_i, \widetilde{\theta}, \widehat{\gamma})(\widehat{\theta} - \theta_0),$$

where $\widetilde{\theta}$ is between $\widehat{\theta}$ and θ_0. I show the asymptotic normality of the estimator using the delta method as follows:

$$-\frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}} g(\omega_i, \widetilde{\theta}, \widehat{\gamma})\sqrt{n}(\widehat{\theta} - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(\omega_i, \theta_0, \widehat{\gamma}). \tag{45}$$

Lemma 4 implies that the term on the right-hand side converges to a normal distribution as follows:

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

The next step is to show the uniform convergence of the Hessian matrix term on the left-hand side of Eq. (45), i.e.,

$$\sup_{\theta \in \Theta} \sup_{\|\gamma - \gamma_0\|_\infty \leqslant \varepsilon} \left| \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}} g(\omega_i, \widetilde{\theta}, \widehat{\gamma}) - \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}} g(\omega_i, \theta_0, \gamma_0) \right| = \mathrm{o_p}(1).$$

The left-hand side is bounded by

$$\begin{aligned} &\leqslant \sup_{\theta \in \Theta} \sup_{\|\gamma - \gamma_0\|_\infty \leqslant \varepsilon} \left| \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}} g(\omega_i, \widetilde{\theta}, \widehat{\gamma}) - \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}} g(\omega_i, \widetilde{\theta}, \gamma_0) \right| \\ &\quad + \sup_{\theta \in \Theta} \left| \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}} g(\omega_i, \widetilde{\theta}, \gamma_0) - \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}} g(\omega_i, \theta_0, \gamma_0) \right|. \end{aligned} \tag{46}$$

Because $\widetilde{\theta} \xrightarrow{\mathrm{p}} \theta_0$ and the function $g(\omega, \theta, \gamma_0)$ is continuously differentiable in θ by Assumption 4.2, the second term in Eq. (46) is $\mathrm{o_p}(1)$. Now consider the first term in Eq. (46):

$$\begin{aligned} D &\equiv \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}} g(\omega_i, \widetilde{\theta}, \widehat{\gamma}) - \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}} g(\omega_i, \widetilde{\theta}, \gamma_0) \\ &= \frac{1}{n}\sum_{i=1}^{n} \tau(z_i, w_i)\left([y_i - m(z_i, w_i; \widetilde{\theta}, \widehat{\gamma})] \frac{\mathrm{d}^2}{\mathrm{d}\theta\,\mathrm{d}\theta^{\mathrm{T}}} m(z_i, w_i; \widetilde{\theta}, \widehat{\gamma}) - [y_i - m(z_i, w_i; \widetilde{\theta}, \gamma_0)] \frac{\mathrm{d}^2}{\mathrm{d}\theta\,\mathrm{d}\theta^{\mathrm{T}}} m(z_i, w_i; \widetilde{\theta}, \gamma_0) \right) \\ &\quad - \frac{1}{n}\sum_{i=1}^{n} \tau(z_i, w_i)\left(\frac{\mathrm{d}}{\mathrm{d}\theta} m(z_i, w_i; \widetilde{\theta}, \widehat{\gamma}) \frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}} m(z_i, w_i; \widetilde{\theta}, \widehat{\gamma}) - \frac{\mathrm{d}}{\mathrm{d}\theta} m(z_i, w_i; \widetilde{\theta}, \gamma_0) \frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}} m(z_i, w_i; \widetilde{\theta}, \gamma_0) \right). \end{aligned}$$

Using identity

$$\widehat{a}\widehat{b} - ab = (\widehat{a} - a)b + a(\widehat{b} - b) + (\widehat{a} - a)(\widehat{b} - b),$$

I obtain for any $\widetilde{\theta} \in \Theta$,

$$\begin{aligned} D = &-\frac{1}{n}\sum_{i=1}^{n} \tau(z_i, w_i)\left([m(z_i, w_i; \widetilde{\theta}, \widehat{\gamma}) - m(z_i, w_i; \widetilde{\theta}, \gamma_0)] \frac{\mathrm{d}^2}{\mathrm{d}\theta\,\mathrm{d}\theta^{\mathrm{T}}} m(z_i, w_i; \widetilde{\theta}, \gamma_0) \right) \\ &+ \frac{1}{n}\sum_{i=1}^{n} \tau(z_i, w_i)[y_i - m(z_i, w_i; \widetilde{\theta}, \gamma_0)] \end{aligned}$$

$$\times\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta\,\mathrm{d}\theta^{\mathrm{T}}}m(z_i,w_i;\widetilde{\theta},\widehat{\gamma})-\frac{\mathrm{d}^2}{\mathrm{d}\theta\,\mathrm{d}\theta^{\mathrm{T}}}m(z_i,w_i;\widetilde{\theta},\gamma_0)\right)$$
$$-\frac{1}{n}\sum_{i=1}^{n}\tau(z_i,w_i)[m(z_i,w_i;\widetilde{\theta},\widehat{\gamma})-m(z_i,w_i;\widetilde{\theta},\gamma_0)]$$
$$\times\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta\,\mathrm{d}\theta^{\mathrm{T}}}m(z_i,w_i;\widetilde{\theta},\widehat{\gamma})-\frac{\mathrm{d}^2}{\mathrm{d}\theta\,\mathrm{d}\theta^{\mathrm{T}}}m(z_i,w_i;\widetilde{\theta},\gamma_0)\right)$$

$$-\frac{1}{n}\sum_{i=1}^{n}\tau(z_i,w_i)\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta}m(z_i,w_i;\widetilde{\theta},\widehat{\gamma})-\frac{\mathrm{d}}{\mathrm{d}\theta}m(z_i,w_i;\widetilde{\theta},\gamma_0)\right)\frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}}m(z_i,w_i;\widetilde{\theta},\gamma_0)\right]$$
$$-\frac{1}{n}\sum_{i=1}^{n}\tau(z_i,w_i)\left[\frac{\mathrm{d}}{\mathrm{d}\theta}m(z_i,w_i;\widetilde{\theta},\gamma_0)\left(\frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}}m(z_i,w_i;\widetilde{\theta},\widehat{\gamma})-\frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}}m(z_i,w_i;\widetilde{\theta},\gamma_0)\right)\right]$$
$$-\frac{1}{n}\sum_{i=1}^{n}\tau(z_i,w_i)\left(\frac{\mathrm{d}}{\mathrm{d}\theta}m(z_i,w_i;\widetilde{\theta},\widehat{\gamma})-\frac{\mathrm{d}}{\mathrm{d}\theta}m(z_i,w_i;\widetilde{\theta},\gamma_0)\right)$$
$$\times\left(\frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}}m(z_i,w_i;\widetilde{\theta},\widehat{\gamma})-\frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}}m(z_i,w_i;\widetilde{\theta},\gamma_0)\right)$$

$$=-\frac{1}{n}\sum_{i=1}^{n}\tau(z_i,w_i)\left(\left(\sum_{x^*}m^*(x^*,w_i;\widetilde{\theta})[\psi(x^*,\widehat{\gamma})-\psi(x^*,\gamma_0)]\right)\frac{\mathrm{d}^2}{\mathrm{d}\theta\,\mathrm{d}\theta^{\mathrm{T}}}m(z_i,w_i;\widetilde{\theta},\gamma_0)\right)$$
$$+\frac{1}{n}\sum_{i=1}^{n}\tau(z_i,w_i)\left[[y_i-m(z_i,w_i;\widetilde{\theta},\gamma_0)]\left(\sum_{x^*}\frac{\mathrm{d}^2}{\mathrm{d}\theta\,\mathrm{d}\theta^{\mathrm{T}}}m^*(x^*,w_i;\widetilde{\theta})[\psi(x^*,\widehat{\gamma})-\psi(x^*,\gamma_0)]\right)\right]$$
$$-\frac{1}{n}\sum_{i=1}^{n}\tau(z_i,w_i)\left(\sum_{x^*}\frac{\mathrm{d}}{\mathrm{d}\theta}m^*(x^*,w_i;\widetilde{\theta})[\psi(x^*,\widehat{\gamma})-\psi(x^*,\gamma_0)]\right)$$
$$\times\left(\sum_{x^*}\frac{\mathrm{d}^2}{\mathrm{d}\theta\,\mathrm{d}\theta^{\mathrm{T}}}m^*(x^*,w_i;\widetilde{\theta})[\psi(x^*,\widehat{\gamma})-\psi(x^*,\gamma_0)]\right)$$

$$-\frac{1}{n}\sum_{i=1}^{n}\tau(z_i,w_i)\left(\sum_{x^*}\frac{\mathrm{d}}{\mathrm{d}\theta}m^*(x^*,w_i;\widetilde{\theta})[\psi(x^*,\widehat{\gamma})-\psi(x^*,\gamma_0)]\right)\frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}}m(z_i,w_i;\widetilde{\theta},\gamma_0)$$
$$-\frac{1}{n}\sum_{i=1}^{n}\tau(z_i,w_i)\frac{\mathrm{d}}{\mathrm{d}\theta}m(z_i,w_i;\widetilde{\theta},\gamma_0)\left(\sum_{x^*}\frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}}m^*(x^*,w_i;\widetilde{\theta})[\psi(x^*,\widehat{\gamma})-\psi(x^*,\gamma_0)]\right)$$
$$-\frac{1}{n}\sum_{i=1}^{n}\tau(z_i,w_i)\left(\sum_{x^*}\frac{\mathrm{d}}{\mathrm{d}\theta}m^*(x^*,w_i;\widetilde{\theta})[\psi(x^*,\widehat{\gamma})-\psi(x^*,\gamma_0)]\right)$$
$$\times\left(\sum_{x^*}\frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}}m^*(x^*,w_i;\widetilde{\theta})[\psi(x^*,\widehat{\gamma})-\psi(x^*,\gamma_0)]\right).$$

Therefore, the term $|D|$ is bounded by

$$\sup_{\|\widehat{\gamma}-\gamma_0\|_\infty\leqslant\varepsilon}\|\boldsymbol{\psi}(\widehat{\gamma}^v)-\boldsymbol{\psi}(\gamma_0^v)\|_1\left|\frac{1}{n}\sum_{i=1}^{n}\tau(z_i,w_i)\left(\sum_{x^*}|m^*(x^*,w_i;\widetilde{\theta})|\left|\frac{\mathrm{d}^2}{\mathrm{d}\theta\,\mathrm{d}\theta^{\mathrm{T}}}m(z_i,w_i;\widetilde{\theta},\gamma_0)\right|\right)\right|$$
$$+\sup_{\|\widehat{\gamma}-\gamma_0\|_\infty\leqslant\varepsilon}\|\boldsymbol{\psi}(\widehat{\gamma}^v)-\boldsymbol{\psi}(\gamma_0^v)\|_1\left|\frac{1}{n}\sum_{i=1}^{n}\tau(z_i,w_i)|y_i-m(z_i,w_i;\widetilde{\theta},\gamma_0)|\sum_{x^*}\left|\frac{\mathrm{d}^2}{\mathrm{d}\theta\,\mathrm{d}\theta^{\mathrm{T}}}m^*(x^*,w_i;\widetilde{\theta})\right|\right|$$

$$
\begin{aligned}
&+\left(\sup_{\|\widehat{\gamma}-\gamma_0\|_\infty\leqslant\varepsilon}\|\boldsymbol{\psi}(\widehat{\gamma}^v)-\boldsymbol{\psi}(\gamma_0^v)\|_1\right)^2\\
&\times\left|\frac{1}{n}\sum_{i=1}^{n}\tau(z_i,w_i)\left(\sum_{x^*}\left|\frac{\mathrm{d}}{\mathrm{d}\theta}m^*(x^*,w_i;\widetilde{\theta})\right|\right)\left(\sum_{x^*}\left|\frac{\mathrm{d}^2}{\mathrm{d}\theta\,\mathrm{d}\theta^{\mathrm{T}}}m^*(x^*,w_i;\widetilde{\theta})\right|\right)\right|\\
&+\sup_{\|\widehat{\gamma}-\gamma_0\|_\infty\leqslant\varepsilon}\|\boldsymbol{\psi}(\widehat{\gamma}^v)-\boldsymbol{\psi}(\gamma_0^v)\|_1\left|\frac{1}{n}\sum_{i=1}^{n}\tau(z_i,w_i)\sum_{x^*}\left|\frac{\mathrm{d}}{\mathrm{d}\theta}m^*(x^*,w_i;\widetilde{\theta})\right|\left|\frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}}m(z_i,w_i;\widetilde{\theta},\gamma_0)\right|\right|\\
&+\sup_{\|\widehat{\gamma}-\gamma_0\|_\infty\leqslant\varepsilon}\|\boldsymbol{\psi}(\widehat{\gamma}^v)-\boldsymbol{\psi}(\gamma_0^v)\|_1\left|\frac{1}{n}\sum_{i=1}^{n}\tau(z_i,w_i)\left|\frac{\mathrm{d}}{\mathrm{d}\theta}m(z_i,w_i;\widetilde{\theta},\gamma_0)\right|\left(\sum_{x^*}\left|\frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}}m^*(x^*,w_i;\widetilde{\theta})\right|\right)\right|\\
&+\left(\sup_{\|\widehat{\gamma}-\gamma_0\|_\infty\leqslant\varepsilon}\|\boldsymbol{\psi}(\widehat{\gamma}^v)-\boldsymbol{\psi}(\gamma_0^v)\|_1\right)^2\\
&\times\left|\frac{1}{n}\sum_{i=1}^{n}\tau(z_i,w_i)\left(\sum_{x^*}\left|\frac{\mathrm{d}}{\mathrm{d}\theta}m^*(x^*,w_i;\widetilde{\theta})\right|\right)\left(\sum_{x^*}\left|\frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}}m^*(x^*,w_i;\widetilde{\theta})\right|\right)\right|.
\end{aligned}
$$

By Assumption 5.6 and Lemma 2.4 in Newey and McFadden (1994),

$$
\sup_{\theta\in\Theta}\sup_{\|\gamma-\gamma_0\|_\infty\leqslant\varepsilon}|D|=\mathrm{O_p}\left(\sup_{\|\widehat{\gamma}-\gamma_0\|_\infty\leqslant\varepsilon}\|\boldsymbol{\psi}(\widehat{\gamma}^v)-\boldsymbol{\psi}(\gamma_0^v)\|_1\right).
$$

By Lemma 3 and Assumptions 3.4 and 4.4, $\sup_{\theta\in\Theta}\sup_{\|\widehat{\gamma}-\gamma_0\|_\infty\leqslant\varepsilon}|D|=\mathrm{o_p}(1)$. Finally, the result is

$$
\sup_{\theta\in\Theta}\sup_{\|\gamma-\gamma_0\|_\infty\leqslant\varepsilon}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}}g(\omega_i,\widetilde{\theta},\widehat{\gamma})-\frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{d}}{\mathrm{d}\theta^{\mathrm{T}}}g(\omega_i,\theta_0,\gamma_0)\right|=\mathrm{o_p}(1).
$$

Because $\mathrm{E}[\nabla_\theta g(\omega,\theta_0,\gamma_0)]$ exists and is nonsingular, the Slutsky theorem then implies

$$
\sqrt{n}(\hat{\theta}-\theta_0)\overset{\mathrm{d}}{\rightarrow}\mathrm{N}(0,G_\theta^{-1}\Omega G_\theta^{-1\prime}),
$$

where

$$
\begin{aligned}
G_\theta&=\mathrm{E}[\nabla_\theta g(\omega,\theta_0,\gamma_0)],\\
\Omega&=Var[g(\omega,\theta_0,\gamma_0)+\delta(\omega)].\qquad\square
\end{aligned}
$$

References

Aigner, D., 1973. Regression with a binary independent variable subject to errors of observations. Journal of Econometrics 1, 49–60.

Amemiya, Y., 1985a. Instrumental variable estimator for the nonlinear errors-in-variables model. Journal of Econometrics 28, 273–289.

Amemiya, Y., 1985b. Advanced Econometrics. Harvard University Press, Cambridge, MA.

Amemiya, Y., Fuller, W.A., 1988. Estimation for the nonlinear functional relationship. Annals of Statistics 16, 147–160.

Andrew, A., Chu, K., Lancaster, P., 1993. Derivatives of eigenvalues and eigenvectors of matrix functions. SIAM Journal on Matrix Analysis and Applications 14 (4), 903–926.

Black, D., Berger, M.C., Scott, F.A., 2000. Bounding parameter estimates with nonclassical measurement error. Journal of the American Statistical Association 95, 739–748.

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

Bound, J., Brown, C., Mathiowetz, N., 2001. Measurement error in survey data. In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 5.

Buzas, J.S., 1997. Instrumental variable estimation in nonlinear measurement error models. Communications in Statistics, Part A—Theory and Methods 26, 2861–2877.

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

Darolles, S., Florens, J.-P., Renault, E., 2000. Nonparametric instrumental regression. Manuscript, GREMAQ, University of Toulouse.

Fan, Y., Li, Q., 1996. Consistent model specification tests: omitted variables and semiparametric functional forms. Econometrica 64, 865–890.

Freeman, R., 1984. Longitudinal analyses of the effects of trade unions. Journal of Labor Economics 2, 1–26.

Hardle, W., Marron, J., 1990. Semiparametric comparison of regression curves. Annals of Statistics 18, 63–89.

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

Hausman, J., Newey, W.K., Powell, J.L., 1995. Nonlinear errors in variables: estimation of some Engle curves. Journal of Econometrics 65, 205–233.

Hui, S.L., Walter, S.D., 1980. Estimating the error rates of diagnostic tests. Biometrics 36, 167–171.

Kane, T.J., Rouse, C.E., Staiger, D., 1999. Estimating returns to schooling when schooling is misreported. NBER Working Paper #7235.

Levine, P., 1993. CPS contemporaneous and retrospective unemployment compared. Monthly Labor Review 116, 33–39.

Lewbel, A., 1998. Semiparametric latent variable model estimation with endogenous or mismeasured regressors. Econometrica 66, 105–121.

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

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

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

Molinari, F., 2003. Contaminated, corrupted, and missing data. Ph.D. Dissertation, Northwestern University.

Molinari, F., 2005. Partial identification of probability distributions with misclassified data. Mimeo, Cornell University.

Newey, W., 1992. Partial means, kernel estimation, and a general asymptotic variance estimator. Mimeo, MIT.

Newey, W., 1994a. The asymptotic variance of semiparametric estimators. Econometrica 62, 1349–1382.

Newey, W., 1994b. Kernel estimation of partial means and a general variance estimator. Econometric Theory 10, 233–253.

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

Newey, W., McFadden, D., 1994. Large sample estimation and hypothesis testing. In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics, vol. 4, pp. 2111–2245.

Newey, W., Powell, J., 2003. Instrumental variable estimation of nonparametric models. Econometrica 71 (5), 1565–1578.

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

Schennach, S., 2004. Estimation of nonlinear models with measurement error. Econometrica 72 (1), 33–76.

Schennach, S., 2007. Instrumental variable estimation of nonlinear errors-in-variables models. Econometrica 75, 201–239.

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

Wooldridge, J., 2002. Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge, MA.