Estimation of Semiparametric Models with Mismeasured Endogenous Regressors Using Control Variables

Kyoo il Kim[*]
Michigan State University

Suyong Song[†]
University of Iowa

January 2017

Abstract

This paper studies a class of semiparametric models with mismeasured endogenous regressors. In particular we allow for infinite-dimensional parameters to depend on endogenous regressors that are unobservable because of nonclassical measurement errors. For these models we utilize the existence of control variables that ensure conditional mean independence of endogenous regressors and unobserved causes being conditioned on the variables. We provide a set of sufficient conditions for identification of structural parameters, which control for both endogeneity and measurement error using control variables. Based on the identification results we propose a sieve estimation of the parameters. We derive the asymptotic properties of the proposed estimators; consistency, convergence rate, and $\sqrt{n}$-asymptotic normality of the estimator of the finite dimensional parameters. Monte Carlo simulations show that our proposed estimator performs well in finite samples in correcting for endogeneity and measurement error.

Keywords: Endogeneity, Control Variables, Nonclassical Measurement Error, Conditional Independence
JEL Classification: C14, C21, C26

[*] Department of Economics, Michigan State University, Marshall-Adams Hall, 486 W Circle Dr Rm 110, East Lansing, MI 48824, Tel. 517-353-9008, E-mail: kyookim@msu.edu.

[†] Department of Economics, University of Iowa, W360 Pappajohn Business Building, 21 E Market St, Iowa City, IA 52242, Tel. 319-335-0832, E-mail: suyong-song@uiowa.edu.

1 Introduction

We consider a class of semiparametric models with mismeasured endogenous regressors as below and its extensions

$$Y_1 = \pi_0(Z_1) + G(Y_2; \theta_0, h_0) + \varepsilon, \quad Y_2^* = g(Y_2, e) \tag{1}$$

where $Y \equiv (Y_1, Y_2')'$ is a vector of endogenous (or dependent) variables, $W \equiv (Z_1', V')'$ is a vector of conditioning variables where Z_1 is a vector of exogenous regressors and V is a vector of control variables such that Y_2 and ε are conditionally mean independent given W. The control variables V can be exogenous variables or be generated variables. Here $G(\cdot)$ is a known function up to an unknown parameter vector (θ, h) where θ is finite dimensional parameters and h is infinite-dimensional functions. The model includes examples of a parametric model $G(Y_2) = Y_2'\theta$, a partially linear model $G(Y_2) = Y_{2,1}'\theta + h(Y_{2,2})$ such that $Y_2 = (Y_{2,1}', Y_{2,2}')'$, and a nonparametric regression model $G(Y_2) = h(Y_2)$. Specifically, the model allows dependence of $h(\cdot)$ on the endogenous variables Y_2. For this model (θ, h) is the parameters of interest and we let (θ_0, h_0) denote the true parameters. Here Y_1, Y_2, Z_1 and V denote true variables while Y_2 is only measured with error as Y_2^*, which we refer to the mismeasured regressor(s) where e is the measurement error on Y_2. This measurement problem of Y_2 hinders us from using other existing semiparametric approaches and we develop a new approach to tackle the problem.

Specifically given the conditional mean independence of Y_2 and ε given W and identification of relevant density functions, we show that a conditional moment restriction that uniquely determines the true parameters (θ_0, h_0) is obtained from the model (1) as

$$E[(Y_2 - E[Y_2 \mid W])(Y_1 - G(Y_2; \theta_0, h_0)) \mid W] = 0. \tag{2}$$

Utilizing this moment condition we propose to estimate h_0 using a sieve approach. One merit of our sieve approach is that given a sieve approximation of h, the moment condition (2) yields a closed-form solution and the proposed estimator is obtained as a weighted least squares estimator. Therefore, our proposed estimator is easy to implement for this class of model.

Note that in our approach (θ_0, h_0) is separately estimated from $\pi_0(Z_1)$ and it does not require estimating $\pi_0(Z_1)$ in the moment condition. Given (θ_0, h_0), the identification of $\pi_0(Z_1)$ is trivial as $\pi_0(Z_1) = E[Y_1 - G(Y_2; \theta_0, h_0)|Z_1]$ if we further impose $E[\varepsilon|Z_1] = 0$ since Z_1 is exogenous. If π_0 is also a parameter of interest in the model (1), once (θ_0, h_0) is estimated from the moment condition (2), we then go back to (1) and estimate $\pi_0(Z_1)$ using a regression of $Y_1 - G(Y_2; \hat{\theta}_n, \hat{h}_n)$ on Z_1 where $(\hat{\theta}_n, \hat{h}_n)$ denotes a consistent estimator of (θ_0, h_0).

Our identification argument that generates the moment condition (2) is based on the conditional mean independence condition. Conditional mean independence or conditional independence has been utilized for a basis of identification in various econometric problems including estimation of treatment effects (see, among many others, Heckman, Ichimura, and Todd (1998), Dehejia and Wahba (1999), Lechner (2001), Imbens and Newey (2009), and Imbens and Wooldridge (2009)). As a specific example of the conditional independence, consider the problem of estimating the effect of family income on children's health as in Case, Lubotsky, and Paxson (2002), Currie and Stabile (2003), and Condliffe and Link (2008). In this example family income (Y_2) is endogenous because of the dependence between household earning potential and children's health determinants. However, given parental education as a control variable (V), which acts as a proxy to parental cognitive ability, family income can be treated as being independent of children's health determinants. Because endogeneity problem comes from the common cause (i.e., parental cognitive ability) between household earning potential and children's health determinants, a proxy to parental cognitive ability such as parental education can solve the endogeneity problem.

In estimating the parameters using the moment restriction (2), the difficulty arises because the endogenous regressor Y_2 is not observed but measured with error as Y_2^*. To develop our estimation strategy in this setting we cast the conditional moment restriction (2) into a more general form

$$\begin{aligned} m(W,\theta,h(\cdot)) &\equiv E[\rho(Y,W,\theta,h(\cdot)) \mid W], \\ m(W,\theta_0,h_0(\cdot)) &= 0. \end{aligned} \tag{3}$$

Here an important feature of the model (3) is that because Y_2 is not observed, the conditional moment function $m(W,\theta,h(\cdot))$ is not directly observable from data (given θ and h). Therefore, in the specific example above, our setting allows family income to be only measured with error. Under a set of exclusion restrictions stating that (i) given the true regressors and the control variables, mismeasured regressors do not provide further information on dependent variables, (ii) given the true regressors, control variables do not provide further information on dependent variables, and (iii) given the true regressors, control variables do not provide further information on mismeasured regressors, we show that the conditional moment function is identified from data on (Y_1, Y_2^*, W) by means of recovering relevant conditional density functions from the observables. Our approach builds on an operator-based approach for nonclassical measurement errors (e.g. Hu and Schennach 2008). Given the identified conditional moment function, we propose to estimate the model parameters (θ_0, h_0) using a sieve Minimum Distance (MD) method.

Another class of model that fits into our framework is a triangular nonparametric simultaneous equations model (e.g. Newey, Powell, and Vella 1999) with mismeasured endogenous regressor

$$\begin{aligned} Y_1 &= \pi_0(Z_1) + h_0(Y_2) + \varepsilon, \\ Y_2 &= r(Z,V), \\ Y_2^* &= g(Y_2,e), \end{aligned} \tag{4}$$

where e is the measurement error on Y_2 and $Z \equiv (Z_1, Z_2)$ with Z_2 being a vector of excluded instruments. If Y_2 is observable, the control variable V is obtained as the conditional cumulative distribution function (CDF) of Y_2 given Z, $F_{Y_2|Z}(Y_2|Z)$, under the assumption that V is a scalar, $r(\cdot,\cdot)$ is strictly monotonic in V, and Z is independent of V (see e.g. Matzkin 2003 and Imbens and

Newey 2009). Then the conditional mean independence of Y_2 and ε given $W \equiv (Z_1', V')'$ implies

$$E\left[(Y_2 - E[Y_2 \mid W])\,(Y_1 - h_0(Y_2)) \mid W\right] = 0 \tag{5}$$

which also has the form of (3).

In our setting like (4) the key departure from the usual triangular equations model is that the endogenous regressor Y_2 is measured with error. In this setting, besides the regressor Y_2 itself being mismeasured, the measurement error further complicates the problem because, even if the distribution function $F_{Y_2|Z}(\cdot|\cdot)$ is known, the control variable obtained by plugging in the error-laden observation Y_2^* is also contaminated by the measurement error since we would have $V^* \equiv F_{Y_2|Z}(Y_2^*|Z)$. This complication makes other existing control function methods – that use the control variables as additional regressors – not applicable in our setting. Nevertheless, the identification of the CDF $F_{Y_2|Z}(\cdot|\cdot)$ in a pre-stage will suffice for implementing our proposed estimator that utilizes the moment condition (5). In our approach we show that under a set of exclusion restrictions the conditional moment function (5) is identified by recovering relevant conditional density functions from observations on (Y_1, Y_2^*, Z_1, V) where V is a generated variable from other observables. Because in the construction of the control variable $V \equiv F_{Y_2|Z}(Y_2|Z)$ here the dependent variable Y_2 is not observed, this problem can be understood as the measurement error on the left-side variable. We propose two alternative approaches to tackle this problem. One is to recover the CDF using repeated measurements of Y_2. The other is to recover the CDF using an instrumental variable for Y_2.

Note that our approach is different from other control function methods because V is not used as an additional regressor but only plays the role of a conditioning variable, so we do not estimate a nonparametric function of V in the regression equation. In our setting whether or not V is observable is irrelevant as long as we recover required conditional density/distribution functions. Finally, note that imposing the conditional mean independence of Y_2 and ε given W may serve as an alternative to other approaches that are assuming either the sufficiency of control variables $E[\varepsilon \mid V, Z] = E[\varepsilon \mid V]$ as in Newey, Powell, and Vella (1999) or assuming the conditional moment condition $E[\varepsilon \mid Z] = 0$ as in Newey and Powell (2003) and Ai and Chen (2003). Note that none of these modeling assumptions including ours implies the other. For the conditional moment restriction model $E[\varepsilon \mid Z] = 0$ Ai and Chen (2003) develop a semiparametric sieve estimator and establish its asymptotic properties. Song (2015) considers measurement errors in their setting. Because our model is based on the conditional mean independence given control variables, which imposes different forms of moment restrictions, their methods are not applicable in our setting.

Given our identification results we propose a sieve estimation method to estimate the parameters. Our estimation proceeds in two stages. In the first stage we estimate the unknown densities to recover the conditional moment function using a sieve Maximum Likelihood Estimation (MLE), and in the second stage we estimate the structural parameters using a sieve MD estimation. We then derive the asymptotic properties of the estimator, such as consistency, convergence rate, and $\sqrt{n}$-asymptotic normality of the estimator of the finite dimensional parameters. We focus on the case for which the control variables V are observables and then show how the setting can extend to the case for which V are generated variables as in the triangular simultaneous equations model (4). In the latter case, we obtain the asymptotic variance of the estimator accounting for the generated control variables by utilizing the approach from Hahn and Ridder (2013), who study the asymptotic variances of semiparametric estimators with generated regressors.

We run Monte Carlo simulations to illustrate the finite-sample performances of our proposed estimator. We experiment with a partially linear model and with an additively-separable nonparametric regression model. In particular we consider different structures of measurement errors and vary their influences in our experiments. Our proposed estimator shows desirable finite-sample behaviors in correcting for endogeneity as well as measurement errors. On the other hand a conventional sieve instrumental variable estimator which only corrects for endogeneity shows considerably large biases.

The outline of the paper is as follows. Section 2 discusses issues of identification in the presence of mismeasured endogenous regressors. Section 3 develops a sieve estimation of the parameters. Sections 4 and 5 study the asymptotic property of the proposed estimator. We report Monte Carlo simulations in Section 6 to illustrate finite sample performance of the estimator. We then conclude in Section 7. Technical details are gathered in the appendix.

2 Identification Using Control Variables

We develop notations and further articulate the nature of endogeneity and measurement error in the model we consider. We denote the supports of the distributions of the random variables Y_1, Y_2, Y_2^*, Z_1, and V by $\mathcal{Y}_1, \mathcal{Y}_2, \mathcal{Y}_2^*, \mathcal{Z}_1$, and $\mathcal{V}$, respectively. The joint density of Y_1 and (Y_2, Y_2^*, Z_1, V) admits a bounded density with respect to the product measure of some dominating measure μ on $\mathcal{Y}_1$ and the Lebesgue measure on $\mathcal{Y}_2 \times \mathcal{Y}_2^* \times \mathcal{Z}_1 \times \mathcal{V}$. All marginal and conditional densities are also bounded. For notational ease, let $Y \equiv (Y_1, Y_2')' \in \mathcal{Y} \equiv \mathcal{Y}_1 \times \mathcal{Y}_2$, $Y^* \equiv (Y_1, Y_2^{*\prime})' \in \mathcal{Y}^* \equiv \mathcal{Y}_1 \times \mathcal{Y}_2^*$, $W \equiv (Z_1', V')' \in \mathcal{W} \equiv \mathcal{Z}_1 \times \mathcal{V}$. Let $X \equiv (Y', W')' \in \mathcal{X} \equiv \mathcal{Y} \times \mathcal{W}$. Suppose that the true observations $\{(Y_i, W_i) : i = 1, 2, ..., n\}$ are independently drawn from the distribution of (Y, W) with support $\mathcal{Y} \times \mathcal{W}$, where $\mathcal{Y}$ is a subset of $\mathcal{R}^{d_y}$ and $\mathcal{W}$ is a compact subset of $\mathcal{R}^{d_w}$. Let $\alpha_0 \equiv (\theta_0, h_0) \in \mathcal{A} \equiv \Theta \times \mathcal{H}$. We assume that $\Theta \subseteq \mathcal{R}^{d_\theta}$ is compact with nonempty interior, and that $\mathcal{H}$ is a space of continuous functions. We use notations $f_{R_1}(r_1)$, $f_{R_1|R_2}(r_1 \mid r_2)$, and $F_{R_1|R_2}(r_1 \mid r_2)$ to denote the marginal density of variable R_1, the conditional density of R_1 conditional on R_2, and the cumulative distribution function of R_1 conditional on R_2, respectively.

2.1 Conditions for Identification

Our identification results utilize the notion of conditional (mean) independence or equivalent exclusion restrictions. Following Dawid (1979a), we will write $\mathcal{A} \perp \mathcal{B} \mid \mathcal{C}$ to denote that $\mathcal{A}$ is independent of $\mathcal{B}$ being conditioned on $\mathcal{C}$. We first state required conditions for identification of parameters from the moment condition (3) using control variables when Y_2 is measured with error as Y_2^*.

Assumption 2.1 (i) $E[Y_2\varepsilon \mid W] = E[Y_2 \mid W]E[\varepsilon \mid W]$; (ii) (θ_0, h_0) *is the only* $(\theta, h) \in \Theta \times \mathcal{H}$

satisfying the conditional moment restrictions (3).

Assumption 2.2 $Y_1 \perp Y_2^* \mid (Y_2, W)$ *for all* $(Y_1, Y_2, Y_2^*, W) \in \mathcal{Y}_1 \times \mathcal{Y}_2 \times \mathcal{Y}_2^* \times \mathcal{W}$.

Assumption 2.3 $Y_1 \perp V \mid (Y_2, Z_1)$ *for all* $(Y_1, Y_2, W) \in \mathcal{Y}_1 \times \mathcal{Y}_2 \times \mathcal{W}$.

Assumption 2.4 $Y_2^* \perp V \mid (Y_2, Z_1)$ *for all* $(Y_2, Y_2^*, W) \in \mathcal{Y}_2 \times \mathcal{Y}_2^* \times \mathcal{W}$.

Assumption 2.1 ensures identification of the parameters when all true Y are observed and endogeneity is properly controlled. In particular the conditional moment condition in equation (3) can arise from Assumption 2.1 (i), the conditional mean independence between the unobserved cause of Y_1 and the endogenous regressor Y_2 conditional on W. Assumption 2.1 (ii), the uniqueness of the true parameters, holds if the set $\{w \in \mathcal{W} : \ m(w, \theta, h) \neq m(w, \theta_0, h_0)\}$ has positive probability for any $(\theta, h) \neq (\theta_0, h_0) \in \Theta \times \mathcal{H}$. Assumptions 2.2 - 2.4 state additional conditional independence conditions which are equivalent to relevant exclusion restrictions among relationships of variables. Several variants of these conditions have been widely used in the econometrics literature (e.g., Altonji and Matzkin 2005, Heckman and Vytlacil 2005, Imbens and Newey 2009) and various tests of these conditions have been studied in other strand of the literature (e.g., Su and White 2007, Song 2009). In our setting these assumptions are utilized for recovering the (conditional) densities of true variables from the observed ones. Assumption 2.2 can be equivalently stated in terms of density functions as $f_{Y_1|Y_2Y_2^*W}(y_1 \mid y_2, y_2^*, w) = f_{Y_1|Y_2W}(y_1 \mid y_2, w)$, and Assumption 2.3 is equivalent to $f_{Y_1|Y_2Z_1V}(y_1 \mid y_2, z_1, v) = f_{Y_1|Y_2Z_1}(y_1 \mid y_2, z_1)$. Assumption 2.2 states that given the true regressors and the control variables, mismeasured regressors do not provide further information on dependent variables. Assumption 2.3 states that given the true regressors, control variables do not provide further information on dependent variables. Similarly, Assumption 2.4 can be equivalently written as $f_{Y_2^*|Y_2Z_1V}(y_2^* \mid y_2, z_1, v) = f_{Y_2^*|Y_2Z_1}(y_2^* \mid y_2, z_1)$, which means that given the true regressors, control variables do not have further information on mismeasured regressors. We further discuss these assumptions for a specific model and provide sufficient conditions for the assumptions in Section 2.3.

2.2 Identification of Parameters by Means of Density Functions

Here we show how to obtain identification of parameters using the conditional (mean) independence conditions by means of density functions. First note that by Assumption 2.1(i), we have

$$E[Y_2\varepsilon \mid W] = E[Y_2 \mid W]E[\varepsilon \mid W],$$

which is followed by

$$E[Y_2\varepsilon \mid W] = E[E[Y_2 \mid W]\varepsilon \mid W].$$

Then by rearranging both sides, we obtain

$$E[(Y_2 - E[Y_2 \mid W])\varepsilon \mid W] = E[(Y_2 - E[Y_2 \mid W])(Y_1 - G(Y_2;\theta,h)) \mid W] = 0.$$

Let $\rho(X,\theta,h) = (Y_2 - E[Y_2 \mid W])(Y_1 - G(Y_2;\theta,h))$ be the residual. Let $\alpha = (\theta,h)$ and let $m(W,\alpha) \equiv E[\rho(X,\theta,h) \mid W]$ denote the conditional mean function of the residual, $\rho(X,\theta,h)$, given W. Then we can cast the above conditional moment restriction into the more general form as (3). For a class of semiparametric models, Ai and Chen (2003) propose a sieve MD estimation which requires a consistent estimator for the conditional moment function. Because Y_2 is not observed in our model (3), their methods are not directly applicable to our case. We instead estimate α through recovering conditional density functions associated with the unobserved true variable Y_2 from the observed data. Our main idea is that the conditional mean function (3) can be written as an integral form

$$m(w,\alpha) \equiv E[\rho(X,\theta,h) \mid W] = \int_{\mathcal{Y}} \rho(x,\theta,h) f_{Y|W}(y \mid w)dy, \tag{6}$$

so identification of the conditional mean function $m(w,\alpha)$ is obtained, given the identification of the conditional density $f_{Y|W}(y \mid w)$ and the residual function $\rho(X,\theta,h)$. By utilizing the conditional independence condition of Assumption 2.3, we note that two densities $f_{Y_1|Y_2Z_1}(y_1 \mid y_2, z_1)$ and $f_{Y_2|W}(y_2 \mid w)$ are sufficient for the identification of the conditional density $f_{Y|W}(y \mid w)$. We further note that the conditional mean function $E[Y_2 \mid W]$ inside the residual can be written as

$$E[Y_2 \mid W = w] = \int_{\mathcal{Y}_2} y_2 f_{Y_2|W}(y_2 \mid w)dy_2, \tag{7}$$

so that $f_{Y_2|W}(y_2 \mid w)$ is also sufficient for the identification of the residual function $\rho(X,\theta,h(\cdot))$, given the parameter α. Rewriting the conditional mean functions in terms of conditional densities makes it clear that once $f_{Y_1|Y_2Z_1}(y_1 \mid y_2,z_1)$ and $f_{Y_2|W}(y_2 \mid w)$ are obtained, we then can estimate α_0 using the conditional moments because α_0 is the unique solution to the equation $m(w,\alpha) = 0$. In Section 3, given identification of $m(w,\alpha)$, we propose a sieve estimator of α_0 that is obtained by minimizing a sample analogue of the population criterion function

$$Q(\alpha) \equiv E\left[m(W_i,\alpha)'[\Sigma(W_i)]^{-1}m(W_i,\alpha)\right] \tag{8}$$

where $\Sigma(W)$ denotes a positive-definite weighting matrix.

Therefore our problem boils down to the problem of recovering the two densities $f_{Y_1|Y_2Z_1}(y_1 \mid y_2, z_1)$ and $f_{Y_2|W}(y_2 \mid w)$ from observables where Y_2 is only measured with error. Note that under Assumptions 2.2-2.4, we can express the conditional density of Y^* given W as an integral equation

$$f_{Y^*|W}(y^* \mid w) = \int_{\mathcal{Y}_2} f_{Y_1|Y_2Z_1}(y_1 \mid y_2,z_1) f_{Y_2^*|Y_2Z_1}(y_2^* \mid y_2,z_1) f_{Y_2|W}(y_2 \mid w)dy_2. \tag{9}$$

Therefore, the problem of identifying the densities $f_{Y_1|Y_2Z_1}(y_1 \mid y_2,z_1)$ and $f_{Y_2|W}(y_2 \mid w)$ becomes the problem of finding the unique solution to the integral equation (9) where $f_{Y^*|W}(y^* \mid w)$ is directly observable from data. We now provide additional conditions that ensure the solution to the integral equation (9) is unique. These conditions are similar to those in Hu and Schennach (2008). Let R_1, R_2, and R_3 denote random variables with supports $\mathcal{R}_1, \mathcal{R}_2$, and $\mathcal{R}_3$, respectively. Let $L_{R_1|R_2r_3}$ denote an integral operator mapping $g \in \mathcal{G}(\mathcal{R}_2)$ to $L_{R_1|R_2r_3}g \in \mathcal{G}(\mathcal{R}_1)$ for a given r_3, defined by $[L_{R_1|R_2r_3}g](r_1) \equiv \int_{\mathcal{R}_2} f_{R_1|R_2R_3}(r_1 \mid r_2, r_3)g(r_2)dr_2$, where $\mathcal{G}(\mathcal{R}_j)$ is the corresponding function space with domain $\mathcal{R}_j$ with $j = 1, 2$.

Assumption 2.5 *The operators $L_{Y_2^*|Y_2z_1}$ and $L_{Y_2^*|Vz_1}$ are one-to-one.*

Assumption 2.6 *For any $z_1 \in \mathcal{Z}_1$ and any $\tilde{y}_2, \bar{y}_2 \in \mathcal{Y}_2$, the set $\{y_1 \in \mathcal{Y}_1 : f_{Y_1|Y_2Z_1}(y_1 \mid \tilde{y}_2, z_1) \neq f_{Y_1|Y_2Z_1}(y_1 \mid \bar{y}_2, z_1)\}$ has positive probability whenever $\tilde{y}_2 \neq \bar{y}_2$.*

Assumption 2.7 *For any given $z_1 \in \mathcal{Z}_1$, there exists a known functional $\mathcal{M}$ such that $\mathcal{M}[f_{Y_2^*|Y_2Z_1}(\cdot \mid y_2, z_1)] = y_2$ for all $y_2 \in \mathcal{Y}_2$.*

Assumption 2.5 states completeness of the family of distributions associated with the operators $L_{Y_2^*|Y_2z_1}$ and $L_{Y_2^*|Vz_1}$ (see Newey and Powell (2003) and Blundell, Chen, and Kristensen (2007) for related discussions). This can be regarded as a nonparametric rank condition for identification. Assumption 2.6 excludes constant distribution of Y_1 at different values of Y_2. For example, in the partially linear model of (1) this assumption holds unless h_0 is a constant function of Y_2. In Assumption 2.7, $\mathcal{M}$ is a general functional mapping a density to a vector, which allows for nonclassical measurement errors so that the true value Y_2 and the measurement error can be dependent. This assumption places restrictions on some measure of the location of a density, denoted by $\mathcal{M}[\cdot]$ such as the mean, mode, and quantiles of the distribution. For example, it reduces to a familiar form $E[Y_2^*|Y_2] = Y_2$ in the classical measurement error case as $Y_2^* = Y_2 + e$ where e denotes the measurement error.

Given the identifying conditions of Assumptions 2.1-2.4 (conditional (mean) independence conditions) and the conditions for identifying required densities (Assumptions 2.5-2.7), the following theorem states identifiability of the parameters in the conditional moment restriction (3).

Theorem 2.1. *Under Assumptions 2.1-2.7, the parameters $\alpha_0 \equiv (\theta_0, h_0)$ are uniquely identified from the observables (Y_1, Y_2^*, Z_1, V).*

2.3 Sufficient Conditions for Identification

In order to understand the meaning of the conditional independence assumptions in a specific context, consider a regression model of the form:

$$Y_1 = \pi(Z_1) + G(Y_2; \theta, h) + \varepsilon, \quad Y_2^* = g(Y_2, e), \quad V = r(Z_1, \eta), \tag{10}$$

where $G(\cdot)$ is known up to (θ, h), Y_1, Z_1 and V are observed random variables, Y_2 is the true endogenous regressor, Y_2^* is the observed mismeasured regressor, and g and r denote true structural functions. ε, e and η are unobserved causes of Y_1, Y_2^* and V, respectively. e can be also interpreted as the measurement error on Y_2. A simple example of the measurement error is an additive error as $Y_2^* = Y_2 + e$. Inclusion of the exogenous variables Z_1 in r clearly illustrates that the control variable V does not need to be independent of the exogenous variables in this setting.

Below we provide sufficient conditions for the identifying assumptions (Assumptions 2.1-2.4) in terms of the model (10). Let $\mathcal{E}, \Xi$, and Υ be the support of the disturbances ε, e and η, respectively. Assumptions below state underlying conditions for unobservable causes such that a control variable is sufficient to control for both endogeneity and measurement error on the regressor Y_2. Recall $W \equiv (Z_1', V')'$.

Assumption 2.1S $Y_2 \perp \varepsilon \mid (Z_1, \eta)$ *for all* $(Y_2, Z_1, \varepsilon, \eta) \in \mathcal{Y}_2 \times \mathcal{Z}_1 \times \mathcal{E} \times \Upsilon$.

Assumption 2.2S $\varepsilon \perp e \mid (Y_2, W)$ *for all* $(Y_2, W, \varepsilon, e) \in \mathcal{Y}_2 \times \mathcal{W} \times \mathcal{E} \times \Xi$.

Assumption 2.3S $\varepsilon \perp \eta \mid (Y_2, Z_1)$ *for all* $(Y_2, Z_1, \varepsilon, \eta) \in \mathcal{Y}_2 \times \mathcal{Z}_1 \times \mathcal{E} \times \Upsilon$.

Assumption 2.4S $e \perp \eta \mid (Y_2, Z_1)$ *for all* $(Y_2, Z_1, e, \eta) \in \mathcal{Y}_2 \times \mathcal{Z}_1 \times \Xi \times \Upsilon$.

Theorem 2.2. *Assumption 2.1 (i) and Assumptions 2.2 - 2.4 are implied by Assumptions 2.1S-2.4S, respectively.*

Assumption 2.1S forms the moment condition $E[(Y_2 - E[Y_2 \mid W])(Y_1 - G(Y_2; \theta, h)) \mid W] = 0$ as discussed in Section 2.2.

Figure 1 provides a graphical depiction of a structure that is consistent with Assumptions 2.1S-2.4S. In the figure dashed circles denote unobservable random variables and complete circles denote observable random variables. Arrows denote direct causal relationships. Straight lines without arrows denote dependence between variables (see Pearl (2009) for graphical depictions of other possible structures). For simplicity, assume there is no exogenous regressor, Z_1, for a moment. Let ν be the unobserved cause of Y_2, which depends on η. Then Y_2 is endogenous because of the common cause, η, between ν and ε. V, observable proxy to η, controls for endogeneity by ensuring conditional independence between Y_2 and ε given V. Moreover, Y_2 is a dashed circle because it is not observed. Instead, error-laden Y_2^* is observed. Key conditions for the identification of the effect of Y_2 on Y_1 are: First, ν needs to be independent of unobserved drivers of dependent variable ε conditioning on V (or η); Second, given Y_2 and V, measurement error e needs to be independent of ε; Third, given Y_2, η needs to be independent of ε; Fourth, given Y_2, η needs to be independent of e. The first condition is required to control for endogeneity, while last three conditions are utilized to control for measurement error. Note that as in Figure 1, we allow for nonclassical measurement error since the true regressor and the measurement error can be dependent.

One important literature that utilizes the conditional independence is the problem of estimating the effect of family income on children's health (often called the *gradient*). This has been studied in Case, Lubotsky, Paxson (2002), Currie and Stabile (2003), and Condliffe and Link (2008) among others. Figure 2 shows graphical depiction of the relationship between family income and children's health. Because endogeneity comes from the common cause (i.e., parental cognitive ability) between household earning potential and children's health determinants, an available proxy to parental cognitive ability such as parental education can solve the endogeneity problem. These studies, however, do not consider possible measurement error in family income in nonlinear models. In our framework, although family income is observed with error, we do not require additional measurement on family income or other excluded instrument. Here, one control variable could be sufficient to control for both endogeneity and measurement error if the variable and the structure of the model satisfy our identification conditions.

It is interesting to see that Assumption 2.1S is equivalent to Assumption 2.1S$'$ below since $V = r(Z_1, \eta)$, and Assumption 2.3S is equivalent to Assumption 2.3S$'$ below by Lemma 4.2 (i) of Dawid (1979a).

Assumption 2.1S$'$ $Y_2 \perp \varepsilon \mid (Z_1, V)$ *for all* $(Y_2, W, \varepsilon) \in \mathcal{Y}_2 \times \mathcal{W} \times \mathcal{E}$.

Assumption 2.3S$'$ $\varepsilon \perp V \mid (Y_2, Z_1)$ *for all* $(Y_2, W, \varepsilon) \in \mathcal{Y}_2 \times \mathcal{W} \times \mathcal{E}$.

One may conclude that Assumption 2.1S$'$ and 2.3S$'$ imply ε is jointly independent of Y_2 and V. If this is indeed right, then these assumptions exclude endogeneity of Y_2 as well as dependence of V on ε. As a result, the assumptions seem contradictory to the model (10). However, this statement is one of common fallacious arguments. The following lemma built on Dawid (1979b) clarifies the implication of Assumptions 2.1S$'$ and 2.3S$'$ (equivalently Assumptions 2.1S and 2.3S).

Lemma 2.3. *Assumptions 2.1S$'$ and 2.3S$'$ hold if and only if* $\varepsilon \perp (Y_2, V) \mid T$ *where* T *is the information in common between* (Y_2, Z_1) *and* (Z_1, V).

Lemma 2.3 implies that as long as (Y_2, Z_1) and (Z_1, V) share common information other than Z_1, ε is allowed to depend on Y_2 and V. See Dawid (1979b, 1980) for general interpretation of *common information* and see Appendix A for examples. Therefore, Assumptions 2.1S−2.4S are not mutually contradictory.

2.4 Triangular Simultaneous Equations with Conditional Mean Independence

Here we discuss our identification conditions for the triangular simultaneous equations model (4) with measurement error. Figure 3 describes a version of the model. Again, assume there is no exogenous regressor, Z_1 for ease of notation. Y_2 is endogenous because its cause V and the unobserved cause of Y_1 are interdependent. In addition, the endogenous true Y_2 is observed with error e, as Y_2^*. Typically, the excluded instrument $Z = Z_2$ is assumed to only affect Y_2 where η is the underlying unobserved cause of Z. The first stage of the triangular system allows us to recover the cause V from Y_2 and Z, and we can use this generated V as the control variable. When this control variable V satisfies the required conditions in Sections 2.1-2.2 above, the causal effect of Y_2 on Y_1 can be identified.

A number of applied studies have considered the standard instrumental variable approach to control for both endogeneity and measurement error in linear parametric models. Butcher and Case (1994), for example, consider the effect of women's education on earnings. As they mention, education is endogenous because it is correlated with unobserved ability. Here the completed education variable may be reported with error. They state that sibling sex composition may be used as an instrument to estimate returns to education if it is correlated with educational attainment and uncorrelated with measurement error (Butcher and Case 1994, p. 554). Figure 4 depicts the possibility of using sibling sex composition as a potential instrument. In our setting we would recover the control variable from the first stage equation that includes the sibling sex composition as the excluded instrument. As long as the control variable satisfies the identification conditions in Sections 2.1-2.2, we can identify the causal effect. Thus, our result can be interpreted as a semiparametric analog of their result in their linear parametric model.

There are several interesting features of our framework with this model. First, the reduced form equation provides a source from which the control variate V can be obtained. Second, even though we know what should be the control variable (the conditional distribution of Y_2 given Z, $F_{Y_2|Z}$), a plug-in method to obtain the control variable such as $V \equiv F_{Y_2|Z}(Y_2 \mid Z)$ is not feasible because Y_2 is only measured with error. Below we show that in the presence of measurement error in Y_2, the CDF $F_{Y_2|Z}$ can be still obtained from the data (Y_2^*, Z) with additional assumptions. Moreover, in our setting the identification of $F_{Y_2|Z}$ suffices for recovering the structural parameters although we do not recover individual observations of V. This is because we need only the estimates of the relevant density functions to approximate the population criterion function (8) in our estimation (see Section 3 below). Whether or not individual observations are available is not important for our approach.

Several different approaches can be used to recover $F_{Y_2|Z}$ under alternative sets of assumptions (e.g. repeated measurements, existence of additional instruments, or auxiliary data). We consider two important cases below.

2.4.1 Using Repeated Measurements

Assume we observe two repeated measurements of Y_2 as $Y_{2a}^* = Y_2 + e_a$ and $Y_{2b}^* = Y_2 + e_b$ where e_a and e_b are measurement errors. Then $F_{Y_2|Z}(y_2 \mid z)$ can be identified from the repeated measurements, using a similar argument to (e.g.) Li and Vuong (1998) and Schennach (2004). The following lemma provides a result.

Lemma 2.4. *(i) Suppose that V is a scalar, r is strictly monotonic in V, and Z is independent of V in model (4). Then we have $V = F_{Y_2|Z}(Y_2|Z)$.*

(ii) Suppose further that $E[e_a \mid Y_{2b}^*] = 0$, e_b is independent of (Y_2, Z) and $E[\exp(\mathrm{i}\zeta Y_{2b}^*)]$ is nonvanishing for any real ζ. Then $F_{Y_2|Z}(y_2 \mid z)$ is identified from the observables (Y_{2a}^*, Y_{2b}^*, Z). In

particular we obtain

$$\begin{aligned} & F_{Y_2|Z}(y_2 \mid z) \\ = \quad & \frac{1}{2} + \frac{1}{2\pi}\int_0^{\infty} \frac{E[\exp(-\mathrm{i}\zeta Y_2) \mid Z = z]\exp(\mathrm{i}\zeta y_2) - E[\exp(\mathrm{i}\zeta Y_2) \mid Z = z]\exp(-\mathrm{i}\zeta y_2)}{\mathrm{i}\zeta}\mathrm{d}\zeta \end{aligned} \tag{11}$$

where

$$E[\exp(\mathrm{i}\zeta Y_2) \mid Z] \quad = \quad \frac{E[\exp(\mathrm{i}\zeta Y_{2b}^*) \mid Z]}{E[\exp(\mathrm{i}\zeta Y_{2b}^*)]}\exp\left(\int_0^{\zeta} \frac{\mathrm{i}E[Y_{2a}^*\exp(\mathrm{i}\xi Y_{2b}^*)]}{E[\exp(\mathrm{i}\xi Y_{2b}^*)]}\mathrm{d}\xi\right).$$

Note that in Lemma 2.4 the assumption $E[e_a \mid Y_{2b}^*] = 0$ states the first measurement error has conditional mean zero given Y_{2b}^*. The assumption $e_b \perp (Y_2, Z)$ implies the independence of the second measurement error and the true value. Here the second measurement error does not need to have mean zero. It allows for systematic drift term in the second measurement Y_{2b}^*, which may be a useful property when panel data is used for estimation. Lemma 2.4 (or other equivalent results) allows us to recover values of the control variable for any given values of (y_2, z) when Y_2 is only measured with error. In particular we can estimate $F_{Y_2|Z}(y_2 \mid z)$ as a sample analogue to (11) by approximating the (conditional) expectations with corresponding sample (conditional) means.

2.4.2 Using Instrumental Variables

Suppose we observe instrumental variables U for the unobservable Y_2. Then using a similar argument in Theorem 2.1, $F_{Y_2|Z}$ can be identified from the observables (Y_2^*, Z, U). To see this we rewrite the CDF as

$$F_{Y_2|Z}(y_2 \mid z) \quad \equiv \quad \int_{-\infty}^{y_2} f_{Y_2|Z}(\tilde{y}_2 \mid z)\mathrm{d}\tilde{y}_2$$

where

$$f_{Y_2|Z}(y_2 \mid z) \quad = \quad \frac{f_{Y_2Z}(y_2, z)}{f_Z(z)}.$$

Then since $f_Z(z)$ is identified from the data, the identification of $F_{Y_2|Z}(y_2 \mid z)$ rests on the identification of $f_{Y_2Z}(y_2, z)$. We state conditions similar to Assumptions 2.2-2.7 for the identification of $f_{Y_2Z}(y_2, z)$.

Assumption 2.8 *$Z \perp Y_2^* \mid (Y_2, U)$ for all $(U, Y_2, Y_2^*, Z) \in \mathcal{U} \times \mathcal{Y}_2 \times \mathcal{Y}_2^* \times \mathcal{Z}$.*

Assumption 2.9 *$Z \perp U \mid Y_2$ for all $(U, Y_2, Z) \in \mathcal{U} \times \mathcal{Y}_2 \times \mathcal{Z}$.*

Assumption 2.10 *$Y_2^* \perp U \mid Y_2$ for all $(Y_2, Y_2^*, U) \in \mathcal{Y}_2 \times \mathcal{Y}_2^* \times \mathcal{U}$.*

Assumption 2.11 *The operators $L_{Y_2^*|Y_2}$ and $L_{Y_2^*|U}$ are one-to-one.*

Assumption 2.12 *For any* $\tilde{y}_2, \bar{y}_2 \in \mathcal{Y}_2$*, the set* $\{z \in \mathcal{Z} : f_{Z|Y_2}(z \mid \tilde{y}_2) \neq f_{Z|Y_2}(z \mid \bar{y}_2)\}$ *has positive probability whenever* $\tilde{y}_2 \neq \bar{y}_2$*.*

Assumption 2.13 *There exists a known functional* $\tilde{\mathcal{M}}$ *such that* $\tilde{\mathcal{M}}[f_{Y_2^*|Y_2}(\cdot \mid y_2)] = y_2$ *for all* $y_2 \in \mathcal{Y}_2$*.*

The following lemma provides identification of $f_{Y_2 Z}(y_2, z)$ which is sufficient for the identification of the CDF $F_{Y_2|Z}(y_2 \mid z)$ given that $f_Z(z)$ is observable from the data.

Lemma 2.5. *Under Assumptions 2.8-2.13, the density functions* $(f_{Y_2 Z}, f_{Y_2^*|Y_2}, f_{U|Y_2})$ *are uniquely identified from the observables* (Y_2^*, Z, U)*.*

3 Estimation

Based on our identification results in Section 2 we propose a sieve-based estimator of $\alpha_0 \equiv (\theta_0, h_0)$ for the model of (3). First we focus on the case for which the control variables V are observables and then show how the setting can extend to the case for which V are generated variables as the triangular simultaneous equations model (4).

In our approach, from observations on Y_1, Y_2^*, Z_1, and V, we first estimate the unknown densities $f_{Y_1|Y_2 Z_1}(y_1 \mid y_2, z_1)$ and $f_{Y_2|W}(y_2 \mid w)$ in the conditional moment function (6) using a sieve MLE, and in the second stage we estimate parameters of interest (θ_0, h_0) using a sieve MD estimation.

We introduce additional notation. Let

$$f_{Y^*|W}(y^* \mid w; \beta_0) = \int_{\mathcal{Y}_2} f_{Y_1|Y_2 Z_1}(y_1 \mid y_2, z_1) f_{Y_2^*|Y_2 Z_1}(y_2^* \mid y_2, z_1) f_{Y_2|W}(y_2 \mid w) dy_2$$

and let $\beta_0 \equiv (f_1, f_2, f_3) \in \mathcal{B} \equiv \mathcal{F}_1 \times \mathcal{F}_2 \times \mathcal{F}_3$ denote the first stage parameters such that $f_1 \equiv f_{Y_1|Y_2 Z_1}(y_1 \mid y_2, z_1)$, $f_2 \equiv f_{Y_2^*|Y_2 Z_1}(y_2^* \mid y_2, z_1)$, and $f_3 \equiv f_{Y_2|W}(y_2 \mid w) = f_{Y_2|V Z_1}(y_2 \mid v, z_1)$. Define a weighted Hölder ball of radius c by $\Lambda_c^{\gamma,\omega}(\mathcal{V}) \equiv \{\mathrm{g} \in \Lambda^{\gamma,\omega}(\mathcal{V}) : \|\mathrm{g}\|_{\Lambda^{\gamma,\omega}} \leq c < \infty\}$, where $\Lambda^{\gamma,\omega}(\mathcal{V})$ is the weighted Hölder space of order $\gamma > 0$ with a weight function ω (see Ai and Chen (2003) and Chen, Hong, and Tamer (2005) for more details and examples).

Following assumptions impose restrictions on the parameter spaces $\mathcal{F}_1, \mathcal{F}_2$, and $\mathcal{F}_3$.

Assumption 3.1 $f_1 \in \Lambda_c^{\gamma_1,\omega}(\mathcal{Y}_1 \times \mathcal{Y}_2 \times \mathcal{Z}_1)$ *where* $\gamma_1 > 1$ *and* $\int_{\mathcal{Y}_1} f_1(y_1 \mid y_2, z_1) dy_1 = 1$ *for all* $y_2 \in \mathcal{Y}_2, z_1 \in \mathcal{Z}_1$*.*

Assumption 3.2 $f_2 \in \Lambda_c^{\gamma_1,\omega}(\mathcal{Y}_2^* \times \mathcal{Y}_2 \times \mathcal{Z}_1)$ *where* $\gamma_1 > 1$ *and* $\int_{\mathcal{Y}_2^*} f_2(y_2^* \mid y_2, z_1) dy_2^* = 1$ *for all* $y_2 \in \mathcal{Y}_2, z_1 \in \mathcal{Z}_1$*.*

Assumption 3.3 $f_3 \in \Lambda_c^{\gamma_1,\omega}(\mathcal{Y}_2 \times \mathcal{V} \times \mathcal{Z}_1)$ *where* $\gamma_1 > 1$ *and* $\int_{\mathcal{Y}_2} f_3(y_2 \mid v, z_1) dy_2 = 1$ *for all* $v \in \mathcal{V}, z_1 \in \mathcal{Z}_1$*.*

Then the parameter spaces for the density functions can be defined respectively as:

$$\begin{aligned}
\mathcal{F}_1 &= \{f_1(\cdot \mid \cdot, \cdot) : \textit{Assumption 3.1 holds}\}, \\
\mathcal{F}_2 &= \{f_2(\cdot \mid \cdot, \cdot) : \textit{Assumptions 2.5, 2.7, and 3.2 hold}\}, \\
\mathcal{F}_3 &= \{f_3(\cdot \mid \cdot, \cdot) : \textit{Assumptions 2.5 and 3.3 hold}\}.
\end{aligned}$$

Let p_j denote a sequence of known basis functions (such as power series, splines, Fourier series, etc.). A tensor-product multivariate linear sieve basis, denoted by $p^l(\cdot, \cdot, \cdot) = (p_1(\cdot, \cdot, \cdot), \ldots, p_l(\cdot, \cdot, \cdot))'$ is used to approximate functions of three variables. Suppose we have n observations of the sample $\{y_{1i}, y_{2i}^*, z_{1i}, v_i\}$. Then based on the identification results of Section 2.2, we estimate β_0 using the sieve MLE as

$$\begin{aligned}
\hat{\beta}_n &= (\hat{f}_{1n}, \hat{f}_{2n}, \hat{f}_{3n}) \qquad (12) \\
&= \arg \max_{(f_1, f_2, f_3) \in \mathcal{B}_n} \frac{1}{n} \sum_{i=1}^{n} \ln \int_{\mathcal{Y}_2} f_1(y_{1i} \mid y_2, z_{1i}) f_2(y_{2i}^* \mid y_2, z_{1i}) f_3(y_2 \mid v_i, z_{1i}) dy_2,
\end{aligned}$$

where $\mathcal{B}_n = \mathcal{F}_{1n} \times \mathcal{F}_{2n} \times \mathcal{F}_{3n}$ is a sieve space approximating $\mathcal{B}$ with the sample size n, and where $\mathcal{F}_{1n}, \mathcal{F}_{2n}$ and $\mathcal{F}_{3n}$ are defined as:

$$\begin{aligned}
\mathcal{F}_{1n} &= \{f_1(y_1 \mid y_2, z_1) = p^{l_{n1}}(y_1, y_2, z_1)'\pi^1 \ \textit{ for all } \ \pi^1 \ \textit{ s.t. Assumption 3.1 holds}\}, \\
\mathcal{F}_{2n} &= \{f_2(y_2^* \mid y_2, z_1) = p^{l_{n2}}(y_2^*, y_2, z_1)'\pi^2 \ \textit{ for all } \ \pi^2 \ \textit{ s.t. Assumptions 2.5, 2.7, and 3.2 hold}\}, \\
\mathcal{F}_{3n} &= \{f_3(y_2 \mid v, z_1) = p^{l_{n3}}(y_2, v, z_1)'\pi^3 \ \textit{ for all } \ \pi^3 \ \textit{ s.t. Assumptions 2.5 and 3.3 hold}\}.
\end{aligned}$$

Using these sieve approximations, in the first stage, we estimate $f_{Y_1|Y_2Z_1}(y_1 \mid y_2, z_1)$ and $f_{Y_2|VZ_1}(y_2 \mid v, z_1)$ from (12). Using these estimated densities we construct $\hat{m}(w_i, \alpha)$ as the plug-in estimator of $m(w_i, \alpha)$:

$$\begin{aligned}
&\hat{m}(w_i, \alpha) \qquad (13) \\
&\equiv \int_{\mathcal{Y}_2} \left[\int_{\mathcal{Y}_1} \rho(y_1, y_2, z_{1i}, v_i, \theta, h) \hat{f}_{Y_1|Y_2Z_1}(y_1 \mid y_2, z_{1i}) dy_1 \right] \hat{f}_{Y_2|VZ_1}(y_2 \mid v_i, z_{1i}) dy_2
\end{aligned}$$

where $\hat{f}_{1n} = \hat{f}_{Y_1|Y_2Z_1}(y_1 \mid y_2, z_{1i})$ and $\hat{f}_{3n} = \hat{f}_{Y_2|VZ_1}(y_2 \mid v_i, z_{1i})$ are obtained from (12). Let $\mathcal{H}$ be a space of smooth functions (e.g. Hölder ball) that contains the true h_0 and let $\mathcal{H}_n$ be some finite-dimensional approximation space that becomes dense in $\mathcal{H}$ as the sample size n tends to infinity (e.g., Fourier series, power series, splines, wavelets, etc.). Then in the second stage, we obtain the (penalized) sieve MD estimator of $\alpha_0 \equiv (\theta_0, h_0)$ as

$$\hat{\alpha}_n \equiv (\hat{\theta}_n, \hat{h}_n) = \arg \inf_{\alpha = (\theta, h) \in \Theta \times \mathcal{H}_n} \hat{Q}_n(\alpha), \qquad (14)$$

where

$$\hat{Q}_n(\alpha) \quad \equiv \quad \left\{ \frac{1}{n} \sum_{i=1}^{n} \hat{m}(W_i, \alpha)' [\hat{\Sigma}(W_i)]^{-1} \hat{m}(W_i, \alpha) + \lambda_n \hat{P}_n(h) \right\}$$

and where $\hat{\Sigma}(W)$ is a consistent estimator of $\Sigma(W)$ (a positive-definite weighting matrix), $\lambda_n \geq 0$ is a penalization tuning parameter such that $\lambda_n = o(1)$, and $\hat{P}_n(h) \geq 0$ is a possibly random penalty function as in Chen and Pouzo (2012).

Our estimation method can accommodate the case when V are generated variables. For example, in the case of the triangular model (4) we can estimate the densities using a sieve MLE as

$$\begin{aligned} \hat{\beta}_n &= (\hat{f}_{1n}, \hat{f}_{2n}, \hat{f}_{3n}) \\ &= \arg \max_{(f_1, f_2, f_3) \in \mathcal{B}_n} \frac{1}{n} \sum_{i=1}^{n} \ln \int_{\mathcal{Y}_2} f_1(y_{1i} \mid y_2, z_{1i}) f_2(y_{2i}^* \mid y_2, z_{1i}) f_3(y_2 \mid \hat{v}_i(y_2), z_{1i}) dy_2, \end{aligned} \tag{15}$$

where we define $\hat{v}_i(y_2) \equiv \hat{F}_{Y_2|Z}(y_2|z_i)$ and $\hat{F}_{Y_2|Z}(y_2|z_i)$ is an estimator of $F_{Y_2|Z}(y_2|z_i)$ based on (e.g.) the identification results in Section 2.4.1 and Section 2.4.2.

Similarly, for the triangular model case we can estimate the conditional mean function as

$$\begin{aligned} &\hat{m}(\hat{w}_i, \alpha) \\ &\equiv \int_{\mathcal{Y}_2} \left[\int_{\mathcal{Y}_1} \rho(y_1, y_2, z_{1i}, \hat{v}_i(y_2), \theta, h) \hat{f}_{Y_1|Y_2Z_1}(y_1 \mid y_2, z_{1i}) dy_1 \right] \hat{f}_{Y_2|VZ_1}(y_2 \mid \hat{v}_i(y_2), z_{1i}) dy_2 \end{aligned}$$

where $\hat{f}_{1n} = \hat{f}_{Y_1|Y_2Z_1}(y_1 \mid y_2, z_{1i})$ and $\hat{f}_{3n} = \hat{f}_{Y_2|VZ_1}(y_2 \mid \hat{v}_i(y_2), z_{1i})$ are obtained from (15). The sieve MD estimator is then obtained by replacing $\hat{m}(w_i, \alpha)$ with $\hat{m}(\hat{w}_i, \alpha)$ in the construction of the sample criterion function $\hat{Q}_n(\alpha)$ in (14).

4 Consistency

We obtain consistency of our estimator $\hat{\alpha}_n \equiv (\hat{\theta}_n, \hat{h}_n)$ defined in the equation (14) allowing for a penalty function. Let $(Y^{*\prime}, W')'$ be a vector of observed variables for $Y^* \in \mathcal{Y}^*, W \in \mathcal{W}$. We denote the smoothing parameter in the first-stage estimation by l_n which is the total number of sieve coefficients, $l_n = l_{n1} + l_{n2} + l_{n3}$. We also denote another smoothing parameter in the sieve approximation for h by $k_n = \dim(\mathcal{H}_n)$. Define $\|\beta\|_{s,\beta} \equiv \|f_1\|_{\infty,\omega} + \|f_2\|_{\infty,\omega} + \|f_3\|_{\infty,\omega}$ where $\|g\|_{\infty,\omega} \equiv \sup_\xi |g(\xi)\omega(\xi)|$ with a weight function $\omega(\xi) = (1 + \|\xi\|_E^2)^{-\varsigma/2}, \varsigma > \gamma_1 > 0$. Note that the generic variable ξ depends on the domain of g (e.g., when $g = f_1$, $\xi = (y_1, y_2, z_1)$). We also denote $\|\alpha\|_{s,\alpha} \equiv \|\theta\|_E + \|h\|_{s,\alpha}$. For any vector- or matrix-valued A, sometimes we use $|A| = ||A||_E = \sqrt{tr(A'A)}$ for ease of notation. We make the following assumptions. A subset of the assumptions are from Newey and Powell (2003) and Chen and Pouzo (2012).

Assumption 4.1 (i) *The data* $\{(Y_i^*, W_i)_{i=1}^n\}$ *are i.i.d.* (ii) *The density of* $(Y^{*\prime}, W')'$, f_{Y^*W}, *sat-*

isfies $\int \omega(y^*, w) f_{Y^*W}(y^*, w) d(y^*, w) < \infty$. (iii) $E[|\ln f_{Y^*|W}(Y^* \mid W)|^2]$ *is bounded.* (iv) *There exists a measurable function* $C_1(y^*, w)$ *with* $E[|C_1(Y^*, W)|^2] < \infty$ *such that for any* $\bar{\beta} = (\bar{f}_1, \bar{f}_2, \bar{f}_3)' \in \mathcal{B}$,

$$\left| \frac{f^{|1|}_{Y^*|W}(y^* \mid w; \bar{\beta}, \bar{\omega})}{f_{Y^*|W}(y^* \mid w; \bar{\beta})} \right| \leq C_1(y^*, w),$$

where the path-wise first derivative $f^{|1|}_{Y^*|W}(y^* \mid w; \bar{\beta}, \bar{\omega})$ *as well as the term* $\bar{\omega}$ *is defined in the proof of Theorem 4.1.*

Assumption 4.2 (i) $\mathcal{A} \equiv \Theta \times \mathcal{H}$, Θ *is a compact subset of* $\mathcal{R}^{d_\theta}$, *and* $\mathcal{H} \subseteq \mathbf{H}$, $\mathbf{H}$ *is a separable Banach space under a metric* $\|\cdot\|_{s,\alpha}$. (ii) $\mathcal{B} \equiv \mathcal{F}_1 \times \mathcal{F}_2 \times \mathcal{F}_3$ *is compact under a metric* $\|\cdot\|_{s,\beta}$ *and Assumptions 3.1-3.3 hold for* (f_1, f_2, f_3) *in a neighborhood of* β_0. (iii) $E[\rho(X, \alpha_0) \mid W] = 0$, *and* $\|\theta_0 - \theta\|_E + \|h_0 - h\|_{s,\alpha} = 0$ *for any* $\alpha = (\theta, h) \in \mathcal{A}$ *with* $E[\rho(X, \alpha) \mid W] = 0$.

Assumption 4.3 (i) $\mathcal{A}_n \equiv \Theta \times \mathcal{H}_n, n \geq 1$, *are the sieve space which is a nonempty closed subset of* $(\mathcal{A}, \|\cdot\|_{s,\alpha})$ *satisfying* $\mathcal{H}_n \subseteq \mathcal{H}_{n+1} \subseteq \mathcal{H}$, *and there exists a function* $\Pi_n h_0 \in \mathcal{H}_n$ *such that* $\|\Pi_n h_0 - h_0\|_{s,\alpha} = o(1)$ *with* $k_n/n \to 0$. (ii) $E[m(W, \alpha)'\Sigma(W)^{-1} m(W, \alpha)]$ *is continuous at* α_0 *under* $\|\cdot\|_{s,\alpha}$. (iii) $\mathcal{B}_n \equiv \mathcal{F}_{1n} \times \mathcal{F}_{2n} \times \mathcal{F}_{3n}, n \geq 1$, *are the sieve space which is a nonempty closed subset of* $(\mathcal{B}, \|\cdot\|_{s,\beta})$ *satisfying* $\mathcal{F}_{i\,n} \subseteq \mathcal{F}_{i\,n+1} \subseteq \mathcal{F}_i$, $i \in \{1, 2, 3\}$, *and there exists a function* $\Pi_n \beta_0 \in \mathcal{B}_n$ *such that* $\|\Pi_n \beta_0 - \beta_0\|_{s,\beta} = o(1)$ *with* $l_n/n \to 0$.

Assumption 4.4 *One of the following conditions holds:* (i) $\lambda_n = 0$. (ii) $\lambda_n \sup_{h \in \mathcal{H}_n} |\hat{P}_n(h) - P(h)| = O_P(\lambda_n)$ *and* $\lambda_n |P(\Pi_n h_0) - P(h_0)| = O(\lambda_n)$, *with* $\lambda_n > 0$, $\lambda_n = o(1)$ *and* $P(\cdot)$ *a non-negative real-valued measurable function of* $h \in \mathcal{H}$, $P(h_0) < \infty$.

Assumption 4.1 (i) is about the data. Assumption 4.1 (ii) imposes a tail-behavior restriction on f_{Y^*W}. Assumptions 4.1 (iii)-(iv) impose an envelop condition on the first derivative of the log likelihood function, $\ln f_{Y^*|W}(y^* \mid w)$. Assumptions 4.2 (i)-(ii) impose restrictions on the parameter spaces, $\mathcal{A}$ and $\mathcal{B}$. Assumption 4.2 (iii) is an identification condition of the parameter α_0. Assumptions 4.3 (i) and (iii) are the definitions of sieve spaces, stating the sieves can approximate the true α_0 and β_0 arbitrarily well. Assumption 4.3 (ii) is a sufficient condition for $E[\|m(W, \Pi_n \alpha_0)\|_E^2] = o(1)$ for $\Pi_n \alpha_0 \equiv (\theta_0, \Pi_n h_0) \in \mathcal{A}_n$. Assumption 4.4 (i) allows for no penalty case and Assumption 4.4 (ii) simultaneously states a property of the tuning parameter and penalty function when $\lambda_n > 0$. It states that $\hat{P}_n(h)$ is a consistent estimator of $P(h)$ and $|P(\Pi_n h_0) - P(h_0)| = O(1)$ is satisfied if $P(\cdot)$ is continuous at h_0. Let $\{\delta_{m,n}\}_n$ be real-valued positive sequence decreasing to zero as $n \to \infty$, denoting the convergence rate of $\hat{m}(w, \alpha)$ to $m(w, \alpha)$, namely, $\sup_{\alpha \in \mathcal{A}_n} E[\|\hat{m}(W, \alpha) - m(W, \alpha)\|_E^2] \equiv O(\delta_{m,n}^2)$. We denote $\xi_{0n} \equiv \sup_{(\xi_1, \xi_2, \xi_3) \in ((\mathcal{Y}_1 \times \mathcal{Y}_2 \times \mathcal{Z}_1) \cup (\mathcal{Y}_2^* \times \mathcal{Y}_2 \times \mathcal{Z}_1) \cup (\mathcal{Y}_2 \times \mathcal{V} \times Z_1))} \|p^{l_n}(\xi_1, \xi_2, \xi_3)\|_E$, which is nondecreasing in l_n, and denote $\Pi_n \beta \equiv (\Pi_n f_1, \Pi_n f_2, \Pi_n f_3) \in \mathcal{B}_n \equiv \mathcal{F}_{1n} \times \mathcal{F}_{2n} \times \mathcal{F}_{3n}$. Let $\{b_{m,l_n}\}_n$ be a real-valued positive sequence decreasing to zero as $l_n \to \infty$, denoting the approximation bias of $m(\cdot, \alpha)$ using the sieve ML estimator as (13).

Assumption 4.5 (i) $\hat{\Sigma}(w) = \Sigma(w) + o_p(1)$ *uniformly over* $w \in \mathcal{W}$. (ii) $\hat{\Sigma}(w)$ *is positive definite, and its smallest and largest eigenvalues are finite positive uniformly in* $w \in \mathcal{W}$ *with probability approaching one.* (iii) $\Sigma(w)$ *is positive definite, and its smallest and largest eigenvalues are finite positive uniformly in* $w \in \mathcal{W}$.

Assumption 4.6 (i) *There is a finite constant* c *such that* $\sup_{\alpha \in \mathcal{A}_n} \sup_w Var[\rho_j(X, \alpha) \mid W = w] \leq c < \infty$ *for all* $j = 1, \ldots, d_\rho$. (ii) *For any* $g \in \{m(\cdot, \alpha) : \alpha \in \mathcal{A}_n\}$, *there exists* $\tilde{g}(W) \equiv \int \rho(y, W, \alpha) f_{Y|W}(y|W; \tilde{\beta}) dy$ *for some* $\tilde{\beta} \in \mathcal{B}_n$ *such that, uniformly over* $\alpha \in \mathcal{A}_n$, $E[|g(W) - \tilde{g}(W)|^2] = O(b^2_{m,l_n})$ *for the* $p^{l_n}(\xi_1, \xi_2, \xi_3)$ *sieve with* $\xi^2_{0n} = O(l_n)$. (iii) *There are finite constants* $c_1, c_2 > 0$ *such that uniformly over* $\alpha \in \mathcal{A}_n$, $c_1 E[\|\hat{m}(W, \alpha)\|^2_E] \leq n^{-1} \sum_{i=1}^n \|\hat{m}(W_i, \alpha)\|^2_E \leq c_2 E[\|\hat{m}(W, \alpha)\|^2_E]$ *with probability approaching one as* $n \to \infty$.

Assumptions 4.5 is about the weighting matrix $\Sigma(w)$ and its consistent estimator $\hat{\Sigma}(w)$. Assumption 4.6 (i) requires a finite conditional variance of the residual function ρ. Assumption 4.6 (ii) quantifies the bias of the estimator $\tilde{g}(w)$ and is satisfied by the class of typical smooth functions. The conditional mean function estimator $\hat{m}(W, \alpha)$ defined in (13) using the first stage sieve ML estimator can be shown to satisfy $\sup_{\alpha \in \mathcal{A}_n} E[\|\hat{m}(W, \alpha) - m(W, \alpha)\|^2_E] \equiv O(\delta^2_{m,n})$ with $\delta^2_{m,n} = \max\{\frac{l_n}{n}, b^2_{m,l_n}\}$ under Assumptions 4.1 and 4.6.

It is worth noting that if the original parameter space $\mathcal{B}$ is too large, it is useful to introduce another penalty function for the first-stage parameter, β, and estimate it via a penalized sieve MLE as in Shen (1997). For the sake of concise results, we maintain the assumption that $\mathcal{B}$ is a compact space. Define $\mathcal{A}_n^{M_0} \equiv \{\alpha \in \mathcal{A}_n : \lambda_n P(h) \leq \lambda_n M_0\}$ for a finite $M_0 \equiv M_0(\varepsilon) > 0$ such that $\Pi_n \alpha_0 \in \mathcal{A}_n^{M_0}$ and $Pr(\hat{\alpha}_n \notin \mathcal{A}_n^{M_0}) < \varepsilon$ for all $\varepsilon > 0$ and all sufficiently large n. Then we obtain the following consistency results.

Theorem 4.1. *Let* $\hat{\alpha}_n$ *be the PSMD estimator defined in (14) with* $\lambda_n \geq 0$, $\lambda_n = o(1)$. *Suppose Assumptions 2.1-2.7, 3.1-3.3, and 4.1-4.6 hold,* $E[|m(W, \alpha)|^2]$ *is lower semicontinuous in* $\|\cdot\|_{s,\alpha}$ *on* $\mathcal{A}_n$, *and* $\max\{\delta^2_{m,n}, E\left[|m(W, \Pi_n \alpha_0)|^2\right], \lambda_n\} / \inf_{\alpha \in \mathcal{A}_n^{M_0} : \|\alpha - \alpha_0\|_{s,\alpha} \geq \varepsilon} E\left[|m(W, \alpha)|^2\right] = o(1)$ *is satisfied for any* $\varepsilon > 0$. *Then* $\|\hat{\alpha}_n - \alpha_0\|_{s,\alpha} = o_p(1)$.

See Appendix F for the proof.

5 Convergence Rate and Asymptotic Normality

To obtain the convergence rate of the second-stage sieve estimator $\hat{\alpha}_n \equiv (\hat{\theta}_n, \hat{h}_n)$ and the asymptotic normality of the finite-dimensional plug-in estimators as functionals of α such as the asymptotic normality of the finite-dimensional parameter $\hat{\theta}_n$, we extend the asymptotic results in Chen and Pouzo (2012) to the case with mismeasured endogenous regressors. Chen and Pouzo use the series least-squares estimator for the conditional mean function $m(w, \alpha)$, which is infeasible in the presence of measurement error. Nevertheless, their asymptotic results are applicable to our case because they establish the asymptotic properties of their proposed estimators, which can accommodate any consistent estimators for the conditional mean function. Therefore, we can derive the asymptotic properties of the second-stage estimator $\hat{\alpha}_n$ and its plug-in estimators as functionals of $\hat{\alpha}_n$ using Chen and Pouzo's framework.

5.1 Convergence Rate

To derive the convergence rates, we introduce additional notations and weaker metrics for which we derive the rate results, and make a few additional assumptions. Denote $\mathcal{B}_{os} \equiv \{\beta \in \mathcal{B} : \|\beta - \beta_0\|_{s,\beta} = o(1)\}$ and $\mathcal{B}_{osn} \equiv \mathcal{B}_{os} \cap \mathcal{B}_n$. For any $\beta \in \mathcal{B}_{os}$, denote the first path-wise derivative of $\ln f_{Y^*|W}(y^* \mid w; \beta_0)$ at the direction $[\beta - \beta_0]$ evaluated at β_0 by:

$$\frac{d \ln f_{Y^*|W}(y^* \mid w; \beta_0)}{d\beta}[\beta - \beta_0] \equiv \left. \frac{d \ln f_{Y^*|W}(y^* \mid w; (1-\tau)\beta_0 + \tau\beta)}{d\tau} \right|_{\tau=0}$$

almost everywhere (under the probability measure of (Y^*, W)) and for $\beta_1, \beta_2 \in \mathcal{B}_{os}$ denote

$$\frac{d \ln f_{Y^*|W}(y^* \mid w; \beta_0)}{d\beta}[\beta_1 - \beta_2] \equiv \frac{d \ln f_{Y^*|W}(y^* \mid w; \beta_0)}{d\beta}[\beta_1 - \beta_0] - \frac{d \ln f_{Y^*|W}(y^* \mid w; \beta_0)}{d\beta}[\beta_2 - \beta_0].$$

Specifically, the path-wise derivative is denoted by:

$$\begin{aligned}
& \frac{d \ln f_{Y^*|W}(y^* \mid w; \beta_0)}{d\beta}[\beta - \beta_0] \\
= \; & \frac{1}{f_{Y^*|W}(y^* \mid w; \beta_0)} \Bigg\{ \int_{\mathcal{Y}_2} [f_1(y_1|y_2,z_1) - f_{Y_1|Y_2Z_1}(y_1|y_2,z_1)] f_{Y_2^*|Y_2Z_1}(y_2^* \mid y_2, z_1) f_{Y_2|VZ_1}(y_2 \mid v, z_1) dy_2 \\
& + \int_{\mathcal{Y}_2} f_{Y_1|Y_2Z_1}(y_1 \mid y_2, z_1)[f_2(y_2^* \mid y_2, z_1) - f_{Y_2^*|Y_2Z_1}(y_2^* \mid y_2, z_1)] f_{Y_2|VZ_1}(y_2 \mid v, z_1) dy_2 \\
& + \int_{\mathcal{Y}_2} f_{Y_1|Y_2Z_1}(y_1 \mid y_2, z_1) f_{Y_2^*|Y_2Z_1}(y_2^* \mid y_2, z_1)[f_3(y_2 \mid v, z_1) - f_{Y_2|VZ_1}(y_2 \mid v, z_1)] dy_2 \Bigg\}.
\end{aligned}$$

For any $\beta_1, \beta_2 \in \mathcal{B}_{os}$, the metric is defined as

$$\|\beta_1 - \beta_2\|_\beta \equiv \sqrt{E\left\{ \left(\frac{d \ln f_{Y^*|W}(Y^* \mid W; \beta_0)}{d\beta}[\beta_1 - \beta_2] \right)^2 \right\}}.$$

Denote $\mathcal{A}_{os} \equiv \{\alpha \in \mathcal{A} : \|\alpha - \alpha_0\|_{s,\alpha} = o(1), P(h) \leq c\}$ for a constant $c > 0$ and $\mathcal{A}_{osn} \equiv \mathcal{A}_{os} \cap \mathcal{A}_n$. For any $\alpha \in \mathcal{A}_{os}$, we define the first path-wise derivative of $\rho(X, \alpha)$ at the direction $[\alpha - \alpha_0]$ evaluated at α_0 by:

$$\frac{d\rho(X, \alpha_0)}{d\alpha}[\alpha - \alpha_0] \equiv \left. \frac{d\rho(X, (1-\tau)\alpha_0 + \tau\alpha)}{d\tau} \right|_{\tau=0}$$

almost everywhere (under the probability measure of X) and for any $\alpha_1, \alpha_2 \in \mathcal{A}_{os}$ denote

$$\begin{aligned}
\frac{d\rho(X,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2] &\equiv \frac{d\rho(X,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_0] - \frac{d\rho(X,\alpha_0)}{d\alpha}[\alpha_2 - \alpha_0], \\
\frac{dm(W,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2] &\equiv E\left\{ \frac{d\rho(X,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2] \middle| W \right\}.
\end{aligned}$$

Also, for any $\alpha_1, \alpha_2 \in \mathcal{A}_{os}$, the metric $\|\cdot\|_\alpha$ is defined as

$$\|\alpha_1 - \alpha_2\|_\alpha \equiv \sqrt{E\left\{ \left(\frac{dm(W,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2]\right)' \Sigma(W)^{-1} \frac{dm(W,\alpha_0)}{d\alpha}[\alpha_1 - \alpha_2] \right\}}.$$

The metrics $\|\cdot\|_\beta$ and $\|\cdot\|_\alpha$ are weaker than the norms $\|\cdot\|_{s,\beta}$ and $\|\cdot\|_{s,\alpha}$, respectively, in the sense that $\|\cdot\|_\beta \leq \|\cdot\|_{s,\beta}$ and $\|\cdot\|_\alpha \leq \|\cdot\|_{s,\alpha}$. The convergence rates of $\hat{\beta}$ and $\hat{\alpha}$ are analyzed under the weaker metrics $\|\cdot\|_\beta$ and $\|\cdot\|_\alpha$, respectively. Therefore, given the consistency results, now we can treat $\mathcal{B}_{os}$ and $\mathcal{A}_{os}$ as the new parameter spaces, while $\mathcal{B}_{osn}$ and $\mathcal{A}_{osn}$ are considered as their sieve spaces, respectively. We make the following assumptions.

Assumption 5.1 (i) *$\ln f_{Y^*|W}(y^* \mid w;\beta)$ satisfies an envelope condition in $\beta \in \mathcal{B}_{osn}$.* (ii) *$\ln f_{Y^*|W}(y^* \mid w;\beta) \in \Lambda_c^{\gamma,\omega}(\mathcal{Y}^* \times \mathcal{W})$ for some constant $c > 0$ with $\gamma > d_{(Y^*,W)}/2$, for all $\beta \in \mathcal{B}_{osn}$, where $d_{(Y^*,W)}$ is the dimension of (Y^*, W).* (iii) *There is a constant γ_1 such that for any $\beta \in \mathcal{B}_{os}$, there exists $\Pi_n\beta \in \mathcal{B}_{osn}$ satisfying $\|\Pi_n\beta - \beta\|_\beta = O(l_n^{-\gamma_1/3})$, and $l_n^{-\gamma_1/3} = o(n^{-1/4})$.*

Assumption 5.1 (i) imposes a dominance condition on the log likelihood function and Assumption 5.1 (ii) imposes a smoothness condition on the function. Assumption 5.1 (iii) quantifies the approximation error of β by $\Pi_n\beta \in \mathcal{B}_{osn}$. This condition is usually satisfied by the commonly used sieve approximations (e.g., power series, Fourier series, splines, wavelet, etc). Let $N(\varepsilon, \mathcal{B}_{osn}, \|\cdot\|_{s,\beta})$ denote the minimal number of radius ε covering balls of $\mathcal{B}_{osn}$ under the metric $\|\cdot\|_{s,\beta}$.

Assumption 5.2 (i) *$l_n \times \ln n \times \xi_{0n}^2 \times n^{-1/2} = o(1)$.* (ii) *For a constant $c > 0$, $\ln[N(\varepsilon, \mathcal{B}_{osn}, \|\cdot\|_{s,\beta})] \leq c \times l_n \times \ln(l_n/\varepsilon)$.*

Assumption 5.3 (i) *$\mathcal{A}_{os}$ and $\mathcal{A}_{osn}$ are convex in α_0 and $m(W,\alpha)$ is continuously path-wise differentiable with respect to $\alpha \in \mathcal{A}_{os}$.* (ii) *$\mathcal{B}_{os}$ and $\mathcal{B}_{osn}$ are convex in β_0 and $f_{Y^*|W}(y^* \mid w;\beta)$ is continuously path-wise differentiable with respect to $\beta \in \mathcal{B}_{os}$.*

Assumption 5.4 (i) *There are finite constants $c_1, c_2 > 0$ such that $c_1 E[\|m(W,\alpha)\|_E^2] \leq \|\alpha - \alpha_0\|_\alpha^2 \leq c_2 E[\|m(W,\alpha)\|_E^2]$ for all $\alpha \in \mathcal{A}_{os}$.* (ii) *There are finite constants $c_1, c_2 > 0$ such that*

$$c_1 E\left\{ \ln \frac{f_{Y^*|W}(y^* \mid w;\beta_0)}{f_{Y^*|W}(y^* \mid w;\beta)} \right\} \leq \|\beta - \beta_0\|_\beta^2 \leq c_2 E\left\{ \ln \frac{f_{Y^*|W}(y^* \mid w;\beta_0)}{f_{Y^*|W}(y^* \mid w;\beta)} \right\}$$

holds for all $\beta \in \mathcal{B}_{os}$.

Assumption 5.2 (i) imposes a restriction on the divergence rate of l_n and Assumption 5.2 (ii) imposes a restriction on the size of the sieve space $\mathcal{B}_{osn}$ such that it does not grow too fast in terms of the covering number. Assumption 5.3 ensures that the weak metrics $\|\cdot\|_\beta$ and $\|\cdot\|_\alpha$ are well-defined, respectively. Assumption 5.4 imposes that the population criterion function and the log likelihood function can be approximated locally by the weak metrics.

Let $\overline{\mathbf{B}}$ denote the closure of the linear span of $\mathcal{A}_{os}-\{\alpha_0\}$ under the metric $\|\cdot\|_\alpha$ (i.e., $\overline{\mathbf{B}} = \mathcal{R}^{d_\theta}\times\overline{\Phi}$ with $\overline{\Phi}\equiv\overline{\mathcal{H}}-\{h_0\}$). Then $(\overline{\mathbf{B}},\|\cdot\|_\alpha)$ is a Hilbert space with the inner product:

$$\langle b_1, b_2\rangle_\alpha \;=\; E\left\{\left(\frac{dm(W,\alpha_0)}{d\alpha}[b_1]\right)'\Sigma(W)^{-1}\left(\frac{dm(W,\alpha_0)}{d\alpha}[b_2]\right)\right\}.$$

The path-wise derivative at α_0 is defined as

$$\frac{dm(W,\alpha_0)}{d\alpha}[\alpha-\alpha_0] \;\equiv\; \frac{dm(W,\alpha_0)}{d\theta'}(\theta-\theta_0)+\frac{dm(W,\alpha_0)}{dh}[h-h_0].$$

For each component θ_j of θ, $j=1,2,\ldots,d_\theta$, we define $\phi_j^*\in\overline{\Phi}$ as

$$\phi_j^* \;\equiv\; \arg\inf_{\phi_j\in\overline{\Phi}} E\left\{\left(\frac{dm(W,\alpha_0)}{d\theta_j}-\frac{dm(W,\alpha_0)}{dh}[\phi_j]\right)'\Sigma(W)^{-1}\left(\frac{dm(W,\alpha_0)}{d\theta_j}-\frac{dm(W,\alpha_0)}{dh}[\phi_j]\right)\right\}.$$

Define

$$\begin{aligned}\phi^* \;&=\; (\phi_1^*,\phi_2^*,...,\phi_{d_\theta}^*),\\ \frac{dm(W,\alpha_0)}{dh}[\phi^*] \;&=\; \left(\frac{dm(W,\alpha_0)}{dh}[\phi_1^*],\ldots,\frac{dm(W,\alpha_0)}{dh}[\phi_{d_\theta}^*]\right),\end{aligned}$$

and let

$$G_{\phi^*}(W,\alpha_0) \;\equiv\; \frac{dm(W,\alpha_0)}{d\theta'}-\frac{dm(W,\alpha_0)}{dh}[\phi^*].$$

We impose the following assumptions.

Assumption 5.5 (i) $E\left[\left\|\Sigma(W)^{-\frac{1}{2}}\left\{\frac{dm(W,\alpha_0)}{d\theta'}\right\}\right\|_E^2\right]$ *is finite.* (ii) $E\left[\left\|\Sigma(W)^{-\frac{1}{2}}G_{\phi^*}(W,\alpha_0)\right\|_E^2\right]$ *exists, is bounded, and is positive-definite.*

Assumption 5.5 is related to a local identification condition for θ. Define $s(\alpha)\equiv\lambda'\theta$ for $\lambda\in\mathcal{R}^{d_\theta}$ and $\lambda\neq 0$. Since $s(\alpha)\equiv\lambda'\theta$ is bounded if and only if $E[G_{\phi^*}(W,\alpha_0)'\Sigma(W)^{-1}G_{\phi^*}(W,\alpha_0)]$ is finite positive-definite, we have, by Assumption 5.5,

$$s(\alpha)-s(\alpha_0) \;\equiv\; \lambda'(\theta-\theta_0)=\langle b^*,\alpha-\alpha_0\rangle_\alpha$$

for all $\alpha\in\mathcal{A}$ where $b^*\equiv(b_\theta^*,b_h^*)\in\overline{\mathbf{B}}$, $b_\theta^*=\tilde{J}^{-1}\lambda$ with $\tilde{J}=E[G_{\phi^*}(W,\alpha_0)'\Sigma(W)^{-1}G_{\phi^*}(W,\alpha_0)]$, and $b_h^*=-\phi^*\times b_\theta^*$.

Let $\mathcal{H}_{os} \equiv \{h \in \mathcal{H} : \|h - h_0\|_{s,\alpha} = o(1), P(h) \leq c\}$ and $\mathcal{H}_{osn} \equiv \mathcal{H}_{os} \cap \mathcal{H}_n$. For any $h_1, h_2 \in \mathcal{H}_{os}$, we define

$$\|h_1 - h_2\|_\alpha^2 \equiv E\left[\left(\frac{dm(W,\alpha_0)}{dh}[h_1 - h_2]\right)' \Sigma(X)^{-1} \left(\frac{dm(W,\alpha_0)}{dh}[h_1 - h_2]\right)\right].$$

Assumption 5.6 (i) *$\mathcal{H} \subseteq \mathbf{H}, (\mathbf{H}, \|\cdot\|_{s,\alpha})$ is a Hilbert space with $\langle\cdot,\cdot\rangle_{s,\alpha}$ the inner product and $\{q_j\}_{j=1}^\infty$ a Riesz basis.* (ii) *$\mathcal{H}_n = clsp\{q_1, \ldots, q_n\}$.*

Assumption 5.7 *There are finite constants $c_1, c_2 > 0$ and a non-increasing positive sequence $\{b_j\}_{j=1}^\infty$ such that* (i) *$\|h\|_\alpha^2 \geq c_1 \sum_{j=1}^\infty b_j |\langle h, q_j\rangle_{s,\alpha}|^2$ for all $h \in \mathcal{H}_{osn}$ and* (ii) *$c_2 \sum_{j=1}^\infty b_j |\langle h_0 - \Pi_n h_0, q_j\rangle_{s,\alpha}|^2 \geq \|h_0 - \Pi_n h_0\|_\alpha^2$.*

Assumption 5.6 suggests that $\mathcal{H}_n = clsp\{q_1, \ldots, q_n\}$ is a natural sieve space for the estimation of h_0 where $clsp\,\{\cdot\}$ is the closure of the linear span under the metric $\|\cdot\|_{s,\alpha}$. Assumption 5.7 (i) links the weak metric to its corresponding strong metric and (ii) is the so-called stability condition (see Chen and Pouzo (2012) for further discussions on these conditions). The following theorem states a convergence rate result for the estimator in (14).

Theorem 5.1. *(i) Let all the Assumptions of Theorem 4.1 hold. Let Assumptions 5.1 - 5.5 hold, and $\sup_{h \in \mathcal{H}_{osn}} |\hat{P}_n(h) - P(h)| = o_p(1)$. Then $\|\hat{\alpha}_n - \alpha_0\|_\alpha = O_p(\max\{\delta_{m,n}, \|h_0 - \Pi_n h_0\|_\alpha, \sqrt{\lambda_n}\})$. (ii) Further, let Assumptions 5.6- 5.7 hold, $\|h_0 - \Pi_n h_0\|_\alpha = o(n^{-1/4})$, and $\max\{\delta_{m,n}, \sqrt{\lambda_n}\} = \delta_{m,n}$. Then $\|\hat{\alpha}_n - \alpha_0\|_\alpha = o_p(n^{-1/4})$.*

See Appendix G for the proof.

5.2 Asymptotic Normality

Based on the asymptotic results in the previous sections, we now establish the $\sqrt{n}$-normality of the PSMD estimator $\hat{\theta}_n$ by extending the results in Ai and Chen (2003). Under Assumption 5.5, for $\lambda \in \mathcal{R}^{d_\theta}$ and $\lambda \neq 0$, there exists a Riesz representer $b^* \equiv (b_\theta^*, b_h^*) \in \overline{\mathbf{B}}$ of $\lambda'(\theta - \theta_0) = \langle b^*, \alpha - \alpha_0\rangle_\alpha$ where $b_\theta^* = \tilde{J}^{-1}\lambda$ with $\tilde{J} = E[G_{\phi^*}(W,\alpha_0)'\Sigma(W)^{-1}G_{\phi^*}(W,\alpha_0)]$, and $b_h^* = -\phi^* \times b_\theta^*$. Define $\mathcal{N}_0 \equiv \{\alpha \in \mathcal{A}_{os} : \|\alpha - \alpha_0\|_\alpha = o(n^{-1/4}), \|\alpha - \alpha_0\|_{s,\alpha} = o(1)\}$ and $\mathcal{N}_{0n} \equiv \mathcal{N}_0 \cap \mathcal{A}_n$. We denote

$$\frac{d\rho(X,\alpha)}{d\alpha}[b] \equiv \left.\frac{d\rho(X,\alpha + \tau b)}{d\tau}\right|_{\tau=0} \quad \text{a.s. } X,$$

and

$$\frac{dm(W,\alpha)}{d\alpha}[b] \equiv \left.\frac{dm(W,\alpha + \tau b)}{d\tau}\right|_{\tau=0} \quad \text{a.s. } W,$$

for any $b \in \overline{\mathbf{B}}$. We make the following assumptions.

Assumption 5.8 (i) *There exists a measurable function $C_2(W)$ with $E[|C_2(W)|] < \infty$ and constants $\kappa \in (0,1], r \geq 1$ such that for all $\delta > 0$ and $\alpha \in \mathcal{N}_{0n}$*

$$\sup_{\|\alpha-\alpha_0\|_{s,\alpha} \leq \delta} \int \left| \frac{\rho(x,\alpha) - \rho(x,\alpha_0)}{\delta^\kappa} \right|^r dF_{Y|W=w}(y) \quad \leq \quad C_2(w)^r$$

and $\|\alpha - \alpha_0\|^\kappa_{s,\alpha} = o(n^{-1/2})$. (ii) *There exists a measurable function $C_3(X)$ with a constant $c > 0$ such that $\sup_{\alpha \in \mathcal{N}_0} |\rho(X,\alpha)| \leq C_3(X)$ and $E[C_3(X) \mid W] \leq c < \infty$.*

Assumption 5.9 (i) *$\theta_0 \in int(\Theta)$.* (ii) *$\Sigma_0(W) \equiv Var[\rho(X,\alpha_0) \mid W]$ is positive-definite for all $W \in \mathcal{W}$.* (iii) *There is a $b^*_n = (b^*_\theta, -\Pi_n\phi^* \times b^*_\theta) \in \mathcal{A}_n - \{\alpha_0\}$ such that $\|b^*_n - b^*\|_\alpha = o(n^{-1/4})$.*

Assumption 5.10 (i) *$\hat{\Sigma}(w) = \Sigma(w) + o_p(n^{-1/4})$ uniformly over $w \in \mathcal{W}$.* (ii) *There exists a positive sequence $\epsilon_n = o(n^{-1/2})$ such that $\lambda_n \sup_{\alpha \in \mathcal{N}_{0n}} |\hat{P}_n(h \pm \epsilon_n \phi^*_n b^*_\theta) - \hat{P}_n(h)| = o_p(n^{-1})$.*

Assumption 5.8 is satisfied by typical smooth classes of the residual function. Assumption 5.9 (iii) requires no asymptotic bias of b^*_n. Assumptions 5.10 (i)-(ii) quantify the estimation error of the weighting matrix and the approximation error of the sieve in the penalty function, respectively. Define $g(W,b^*) \equiv \left(\frac{dm(W,\alpha_0)}{d\alpha}[b^*]\right)' \Sigma(W)^{-1}$ and its projection onto the integral function using the estimated densities as $\tilde{g}(W,b^*) \equiv \int [\int g(W,b^*) \hat{f}_{Y_1|Y_2Z_1}(y_1 \mid y_2, z_1) dy_1] \hat{f}_{Y_2|VZ_1}(y_2 \mid v, z_1) dy_2$. Similarly define the projection of $m(W,\alpha)$ as $\tilde{m}(W,\alpha) \equiv \int [\int m(W,\alpha) \hat{f}_{Y_1|Y_2Z_1}(y_1 \mid y_2, z_1) dy_1] \hat{f}_{Y_2|VZ_1}(y_2 \mid v, z_1) dy_2$.

Assumption 5.11 (i) *$m(W,\alpha)$ is path-wise differentiable in $\alpha \in \mathcal{N}_{0n}$ and uniformly over $\alpha \in \mathcal{N}_{0n}$, $E\left[\left\| \frac{d\tilde{m}(W,\alpha)}{d\alpha}[b^*_n] - \frac{dm(W,\alpha)}{d\alpha}[b^*_n] \right\|^2_E\right] = o_p(n^{-1/2})$.* (ii) *$E\left[\|\tilde{g}(W,b^*) - g(W,b^*)\|^2_E\right] = o_p(n^{-1/2})$.*

Assumption 5.12 *$\left\{ \left(\frac{dm(W,\alpha_0)}{d\alpha}[b^*]\right)' \Sigma(W)^{-1} m(W,\alpha) : \alpha \in \mathcal{N}_{0n}, m \in \Lambda^{\gamma,\omega}_c(\mathcal{W}) \right\}$ is a Donsker class for some constant $c > 0$ with $\gamma > d_W/2$, where d_W is the dimension of W.*

Assumption 5.13 (i) *$m(W,\alpha)$ is twice path-wise differentiable in $\alpha \in \mathcal{N}_{0n}$, and there exists a measurable function $C_4(W)$ with $E[C_4(W)^2] < \infty$ such that $\frac{d^2m(W,\alpha)}{d\alpha d\alpha}[b^*_n, b^*_n]$ is bounded by $C_4(W)$ uniformly over $\alpha \in \mathcal{N}_{0n}$.* (ii) *$E\left[\sup_{\alpha \in \mathcal{N}_{0n}} \left\| \frac{dm(W,\alpha)}{d\alpha}[b^*_n] - \frac{dm(W,\alpha_0)}{d\alpha}[b^*_n] \right\|^2_E\right] = o(n^{-1/2})$.* (iii) *Uniformly over $\alpha \in \mathcal{N}_{0n}, \bar{\alpha} \in \mathcal{N}_0$,*

$$E\left[\left(\frac{dm(W,\alpha_0)}{d\alpha}[b^*]\right)' \Sigma(W)^{-1} \left(\frac{dm(W,\bar{\alpha})}{d\alpha}[\alpha - \alpha_0] - \frac{dm(W,\alpha_0)}{d\alpha}[\alpha - \alpha_0]\right)\right] \quad = \quad o(n^{-1/2}).$$

Assumptions 5.11 and 5.12 are required to control the asymptotic bias when the parameter α enters the residual $\rho(\cdot)$ nonlinearly. Assumption 5.13 is required to control the higher order terms in a mean value expansion to derive the influence function representation result below.

Under these assumptions, we can obtain the influence function representation of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ as

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_n - \theta_0) &= \sqrt{n}\langle b^*, \hat{\alpha}_n - \alpha_0\rangle_\alpha \qquad (16) \\
&= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{dm(W_i,\alpha_0)}{d\alpha}[b^*]\right\}'\Sigma(W_i)^{-1}\rho(X_i,\alpha_0) + o_p(1) \\
&= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\tilde{J}^{-1}G_{\phi^*}(W_i,\alpha_0)\right\}'\Sigma(W_i)^{-1}\rho(X_i,\alpha_0) + o_p(1)
\end{aligned}$$

where $\tilde{J} = E[G_{\phi^*}(W,\alpha_0)'\Sigma(W)^{-1}G_{\phi^*}(W,\alpha_0)]$, which yields the following $\sqrt{n}$-asymptotic normality of $\hat{\theta}_n$.

Theorem 5.2. *Let $\hat{\theta}_n$ be the PSMD estimator of the finite dimensional parameters with $\lambda_n \geq 0, \lambda_n = o(1)$. Suppose that all the Assumptions of Theorem **5.1** and Assumptions 5.8 - 5.13 hold. Then $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, J^{-1})$, where*

$$\begin{aligned}
J &= E[G_{\phi^*}(W,\alpha_0)'\Sigma(W)^{-1}G_{\phi^*}(W,\alpha_0)] \\
&\quad \times(E[G_{\phi^*}(W,\alpha_0)'\Sigma(W)^{-1}\Sigma_0(W)\Sigma(W)^{-1}G_{\phi^*}(W,\alpha_0)])^{-1} \\
&\quad \times E[G_{\phi^*}(W,\alpha_0)'\Sigma(W)^{-1}G_{\phi^*}(W,\alpha_0)].
\end{aligned}$$

See Appendix H for the proof.

Note that by taking the weighting matrix $\hat{\Sigma}(W)$ in (14) as a consistent estimator of $\Sigma_0(W)$, the asymptotic variance can be reduced to $J^{-1} = \{E[G_{\phi^*}(W,\alpha_0)'\Sigma_0(W)^{-1}G_{\phi^*}(W,\alpha_0)]\}^{-1}$.

5.3 Correcting Asymptotic Variance for Generated Control Variable

Next, by building on the approach from Hahn and Ridder (2013), we obtain the asymptotic variance of the estimator $\hat{\theta}_n$ that uses the generated control variable as in the triangular model (4) for which the first stage equation yields the control variable (e.g.) $V = F_{Y_2|Z}$ (Imbens and Newey 2009) or $V = Y_2 - E[Y_2|Z]$ (Newey, Powell, and Vella 1999). For generic notation, below we write $V = \varphi(Y_2, Z)$ and let $V_* = \varphi_*(Y_2, Z)$ denote the true value. To use the framework of Hahn and Ridder (2013) and obtain the potential influence of the first stage to the asymptotic expansion of $\hat{\theta}_n$, define

$$\rho(X, \alpha_0(V_1; V_2)) \equiv \rho(Y, Z_1, \varphi(Y_2, Z) = V_1, \alpha_0(\varphi(Y_2, Z) = V_1; V_2))$$

where $V_2 \equiv \varphi(Y_2, Z)$. In this definition the two roles of V are made explicit. First, it enters in the variables at which both $\rho(\cdot, V_1, \cdot)$ and $\alpha_0(V_1; \cdot)$ are evaluated. Second, it determines the functional form of the parameter α_0, $\alpha_0(\cdot; V_2)$. Note that $V_1 = V_2 = V$, so the notation V_1, V_2 is just an expositional device to distinguish two roles of V. In the view of Hahn and Ridder (2013), the influence function in (16) already accounts for the estimation of α in the second step as a pre-step to obtain the plug-in estimator $\hat{\theta}_n$ in the final step. Therefore, we have only to account for the contribution of the sampling variation in $\hat{V}$ while taking the function $\alpha_0(V_1; V_2)$ is known.

To account for the first stage estimation of V, now we can make the adjustment to the influence function following standard arguments as

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n - \theta_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{H}_i\left\{\rho(X_i,\alpha_0)\right\} \\ &\quad +\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{H}_i\left\{\rho(X_i,\alpha_0(\hat{V}_{1i};\hat{V}_{2i})) - \rho(X_i,\alpha_0(V_{1i};V_{2i}))\right\} \\ &\quad +\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{H}_i\left\{\rho(Y_i,Z_{1i},\hat{V}_i,\alpha_0) - \rho(X_i,\alpha_0)\right\} + o_p(1) \end{aligned} \tag{17}$$

where $\mathbf{H}_i \equiv -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{dm(W_i,\alpha_0)}{d\alpha}[b^*]\right\}'\Sigma(W_i)^{-1}$ and $\hat{V}_1 = \hat{V}_2 = \hat{V}$.

Define

$$\kappa(z_1,v) = E\left[\left.\frac{\partial\rho(X,\alpha_0)}{\partial h'}\right| z_1, \varphi_*(y_2,z) = v\right]$$

and then, following similar arguments to Hahn and Ridder (2013) (Section 2.3, Theorem 5 and Remark 4), we obtain the approximation of the second term in the right-hand side of (17) as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{H}_i\delta(Y_{2i},Z_i)\tilde{V}_i$$

with $\tilde{V}_i = E\left[1(Y_{2i} < Y_{2j}) - F_{Y_2|Z}(Y_{2j}|Z_i)|Y_{2i},Z_i\right]$ (nonseparable first stage) or $\tilde{V}_i = V_i = Y_{2i} - E[Y_{2i}|Z_i]$ (separable first stage) and

$$\begin{aligned} \delta(y_2,z) &= E\left[\left.\left(\frac{\partial\rho(X,\alpha_0)}{\partial h'} - \kappa(Z_1,\varphi_*(Y_2,Z))\right)\frac{\partial h_0(r(Z,\varphi_*(Y_2,Z)),Z_1)}{\partial V}\right| y_2,z\right] \\ &\quad +E\left[\left.\frac{\partial\kappa(Z_1,\varphi_*(Y_2,Z))}{\partial V}\kappa(Z_1,\varphi_*(Y_2,Z))'\Sigma(Z_1,\varphi_*(Y_2,Z))^{-1}E\left[\rho(X,\alpha_0)|Y_2,Z\right]\right| y_2,z\right]. \end{aligned} \tag{18}$$

Note that when the modeling assumptions of the triangular model (4) hold, V is known given (Y_2,Z), so (Z_1,V) becomes a subset of (Y_2,Z). Therefore, in this case, the first term in the right-hand side of (18) can be dropped using the law of iterated expectation by the definition of $\kappa(z_1,v)$. The second term in the right-hand side of (18) can be also dropped if the moment condition $E\left[\rho(X,\alpha_0)|Y_2,Z\right] = 0$ holds, which is stronger than the original moment condition $E\left[\rho(X,\alpha_0)|Z_1,V\right] = 0$.

Finally, the third term in the influence function can be approximated using a standard approach (e.g. Newey 1994) as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{H}_iE\left[\left.\frac{\partial\rho(X_i,\alpha_0)}{\partial V}\right| Y_{2i},Z_i\right]\tilde{V}_i.$$

Combining these results we obtain the asymptotic variance of $\hat{\theta}_n$ in the triangular model as

$$AVar[\sqrt{n}(\hat{\theta}_n - \theta_0)] = Var\left[\mathbf{H}_i\left\{\rho(X_i,\alpha_0) + \delta(Y_{2i},Z_i)\tilde{V}_i + E\left[\left.\frac{\partial\rho(X_i,\alpha_0)}{\partial V}\right| Y_{2i},Z_i\right]\tilde{V}_i\right\}\right].$$

We now summarize the result.

Corollary 5.3. *Suppose that all the Assumptions of* ***Theorem 5.1*** *and Assumptions 5.8 - 5.13 hold. Then for the triangular model (4) we obtain the asymptotic variance of the estimator $\hat{\theta}_n$ as*

$$AVar[\sqrt{n}(\hat{\theta}_n - \theta_0)] = Var\left[\mathbf{H}_i\left\{\rho(X_i,\alpha_0) + \delta(Y_{2i},Z_i)\tilde{V}_i + E\left[\frac{\partial\rho(X_i,\alpha_0)}{\partial V}\middle| Y_{2i},Z_i\right]\tilde{V}_i\right\}\right].$$

6 Simulations

This section conducts Monte Carlo simulations to assess the finite sample performance of the proposed estimator in a few different settings. First we study a partially linear model and investigate the performance of the proposed estimator of a finite dimensional parameter associated with an endogenous and mismeasured regressor. Next, we consider an additively-separable nonparametric regression model and investigate the performance of the estimator for an infinite dimensional parameter.

6.1 Partially Linear Model

We consider a data generating process from the following partially linear model:

$$\begin{aligned} Y_1 &= \pi(Z_1) + \theta Y_2 + \varepsilon, \\ Y_2 &= \phi_1 Z_1 + \phi_2 V + \nu, \\ \varepsilon &= \delta V + \varpi, \end{aligned}$$

where Y_1 is the dependent variable, Z_1 is an exogenous covariate drawn from $N(0, 0.5^2)$, and V is a control variable drawn from $N(0, 0.5^2)$, and where ν and ϖ are mutually independent innovations drawn from $N(0, 0.25^2)$ and $N(0, 0.5^2)$, respectively. The nonparametric function is specified as $\pi(\cdot) = \exp(\cdot)$. Y_2 is an endogenous and unobserved covariate and researchers observe only its mismeasured counterpart $Y_2^* = Y_2 + \sigma_e \exp(-Y_2) \cdot e$ where e is an measurement error and σ_e is its standard deviation. We consider three different structures of measurement error as follows. Design A is a non-additive error with zero mode such that $e = \ln(-\ln(1 - \mathrm{U}))$ where U is a uniformly distributed random variable over $[0, 1]$ support, Design B is a heteroskedastic measurement error with zero mean as $e = N(0, 1)$, and lastly Design C is a non-additive error with zero median such that $e = \ln(\omega + \sqrt{\omega^2 + 2})$ where $\omega = -0.5 + \tan(\pi\mathrm{U} - 0.5)/\exp(-Y_2)$. Coefficients in this model are set to be $\theta = 1.5$, $\phi_1 = 1$, $\phi_2 = 1.5$, and $\delta = 0.5$. For each design we also vary the size of standard deviation σ_e by $0.5, 1$, and 1.5.

We compare the finite-sample performances of the proposed estimator with two other sieve IV estimators; infeasible estimator using the true Y_2 as a benchmark and inconsistent estimator using mismeasured Y_2^*. These two sieve IV estimators control for endogeneity of Y_2 using a set of instruments which is a tensor product polynomial sieve of order 3: $P_i \equiv (1, Z_{1i}, V_i, Z_{1i}^2, Z_{1i}V_i, V_i^2, \ldots, V_i^3)'$.

For instance, in order to construct the infeasible estimator, the term $R \equiv Y_2 - E[Y_2 \mid W]$ is estimated as the regression residual of Y_2 on P. Then the estimator of θ is obtained by taking the weighted regression of Y_1 on Y_2 treating R as the weight:

$$\hat{\theta}_{infeasible} = (\sum_{i=1}^{n} \Psi_i P_i' (\sum_{i=1}^{n} P_i P_i')^{-1} \sum_{i=1}^{n} P_i \Psi_i')^{-1} \sum_{i=1}^{n} \Psi_i P_i' (\sum_{i=1}^{n} P_i P_i')^{-1} \sum_{i=1}^{n} P_i \widetilde{Y}_{1i}$$

where

$$\begin{aligned} \widetilde{Y}_{1i} &\equiv \hat{R}_i \times Y_{1i}, \\ \Psi_i &\equiv \hat{R}_i \times Y_{2i}, \\ \hat{R}_i &\equiv Y_{2i} - \hat{E}[Y_2 \mid W = W_i]. \end{aligned}$$

The inconsistent estimator is constructed in a similar fashion by replacing the true Y_2 with mismeasured Y_2^*.

To implement the proposed estimator, approximating sieves for functions of three variables using tensor product bases of univariate trigonometric series are employed to approximate the densities $f_{Y_1|Y_2 Z_1}$ and $f_{Y_2^*|Y_2 Z_1}$. For instance, the sieve approximations are given by

$$\begin{aligned} f_{Y_1|Y_2 Z_1}(y_1 \mid y_2, z_1) &\approx \sum_{j_1=0}^{j_{1n}} \sum_{j_2=0}^{j_{2n}} \sum_{j_3=0}^{j_{3n}} \gamma_{j_1 j_2 j_3} u_{j_1}(y_1 - y_2) u_{j_2}(y_2) u_{j_3}(z_1), \\ f_{Y_2^*|Y_2 Z_1}(y_2^* \mid y_2, z_1) &\approx \sum_{j_1=0}^{j_{1n}} \sum_{j_2=0}^{j_{2n}} \sum_{j_3=0}^{j_{3n}} \vartheta_{j_1 j_2 j_3} u_{j_1}(y_2^* - y_2) u_{j_2}(y_2) u_{j_3}(z_1) \end{aligned}$$

where $u_{j_1}(\cdot)$ is a sine or cosine function, and where $u_{j_2}(\cdot)$ and $u_{j_3}(\cdot)$ are cosine functions. By utilizing desirable properties of the trigonometric series, the identification restriction on $f_{Y_2^*|Y_2 Z_1}$ in Assumption 2.7 can be easily imposed (see e.g. Hu and Schennach 2008). In addition, it can be guaranteed that integral of each density over its support is indeed equal to one. The density $f_{Y_2|V Z_1}$ is specified as a normal density to ease high dimensionality of the nonparametric specification. In the first stage, we estimate $f_{Y_1|Y_2 Z_1}$ and $f_{Y_2|V Z_1}$ using the sieve maximum likelihood estimation as in the equation (12). In the second stage, the estimate of the weight $R \equiv Y_2 - E[Y_2 \mid W]$ is constructed by

$$\tilde{R}(Y_2, V, Z_1) = Y_2 - \int_{\mathcal{Y}_2} y_2 \hat{f}_{Y_2|V Z_1}(y_2 \mid V, Z_1) dy_2.$$

Then, the estimator of θ is obtained by taking the weighted least squares regression of Y_1 on Y_2 through the equation (14) [1] such that

[1] We take $\widehat{\Sigma} = \Sigma = I$ (identity) and $\lambda_n = 0$ for our experiments.

$$\hat{\theta}_{proposed} = (\sum_{i=1}^{n} \widetilde{Y}_{2i}\widetilde{Y}_{2i}')^{-1}\sum_{i=1}^{n} \widetilde{Y}_{2i}\widetilde{Y}_{1i}$$

where

$$\begin{aligned}
\widetilde{Y}_{1i} &\equiv \int_{\mathcal{Y}_2}\int_{\mathcal{Y}_1} \tilde{R}(y_2, V_i, Z_{1i}) y_1 \hat{f}_{Y_1|Y_2 Z_1}(y_1 \mid y_2, Z_{1i}) \hat{f}_{Y_2|VZ_1}(y_2 \mid V_i, Z_{1i}) dy_1 dy_2, \\
\widetilde{Y}_{2i} &\equiv \int_{\mathcal{Y}_2} \tilde{R}(y_2, V_i, Z_{1i}) y_2 \hat{f}_{Y_2|VZ_1}(y_2 \mid V_i, Z_{1i}) dy_2.
\end{aligned}$$

In both stages, we adopt a Gauss-Hermite quadrature method for numerical integrals because the supports of Y_1 and Y_2 are potentially unbounded. As discussed before, the proposed estimator does not require numerical optimization in the second stage since it has a closed-form solution. In addition, it is not necessary to estimate the function $\pi(\cdot)$ associated with exogenous covariates Z_1 when researchers are primarily interested in estimating the effect of endogenous regressor Y_2 on Y_1.

We investigate the finite-sample performances of the three estimators described above by calculating the squared bias (SB), variance (VAR), and mean squared error (MSE). We consider several levels of standard deviation of the measurement error, $\sigma_e \in \{0.5, 1.0, 1.5\}$. By doing so, we can investigate how the degree of severeness of the measurement error affects behaviors of the estimators. The number of observations is 1000 and the number of repetitions for each experiment is 200.

Table 1 reports the estimation results. It mainly shows the proposed estimator outperforms the inconsistent estimator. For example, when the measurement error is non-additive with zero mode and $\sigma_e = 0.5$, SB of the proposed estimator is 0.0188 which is close to that of the infeasible estimator, 0.0034. However, SB from the inconsistent estimator is significantly larger as it becomes 2.1440. For this MC study, because the proposed estimator is a semiparametric two-step estimator that requires more flexible approximation in the first stage while the other two estimators are based on least squares, the proposed estimator tends to produce larger variances.[2] Nevertheless, MSE of the proposed estimator, 0.4813, is much smaller than the one from the inconsistent estimator, which is 2.1457.

The finite sample behaviors of the three estimators are similar in other structures of the measurement errors such as heteroskedastic measurement error with zero mean and non-additive error with zero median. The results show similar patterns across larger standard deviations of the measurement error.

6.2 Additively-separable Nonparametric Model

We consider the following additively-separable nonparametric model for our experiments:

[2] For our proposed estimator we also experimented with different numbers of sieve approximation terms in the first stage. We find the usual trade-off between bias and variance. Using more sieve terms generally reduces biases while increasing variances.

$$\begin{aligned} Y_1 &= h(Y_2) + \pi(Z_1) + \varepsilon, \\ Y_2 &= \phi_1 Z_1 + \phi_2 V + \nu, \\ \varepsilon &= \delta V + \varpi, \end{aligned}$$

where $h(\cdot) = -\frac{\exp(\cdot)}{1+\exp(\cdot)}$ and $\pi(\cdot) = 2\sin(\cdot)$. All other random variables and parameters are the same as those given in Section 6.1. The function h is the primary parameter of interest. As a sieve basis for the function h, we use a power series of fourth order multiplied by the standard normal CDF. We report performances of the three estimators over different structures of measurement error as described in Section 6.1: Design A, Design B, and Design C. We also vary standard deviation of the measurement error σ_e by $0.5, 1$, and 1.5. The finite-sample performances are also evaluated over several numbers of observations, $n \in \{500, 1000\}$. The number of repetitions for each experiment is 200.

Table 2 and Table 3 report the integrated squared bias (ISB), integrated variance (IVAR), and integrated mean squared error (IMSE) of the estimate of $h(\cdot)$, which are computed using numerical integral over a grid ranging from -2 to 2.

From the estimation results we find that in all designs the proposed estimator outperforms the inconsistent estimator. For example, when the measurement error is non-additive with zero mode (Design A) and $\sigma_e = 0.5$ and when the sample size is $n = 1,000$, ISB from the proposed estimator is 0.0299 which is close to that of the infeasible estimator, 0.0102 while ISB of the inconsistent estimator is significantly larger as 2.93. In terms of IMSE our proposed estimator clearly dominates the inconsistent estimator. For the same setting above, IMSE of the proposed estimator is 0.1202 while that of the inconsistent estimator is 47.42. From other designs we observe that the finite sample behaviors of the estimators are all similar and the results are stable across larger standard deviation of the measurement error. Finally, the performance of the proposed estimator improves as the sample size increases from 500 to $1,000$.

We also report graphs of three estimated functions (infeasible, proposed, and inconsistent) along with the true function in Figure G1-G12 for all three designs of measurement errors.

7 Concluding Remarks

We study identification and estimation of regression functions for a class of semiparametric models for which the endogenous regressors are measured with errors. For these models we utilize the existence of control variables - that ensure the conditional mean independence of endogenous regressors and unobservable causes given the control variables. Our framework extends to the triangular simultaneous equations models for which the control variable can be recovered from the first stage reduced form equation. Given our identification results we propose a sieve method to estimate the parameters. Finally we derive the asymptotic properties of the proposed estimator. Monte Carlo simulations illustrate that our proposed estimator performs well in the finite samples.

Appendix

A Examples of Lemma 2.3

In order to understand Lemma **2.3**, consider the following examples. Two of them satisfy the conditional independence condition in Lemma **2.3** while the other two do not satisfy the condition.

Let $\varepsilon, Y_2, V, U_a, U_b, U_\varepsilon, U_{Y_2}, U_V$ be random variables. Assume $U_\varepsilon, U_{Y_2}, U_V$ are mutually independent. For measurable functions p, q, r, assume that ε, Y_2, V are determined by the structural equations in each example.

Example 1)

$$\begin{aligned} \varepsilon &= p(U_a, U_b, U_\varepsilon) \\ Y_2 &= q(U_a, U_b, U_{Y_2}) \\ V &= r(U_a, U_b, U_V). \end{aligned}$$

The information in common between Y_2 and V is $T = \{U_a, U_b\}$. Since both $Y_2 \perp \varepsilon \mid V$ and $V \perp \varepsilon \mid Y_2$ are satisfied, $\varepsilon \perp (Y_2, V) \mid T$ is satisfied as in Lemma **2.3**.

Example 2)

$$\begin{aligned} \varepsilon &= p(U_a, U_\varepsilon) \\ Y_2 &= q(U_a, U_b, U_{Y_2}) \\ V &= r(U_a, U_b, U_V). \end{aligned}$$

The information in common between Y_2 and V is $T = \{U_a, U_b\}$. Since both $Y_2 \perp \varepsilon \mid V$ and $V \perp \varepsilon \mid Y_2$ are satisfied, $\varepsilon \perp (Y_2, V) \mid T$ is satisfied as in Lemma **2.3**.

Example 3)

$$\begin{aligned} \varepsilon &= p(U_a, U_b, U_\varepsilon) \\ Y_2 &= q(U_a, U_b, U_{Y_2}) \\ V &= r(U_a, U_V). \end{aligned}$$

The information in common between Y_2 and V is $T = \{U_a\}$. Since $Y_2 \perp \varepsilon \mid V$ is violated, $\varepsilon \perp (Y_2, V) \mid T$ is not satisfied.

Example 4)

$$\begin{aligned} \varepsilon &= p(U_a, U_b, U_\varepsilon) \\ Y_2 &= q(U_a, U_{Y_2}) \\ V &= r(U_a, U_b, U_V). \end{aligned}$$

The information in common between Y_2 and V is $T = \{U_a\}$. Since $V \perp \varepsilon \mid Y_2$ is violated,

$\varepsilon \perp (Y_2, V) \mid T$ is not satisfied.

These examples clearly illustrate that common information between ε and Y_2 and between ε and V must be included in T in order to satisfy the conditional independence condition $\varepsilon \perp (Y_2, V) \mid T$.

B Proof of Theorem 2.1

By the definition of conditional expectation, we obtain

$$\begin{aligned} m(w, \theta, h) &= \int_{\mathcal{Y}} \rho(x, \theta, h) f_{Y|W}(y \mid w) dy \\ &= \int_{\mathcal{Y}} (y_2 - E[Y_2 \mid W = w])(y_1 - G(y_2; \theta, h)) f_{Y|W}(y \mid w) dy. \end{aligned}$$

From Assumption 2.1, $\alpha_0 \equiv (\theta_0, h_0) \in \Theta \times \mathcal{H}$ is the unique solution for the equation $m(w, \theta, h) = 0$. Thus the identification of $f_{Y|W}(y \mid w)$ and the identification of $\rho(x, \theta, h)$ given α are sufficient for the identification of the parameter α_0 through the moment equation $m(w, \theta, h) = 0$. First, for the identification of $\rho(x, \theta, h) \equiv (y_2 - E[Y_2 \mid W = w])(y_1 - G(y_2; \theta, h))$ we need to recover $E[Y_2 \mid W]$ from the observables. We note that this conditional mean function inside the residual can be written as

$$E[Y_2 \mid W = w] = \int_{\mathcal{Y}_2} y_2 f_{Y_2|W}(y_2 \mid w) dy_2,$$

so that $f_{Y_2|W}(y_2 \mid w)$ is sufficient for the identification of $E[Y_2 \mid W]$. Second, for the identification of $f_{Y|W}(y \mid w)$, we use the fact that $f_{Y|W}(y \mid w) = f_{Y_1|Y_2 W}(y_1 \mid y_2, w) f_{Y_2|W}(y_2 \mid w) = f_{Y_1|Y_2 Z_1}(y_1 \mid y_2, z_1) f_{Y_2|W}(y_2 \mid w)$ where the first equality holds by Bayes rule and the second equality holds by Assumption 2.3. Thus the identification of $m(w, \theta, h)$ is obtained by identifying the two density functions $f_{Y_1|Y_2 Z_1}(y_1 \mid y_2, z_1)$ and $f_{Y_2|W}(y_2 \mid w)$. For the identification of the density functions, we use a similar argument to Hu and Schennach (2008). By Assumptions 2.2-2.4, we have the following integral equation

$$\begin{aligned} f_{Y^*|W}(y^* \mid w) &= \int_{\mathcal{Y}_2} f_{Y^* Y_2|W}(y^*, y_2 \mid w) dy_2 \\ &= \int_{\mathcal{Y}_2} f_{Y_1|Y_2^*, Y_2, W}(y_1 \mid y_2^*, y_2, w) f_{Y_2^* Y_2|W}(y_2^*, y_2 \mid w) dy_2 \\ &= \int_{\mathcal{Y}_2} f_{Y_1|Y_2^*, Y_2, W}(y_1 \mid y_2^*, y_2, w) f_{Y_2^*|Y_2 W}(y_2^* \mid y_2, w) f_{Y_2|W}(y_2 \mid w) dy_2 \\ &= \int_{\mathcal{Y}_2} f_{Y_1|Y_2 Z_1}(y_1 \mid y_2, z_1) f_{Y_2^*|Y_2 Z_1}(y_2^* \mid y_2, z_1) f_{Y_2|W}(y_2 \mid w) dy_2, \end{aligned}$$

where the last equality holds by Assumptions 2.2-2.4. Recall that R_1, R_2, and R_3 denote random variables with supports $\mathcal{R}_1, \mathcal{R}_2$, and $\mathcal{R}_3$, respectively, and $L_{R_1|R_2 r_3}$ denote an integral operator mapping $g \in \mathcal{G}(\mathcal{R}_2)$ to $L_{R_1|R_2 r_3} g \in \mathcal{G}(\mathcal{R}_1)$ for a given r_3 defined by $[L_{R_1|R_2 r_3} g](r_1) \equiv \int_{\mathcal{R}_2} f_{R_1|R_2 R_3}(r_1 \mid r_2, r_3) g(r_2) dr_2$, where $\mathcal{G}(\mathcal{R}_j)$ is the corresponding function space with domain $\mathcal{R}_j$ with $j = 1, 2$. Similarly, let $\triangle_{r_1|R_2 r_3}$ denote a diagonal operator mapping $g \in \mathcal{G}(\mathcal{R}_2)$ to $\triangle_{r_1|R_2 r_3} g \in \mathcal{G}(\mathcal{R}_2)$ for a given (r_1, r_3) such as $\triangle_{r_1|R_2 r_3} g \equiv f_{R_1|R_2R_3}(r_1 \mid r_2, r_3)g(r_2)$. We now show that the densities $(f_{Y_1|Y_2Z_1}, f_{Y_2^*|Y_2Z_1}, f_{Y_2|W})$ are uniquely identified from the joint density $f_{Y^*|W}(y^* \mid w)$ where we observe Y^* instead of Y, by Assumptions 2.5-2.7. Using operator notation, we get

$$\begin{aligned}
[L_{Y^*|Vz_1}(y^* \mid V, z_1)g](y_2^*) &= \int_{\mathcal{V}} f_{Y^*|VZ_1}(y^* \mid v, z_1)g(v)dv \\
&= \int_{\mathcal{V}}\int_{\mathcal{Y}_2} f_{Y_1|Y_2Z_1}(y_1 \mid y_2, z_1) f_{Y_2^*|Y_2Z_1}(y_2^* \mid y_2, z_1) f_{Y_2|W}(y_2 \mid w)dy_2 g(v)dv \\
&= \int_{\mathcal{Y}_2} f_{Y_1|Y_2Z_1}(y_1 \mid y_2, z_1) f_{Y_2^*|Y_2Z_1}(y_2^* \mid y_2, z_1) \int_{\mathcal{V}} f_{Y_2|W}(y_2 \mid w)g(v)dvdy_2 \\
&= [L_{Y_2^*|Y_2z_1} \triangle_{y_1|Y_2z_1} L_{Y_2|Vz_1} g](y_2^*).
\end{aligned}$$

By substituting $L_{Y_2|Vz_1}$ into the equation, which is obtained from an integration of the above equation over all y_1, we have

$$L_{Y^*|Vz_1} L^{-1}_{Y_2^*|Vz_1} = L_{Y_2^*|Y_2z_1} \triangle_{y_1|Y_2z_1} L^{-1}_{Y_2^*|Y_2z_1}$$

where the inverses of $L_{Y_2^*|Vz_1}$ and $L_{Y_2^*|Y_2z_1}$ are guaranteed by Assumption 2.5. Then by Assumptions 2.6-2.7 and a similar argument to the proof of Theorem 1 of Hu and Schennach (2008), the spectral decomposition is unique.

This result is an extension of Hu and Schennach (2008) to the identification of unobservable densities in the model with additional observable exogenous regressors, Z_1, where the unobserved regressors Y_2 are endogenous. Given the identification of the required conditional density functions, we obtain the conditional moment function $m(w, \theta, h)$. Then because the conditional moment restrictions have the unique solution as (θ_0, h_0) (Assumption 2.1), the true parameter is identified. This completes the proof.

C Proof of Theorem 2.2

Let $U_j(\cdot)$ be a generic function, for $j = 1, 2$.

(a) Because Assumption 2.1S implies $Y_2 \perp \varepsilon \mid W$, the conditional mean independence in Assumption 2.1 (i) is trivially satisfied.

(b) From Assumption 2.2S, we have $\varepsilon \perp e \mid (Y_2, W)$. By Lemma 4.1 of Dawid (1979a), $\varepsilon \perp e \mid (Y_2, W)$ is equivalent to $(Y_2, Z_1, \varepsilon) \perp (Y_2, e) \mid (Y_2, W)$. Then by Lemma 4.2 (i) of Dawid (1979a), it follows that $U_1(Y_2, Z_1, \varepsilon)) \perp U_2(Y_2, e) \mid (Y_2, W)$. Then, Assumption 2.2 immediately follows.

(c) From Assumption 2.3S, we have $\varepsilon \perp \eta \mid (Y_2, Z_1)$. By Lemma 4.1 of Dawid (1979a), $\varepsilon \perp \eta \mid (Y_2, Z_1)$ implies $(Y_2, Z_1, \varepsilon) \perp \eta \mid (Y_2, Z_1)$. Then by Lemma 4.2 (i) of Dawid (1979a), it follows that $U_1(Y_2, Z_1, \varepsilon)) \perp U_2(\eta) \mid (Y_2, Z_1)$. Then, Assumption 2.3 immediately follows.

(d) From Assumption 2.4S, we have $e \perp \eta \mid (Y_2, Z_1)$. By Lemma 4.1 of Dawid (1979a), $e \perp \eta \mid (Y_2, Z_1)$ is equivalent to $(Y_2, e) \perp \eta \mid (Y_2, Z_1)$. Then by Lemma 4.2 (i) of Dawid (1979a), it follows that $U_1(Y_2, e) \perp U_2(\eta) \mid (Y_2, Z_1)$. Then, Assumption 2.4 immediately follows.

D Proof of Lemma 2.4

(i) By a similar argument to Matzkin (2003), we have

$$F_V(v) \;=\; F_V(r^{-1}(z,y_2)) = P(V \leq r^{-1}(Z,y_2) \mid Z=z) = P(r(Z,V) \leq y_2 \mid Z=z) = F_{Y_2|Z}(y_2 \mid z)$$

from the monotonicity of $r(Z,V)$ in V and the independence of Z and V. Then by normalizing V such that it follows a uniform distribution over $[0,1]$ as $V = F_V(V)$, we obtain the result.

(ii) Denoting the support of Z by $\mathcal{Z}$, we note that

$$E[\exp(\mathrm{i}\zeta Y_2) \mid Z=z] \;=\; \int f_{Y_2|Z}(y_2 \mid z)\exp(\mathrm{i}\zeta y_2)\mathrm{d}y_2$$

is the Fourier transform of $f_{Y_2|Z}(y_2 \mid z)$ and that

$$\frac{1}{2\pi}\int E[\exp(\mathrm{i}\zeta Y_2) \mid Z=z]\exp(-\mathrm{i}\zeta y_2)\mathrm{d}\zeta$$

is the inverse Fourier transform of $E[\exp(\mathrm{i}\zeta Y_2) \mid Z=z]$ for $(y_2,z) \in \mathcal{Y}_2 \times \mathcal{Z}$. As a result, we get

$$f_{Y_2|Z}(y_2 \mid z) \;=\; \frac{1}{2\pi}\int E[\exp(\mathrm{i}\zeta Y_2) \mid Z=z]\exp(-\mathrm{i}\zeta y_2)\mathrm{d}\zeta.$$

Then the inversion theorem (e.g., Gurland 1948) provides the conditional CDF of Y_2 given $Z=z$

$$\begin{aligned}&F_{Y_2|Z}(y_2 \mid z) \qquad (19)\\ =\;& \frac{1}{2}+\frac{1}{2\pi}\int_0^\infty \frac{E[\exp(-\mathrm{i}\zeta Y_2) \mid Z=z]\exp(\mathrm{i}\zeta y_2) - E[\exp(\mathrm{i}\zeta Y_2) \mid Z=z]\exp(-\mathrm{i}\zeta y_2)}{\mathrm{i}\zeta}\mathrm{d}\zeta.\end{aligned}$$

We now show the identification of $E[\exp(\mathrm{i}\zeta Y_2) \mid Z=z]$. From (19) it is clear that identification of $E[\exp(\mathrm{i}\zeta Y_2) \mid Z=z]$ suffices to recover the CDF, $F_{Y_2|Z}(y_2 \mid z)$.

First observe that

$$\begin{aligned}
\int_0^\zeta \frac{\mathrm{i}E[Y_{2a}^* \exp(\mathrm{i}\xi Y_{2b}^*)]}{E[\exp(\mathrm{i}\xi Y_{2b}^*)]}\mathrm{d}\xi &= \int_0^\zeta \frac{\mathrm{i}E[(Y_2+e_a)\exp(\mathrm{i}\xi Y_{2b}^*)]}{E[\exp(\mathrm{i}\xi(Y_2+e_b))]}\mathrm{d}\xi \\
&= \int_0^\zeta \frac{\mathrm{i}E[Y_2\exp(\mathrm{i}\xi(Y_2+e_b))]+\mathrm{i}E[e_a\exp(\mathrm{i}\xi Y_{2b}^*)]}{E[\exp(\mathrm{i}\xi(Y_2+e_b))]}\mathrm{d}\xi \\
&= \int_0^\zeta \frac{\mathrm{i}E[Y_2\exp(\mathrm{i}\xi(Y_2+e_b))]+\mathrm{i}E[E(e_a\exp(\mathrm{i}\xi Y_{2b}^*)\mid Y_{2b}^*)]}{E[\exp(\mathrm{i}\xi(Y_2+e_b))]}\mathrm{d}\xi \\
&= \int_0^\zeta \frac{\mathrm{i}E[Y_2\exp(\mathrm{i}\xi(Y_2+e_b))]+\mathrm{i}E[E(e_a\mid Y_{2b}^*)\exp(\mathrm{i}\xi Y_{2b}^*)]}{E[\exp(\mathrm{i}\xi(Y_2+e_b))]}\mathrm{d}\xi \\
&= \int_0^\zeta \frac{\mathrm{i}E[Y_2\exp(\mathrm{i}\xi(Y_2+e_b))]}{E[\exp(\mathrm{i}\xi(Y_2+e_b))]}\mathrm{d}\xi \\
&= \int_0^\zeta \frac{\mathrm{i}E[Y_2\exp(\mathrm{i}\xi Y_2)]E[\exp(\mathrm{i}\xi e_b)]}{E[\exp(\mathrm{i}\xi Y_2)]E[\exp(\mathrm{i}\xi e_b)]}\mathrm{d}\xi \\
&= \int_0^\zeta \frac{\mathrm{i}E[Y_2\exp(\mathrm{i}\xi Y_2)]}{E[\exp(\mathrm{i}\xi Y_2)]}\mathrm{d}\xi \\
&= \int_0^\zeta \frac{\partial}{\partial\xi}\ln(E[\exp(\mathrm{i}\xi Y_2)])\mathrm{d}\xi \\
&= \int_0^\zeta (\frac{\partial}{\partial\xi}\ln(E[\exp(\mathrm{i}\xi Y_2)]) - \ln 1)\mathrm{d}\xi \\
&= \ln(E[\exp(\mathrm{i}\zeta Y_2)])
\end{aligned}$$

where the law of iterated expectation is used in the third equality, $E[e_a \mid Y_{2b}^*] = 0$ is used in the fifth equality, $e_b \perp Y_2$ is used in the sixth equality, and $\ln 1 = 0$ is used in the ninth equality. Thus we get

$$E[\exp(\mathrm{i}\zeta Y_2)] = \exp\left(\int_0^\zeta \frac{\mathrm{i}E[Y_{2a}^*\exp(\mathrm{i}\xi Y_{2b}^*)]}{E[\exp(\mathrm{i}\xi Y_{2b}^*)]}\mathrm{d}\xi\right).$$

Further observe that from $e_b \perp Y_2 \mid Z$ (which is implied by $e_b \perp (Y_2, Z)$)

$$\begin{aligned}
E[\exp(\mathrm{i}\zeta Y_2)\mid Z] &= \frac{E[\exp(\mathrm{i}\zeta Y_2)\mid Z]E[\exp(\mathrm{i}\zeta Y_2)]E[\exp(\mathrm{i}\zeta e_b)]}{E[\exp(\mathrm{i}\zeta Y_2)]E[\exp(\mathrm{i}\zeta e_b)]} \\
&= \frac{E[\exp(\mathrm{i}\zeta Y_2)\mid Z]E[\exp(\mathrm{i}\zeta e_b)\mid Z]}{E[\exp(\mathrm{i}\zeta Y_{2b}^*)]}E[\exp(\mathrm{i}\zeta Y_2)] \\
&= \frac{E[\exp(\mathrm{i}\zeta(Y_2+e_b))\mid Z]}{E[\exp(\mathrm{i}\zeta Y_{2b}^*)]}E[\exp(\mathrm{i}\zeta Y_2)] \\
&= \frac{E[\exp(\mathrm{i}\zeta Y_{2b}^*)\mid Z]}{E[\exp(\mathrm{i}\zeta Y_{2b}^*)]}\exp\left(\int_0^\zeta \frac{\mathrm{i}E[Y_{2a}^*\exp(\mathrm{i}\xi Y_{2b}^*)]}{E[\exp(\mathrm{i}\xi Y_{2b}^*)]}\mathrm{d}\xi\right)
\end{aligned}$$

where the right-hand side is a function of all observables, which implies the identification of $E[\exp(\mathrm{i}\zeta Y_2)\mid Z]$. This completes the identification result for $F_{Y_2|Z}(y_2 \mid z)$ through (19).

E Proof of Lemma 2.5

For the identification of $f_{Y_2Z}(y_2, z)$, we use a similar argument to Theorem 2.1. By Assumptions 2.8-2.10, we have the integral equation

$$f_{ZY_2^*|U}(z, y_2^* \mid u) = \int_{\mathcal{Y}_2} f_{Z|Y_2}(z \mid y_2) f_{Y_2^*|Y_2}(y_2^* \mid y_2) f_{Y_2|U}(y_2 \mid u) \mathrm{d}y_2. \tag{20}$$

Then from the proof of Theorem 1 of Hu and Schennach (2008), we note that the densities $(f_{Z|Y_2}, f_{Y_2^*|Y_2}, f_{Y_2|U})$ are uniquely identified from the observable joint density $f_{ZY_2^*|U}(z, y_2^* \mid u)$ by Assumptions 2.11-2.13. Then because $f_{Y_2}(y_2) = \int f_{Y_2|U}(y_2 \mid u) f_U(u) \mathrm{d}u$, $f_{U|Y_2}(u \mid y_2) = f_{Y_2|U}(y_2 \mid u) f_U(u)/f_{Y_2}(y_2)$, $f_{Y_2Z}(y_2, z) = f_{Z|Y_2}(z \mid y_2) f_{Y_2}(y_2)$, and $f_U(u)$ is directly observable from data, we conclude the densities $(f_{Y_2Z}, f_{Y_2^*|Y_2}, f_{U|Y_2})$ are uniquely identified from the observables (Y_2^*, Z, U).

F Proof of Theorem 4.1

We first prove the consistency of the sieve MLE $\hat{\beta}_n$ in the norm $\|\cdot\|_{s,\beta}$ by checking conditions of Theorem 4.1 in Newey and Powell (2003). Their Condition 1 on the identification of β_0 is implied by Assumptions 2.2-2.7. Condition 2 is satisfied by Assumption 4.3(i). Let $\bar{\beta} \equiv (\bar{f}_1, \bar{f}_2, \bar{f}_3)'$ be a mean value between β_1 and β_2. The bound for the path-wise derivative is given by

$$\begin{aligned}
& \left| \frac{d}{dt} \ln f_{Y^*|W}(y^* \mid w; \bar{\beta} + t(\beta_1 - \beta_2)) \right|_{t=0} \\
\leq \;& \frac{1}{|f_{Y^*|W}(y^* \mid w; \bar{\beta})|} \left\{ \int_{\mathcal{Y}_2} \left| \omega^{-1}(y_1, y_2, z_1) \bar{f}_2(y_2^* \mid y_2, z_1) \bar{f}_3(y_2 \mid v, z_1) \right| dy_2 \right. \\
& + \int_{\mathcal{Y}_2} \left| \bar{f}_1(y_1 \mid y_2, z_1) \omega^{-1}(y_2^*, y_2, z_1) \bar{f}_3(y_2 \mid v, z_1) \right| dy_2 \\
& \left. + \int_{\mathcal{Y}_2} \left| \bar{f}_1(y_1 \mid y_2, z_1) \bar{f}_2(y_2^* \mid y_2, z_1) \omega^{-1}(y_2, v, z_1) \right| dy_2 \right\} \|\beta_1 - \beta_2\|_{s,\beta} \\
\equiv \;& \left| \frac{f^{|1|}_{Y^*|W}(y^* \mid w; \bar{\beta}, \bar{\omega})}{f_{Y^*|W}(y^* \mid w; \bar{\beta})} \right| \|\beta_1 - \beta_2\|_{s,\beta},
\end{aligned}$$

where $f^{|1|}_{Y^*|W}(y^* \mid w; \bar{\beta}, \bar{\omega})$ is defined as $\frac{d}{dt} f_{Y^*|W}(y^* \mid w; \bar{\beta} + t\bar{\omega})|_{t=0}$ with $\bar{\omega}(y_1, y_2^*, y_2, v, z_1) = [\omega^{-1}(y_1, y_2, z_1), \omega^{-1}(y_2^*, y_2, z_1), \omega^{-1}(y_2, v, z_1)]'$ and with $\bar{f}_1$, $\bar{f}_2$, and $\bar{f}_3$ being replaced by their absolute values, respectively. Thus, Assumption 4.1 (iv) implies $\ln f_{Y^*|W}(y^* \mid w; \beta)$ is Hölder continuous in $\beta \in \mathcal{B}$ so that their Condition 3 holds with Assumption 4.1 (iii). Condition 4 is satisfied by Assumption 4.2 (ii). Condition 5 is also satisfied by Assumption 4.3 (iii).

To derive the consistency of $\hat{\alpha}_n$, let

$$\begin{aligned}\hat{m}(w,\alpha) &\equiv \int_{\mathcal{Y}_2}\left[\int_{\mathcal{Y}_1}\rho(x,\theta,h)\hat{f}_{Y_1|Y_2Z_1}(y_1\mid y_2,z_1)dy_1\right]\hat{f}_{Y_2|VZ_1}(y_2\mid v,z_1)dy_2\\ \tilde{m}(w,\alpha) &\equiv \int_{\mathcal{Y}_2}\left[\int_{\mathcal{Y}_1}m(w,\theta,h)\hat{f}_{Y_1|Y_2Z_1}(y_1\mid y_2,z_1)dy_1\right]\hat{f}_{Y_2|VZ_1}(y_2\mid v,z_1)dy_2,\end{aligned}$$

where $\hat{\beta}_n=(\hat{f}_{Y_1|Y_2Z_1},\hat{f}_{Y_2^*|Y_2Z_1},\hat{f}_{Y_2|VZ_1})'$ is the sieve ML estimator in (12) and $\tilde{m}(w,\alpha)$ denotes the projection of $m(w,\alpha)$ on the estimated densities $\hat{\beta}_n$. By Lemma 4 of Huang (1998), we have

$$\sup_{\alpha\in\mathcal{A}_n} n^{-1}\sum_{i=1}^{n}\|\hat{m}(W_i,\alpha)-\tilde{m}(W_i,\alpha)\|_E^2 \asymp \sup_{\alpha\in\mathcal{A}_n} E\left[n^{-1}\sum_{i=1}^{n}\|\hat{m}(W_i,\alpha)-\tilde{m}(W_i,\alpha)\|_E^2\right].$$

We also note that for some $\tilde{\beta}=(\tilde{f}_1\equiv\tilde{f}_{Y_1|Y_2Z_1}(y_1\mid y_2,z_1),\tilde{f}_2,\tilde{f}_3\equiv\tilde{f}_{Y_2|VZ_1}(y_2\mid v,z_1))'\in\mathcal{B}_n$ that satisfies Assumptions 4.3 (iii) and 4.6 (ii), we have

$$\begin{aligned}&E\left[n^{-1}\sum_{i=1}^{n}\|\hat{m}(W_i,\alpha)-\tilde{m}(W_i,\alpha)\|_E^2\right]\\ \leq\ &\frac{1}{n}E\Bigg[\sum_{i=1}^{n}\bigg|\int\int\rho(y,W_i,\alpha)\Big(\hat{f}_{Y_1|Y_2Z_1}(y_1\mid y_2,Z_{1i})\hat{f}_{Y_2|VZ_1}(y_2\mid V_i,Z_{1i})\\ &-\tilde{f}_{Y_1|Y_2Z_1}(y_1\mid y_2,Z_{1i})\tilde{f}_{Y_2|VZ_1}(y_2\mid V_i,Z_{1i})\Big)dy_1dy_2\bigg|^2\Bigg]\\ \leq\ &\frac{1}{n}E\Bigg[\sum_{i=1}^{n}\bigg|\int\int\rho(y,W_i,\alpha)dy_1dy_2\bigg|^2\Bigg]\|\hat{\beta}_n-\tilde{\beta}\|_{s,\beta}^2\\ \asymp\ &\|\hat{\beta}_n-\tilde{\beta}\|_{s,\beta}^2=O_p(l_n/n),\end{aligned}$$

by Assumptions 4.3 (iii) and 4.6 (i). Then

$$\begin{aligned}&\sup_{\alpha\in\mathcal{A}_n}E[\|\hat{m}(W,\alpha)-m(W,\alpha)\|_E^2] \qquad (21)\\ \leq\ &2\sup_{\alpha\in\mathcal{A}_n}E[\|\hat{m}(W,\alpha)-\tilde{m}(W,\alpha)\|_E^2]+2\sup_{\alpha\in\mathcal{A}_n}E[\|\tilde{m}(W,\alpha)-m(W,\alpha)\|_E^2]\\ =\ &\delta_{m,n}^2\end{aligned}$$

with $\delta_{m,n}^2=\max\{\frac{l_n}{n},b_{m,l_n}^2\}$ by Assumptions 4.6 (ii)-(iii). Assumption 4.5 (ii) implies that there are finite constants c_1,c_2 such that

$$c_1E\left[\|m(W,\alpha)\|_E^2\right] \leq E\left[\|\Sigma(W)^{-1/2}m(W,\alpha)\|_E^2\right]\leq c_2E\left[\|m(W,\alpha)\|_E^2\right]$$

uniformly over $\alpha\in\mathcal{A}_n$. Then applying the above results, for $\lambda_nP(h)\geq 0$, $\varepsilon>0$ and n sufficiently large, we have

$$
\begin{aligned}
& Pr(\|\hat{\alpha}_n - \alpha_0\|_{s,\alpha} \geq \varepsilon) \\
\leq \quad & Pr(\|\hat{\alpha}_n - \alpha_0\|_{s,\alpha} \geq \varepsilon, \hat{\alpha}_n \in \mathcal{A}_n^{M_0}) + Pr(\hat{\alpha}_n \notin \mathcal{A}_n^{M_0}) \\
\leq \quad & Pr\Bigg(\inf_{\alpha \in \mathcal{A}_n^{M_0} : \|\alpha - \alpha_0\|_{s,\alpha} \geq \varepsilon} \left\{ \frac{1}{n} \sum_{i=1}^{n} \|\hat{\Sigma}(W_i)^{-1/2} \hat{m}(W_i, \alpha)\|_E^2 + \lambda_n \hat{P}_n(h) \right\} \\
& \qquad \leq \frac{1}{n} \sum_{i=1}^{n} \|\hat{\Sigma}(W_i)^{-1/2} \hat{m}(W_i, \Pi_n \alpha_0)\|_E^2 + \lambda_n \hat{P}_n(\Pi_n h_0) \Bigg) + Pr(\hat{\alpha}_n \notin \mathcal{A}_n^{M_0}) \\
\leq \quad & Pr\Bigg(\inf_{\alpha \in \mathcal{A}_n^{M_0} : \|\alpha - \alpha_0\|_{s,\alpha} \geq \varepsilon} \Big\{ E[\|\Sigma(W)^{-1/2} m(W, \alpha)\|_E^2] + \lambda_n P(h) \Big\} \\
& \qquad \leq E[\|\Sigma(W)^{-1/2} m(W, \Pi_n \alpha_0)\|_E^2] + O_p(\delta_{m,n}^2) + \lambda_n P(h_0) + O_p(\lambda_n) \Bigg) + Pr(\hat{\alpha}_n \notin \mathcal{A}_n^{M_0}) \\
\leq \quad & Pr\Bigg(\inf_{\alpha \in \mathcal{A}_n^{M_0} : \|\alpha - \alpha_0\|_{s,\alpha} \geq \varepsilon} \left\{ c_1 E[\|m(W, \alpha)\|_E^2] + \lambda_n P(h) \right\} \\
& \qquad \leq c_2 E[\|m(W, \Pi_n \alpha_0)\|_E^2] + O_p(\delta_{m,n}^2) + \lambda_n P(h_0) + O_p(\lambda_n) \Bigg) + Pr(\hat{\alpha}_n \notin \mathcal{A}_n^{M_0}) \\
\leq \quad & Pr\Bigg(O_p(\max\{\delta_{m,n}^2, E[m(W, \Pi_n \alpha_0)' m(W, \Pi_n \alpha_0)], \lambda_n\}) \\
& \qquad \geq \inf_{\alpha \in \mathcal{A}_n^{M_0} : \|\alpha - \alpha_0\|_{s,\alpha} \geq \varepsilon} E[m(W, \alpha)' m(W, \alpha)] \Bigg) + Pr(\hat{\alpha}_n \notin \mathcal{A}_n^{M_0}) \\
\longrightarrow \quad & 0,
\end{aligned}
$$

where the third inequality holds by Assumptions 4.4, 4.5, and 4.6 and by (21) and where the last result holds since $\max\{\delta_{m,n}^2, E\left[|m(W, \Pi_n \alpha_0)|^2\right], \lambda_n\} / \inf_{\alpha \in \mathcal{A}_n^{M_0} : \|\alpha - \alpha_0\|_{s,\alpha} \geq \varepsilon} E\left[|m(W, \alpha)|^2\right] = o(1)$ for any $\varepsilon > 0$ and we can take $\hat{\alpha}_n \in \mathcal{A}_n^{M_0}$ with probability approaching one. So the consistency result $\|\hat{\alpha}_n - \alpha_0\|_{s,\alpha} = o_p(1)$ follows.

G Proof of Theorem 5.1

To show the first part of the theorem, define $s_n^2 = \max\left\{\delta_{m,n}^2, \|\alpha_0 - \Pi_n \alpha_0\|_{\alpha}^2, \lambda_n |P(\Pi_n h_0) - P(h)|\right\} = o_p(1)$. Note that Assumption 4.5 (ii) implies that there are finite constants c_1, c_2 such that

$$
c_1 \frac{1}{n} \sum_{i=1}^{n} \|\hat{m}(W_i, \alpha)\|_E^2 \quad \leq \quad \frac{1}{n} \sum_{i=1}^{n} \|\hat{\Sigma}(W_i)^{-1/2} \hat{m}(W_i, \alpha)\|_E^2 \leq c_2 \frac{1}{n} \sum_{i=1}^{n} \|\hat{m}(W_i, \alpha)\|_E^2
$$

uniformly over $\alpha \in \mathcal{A}_n$. Also note that $\sup_{h\in\mathcal{H}_{osn}} |\hat{P}_n(h) - P(h)| = o_p(1)$ by the assumption in the statement of the theorem. Since $\hat{\alpha}_n \in \mathcal{A}_{osn}$ with probability approaching one, we get, for all $M > 1$,

$$
\begin{aligned}
& Pr(\|\hat{\alpha}_n - \alpha_0\|_\alpha \geq M s_n) \\
\leq \quad & Pr\Bigg(\inf_{\alpha\in\mathcal{A}_{osn}:\|\alpha-\alpha_0\|_\alpha \geq M s_n} \left\{ \frac{1}{n}\sum_{i=1}^{n} \|\hat{\Sigma}(W_i)^{-1/2}\hat{m}(W_i,\alpha)\|_E^2 + \lambda_n \hat{P}_n(h) \right\} \\
& \quad \leq \frac{1}{n}\sum_{i=1}^{n} \|\hat{\Sigma}(W_i)^{-1/2}\hat{m}(W_i,\Pi_n\alpha_0)\|_E^2 + \lambda_n \hat{P}_n(\Pi_n h_0) \Bigg) \\
\leq \quad & Pr\Bigg(\inf_{\alpha\in\mathcal{A}_{osn}:\|\alpha-\alpha_0\|_\alpha \geq M s_n} \left\{ c_1 \frac{1}{n}\sum_{i=1}^{n} \|\hat{m}(W_i,\alpha)\|_E^2 + \lambda_n \hat{P}_n(h) \right\} \\
& \quad \leq c_2 \frac{1}{n}\sum_{i=1}^{n} \|\hat{m}(W_i,\Pi_n\alpha_0)\|_E^2 + \lambda_n \hat{P}_n(\Pi_n h_0) \Bigg) \\
\leq \quad & Pr\Bigg(\inf_{\alpha\in\mathcal{A}_{osn}:\|\alpha-\alpha_0\|_\alpha \geq M s_n} \left\{ c_1 \frac{1}{n}\sum_{i=1}^{n} \|\hat{m}(W_i,\alpha)\|_E^2 + \lambda_n P(h) \right\} \\
& \quad \leq c_2 \frac{1}{n}\sum_{i=1}^{n} \|\hat{m}(W_i,\Pi_n\alpha_0)\|_E^2 + \lambda_n P(\Pi_n h_0) + o_p(\lambda_n) \Bigg) \\
\leq \quad & Pr\Bigg(\inf_{\alpha\in\mathcal{A}_{osn}:\|\alpha-\alpha_0\|_\alpha \geq M s_n} \left\{ c_1 E[\|m(W,\alpha)]\|_E^2 \right\} \\
& \quad \leq c_2 E[\|m(W,\Pi_n\alpha_0)\|_E^2] + O_p(\delta_{m,n}^2) + \lambda_n P(\Pi_n h_0) - \lambda_n P(h) + o_p(\lambda_n) \Bigg) \\
\leq \quad & Pr\Big(M^2 s_n^2 \leq O_p(\max\left\{\delta_{m,n}^2, \|\alpha_0 - \Pi_n\alpha_0\|_\alpha^2, \lambda_n |P(\Pi_n h_0) - P(h)|\right\}) \Big), \\
\longrightarrow \quad & 0,
\end{aligned}
$$

where the last inequality follows by Assumptions 4.6 and 5.5. As a result, we get $\|\hat{\alpha}_n - \alpha_0\|_\alpha = O_p(\max\left\{\delta_{m,n}, \|\alpha_0 - \Pi_n\alpha_0\|_\alpha, \sqrt{\lambda_n}\right\})$.

To show the second part, we note that $\delta_{m,n} = \max\left\{\sqrt{\frac{l_n}{n}}, b_{m,l_n}\right\} = \sqrt{\frac{l_n}{n}} = const. \times \sqrt{\frac{k_n}{n}} = o(1)$ by Assumptions 5.6-5.7 where $k_n = \dim(\mathcal{H}_n)$. Then under the conditions $\|h_0 - \Pi_n h_0\|_\alpha = o(n^{-1/4})$ and $\max\left\{\delta_{m,n}, \sqrt{\lambda_n}\right\} = \delta_{m,n}$, we have $\|\hat{\alpha}_n - \alpha_0\|_\alpha = O_p(\delta_{m,n})$. We now show $\delta_{m,n} = o(n^{-1/4})$. First, we show that $\|\hat{\beta}_n - \beta_0\|_\beta = o_p(n^{-1/4})$ by checking conditions of Theorem 3.1 in Ai and Chen (2003). Their Conditions 3.5 (iii)-3.6 (iii) are satisfied by Assumption 5.1. Their Conditions 3.7 and 3.8 are satisfied by Assumption 5.2. Assumption 5.4 (ii) implies Condition 3.9 in Ai and Chen

(2003). Thus the $n^{-1/4}$ convergence rate of $\hat{\beta}_n$ in $\|\cdot\|_\beta$ follows. Second, since

$$\begin{aligned}
&\frac{d \ln f_{Y^*|W}(y^* \mid w; \beta_0)}{d\beta}[\beta - \beta_0] \\
= \ & \frac{1}{f_{Y^*|W}(y^* \mid w; \beta_0)} \bigg\{ \int_{\mathcal{Y}_2} [f_1(\scriptstyle y_1|y_2,z_1\textstyle) - f_{Y_1|Y_2Z_1}(\scriptstyle y_1|y_2,z_1\textstyle)] f_{Y_2^*|Y_2Z_1}(y_2^* \mid y_2, z_1) f_{Y_2|VZ_1}(y_2 \mid v, z_1) dy_2 \\
& + \int_{\mathcal{Y}_2} f_{Y_1|Y_2Z_1}(y_1 \mid y_2, z_1)[f_2(y_2^* \mid y_2, z_1) - f_{Y_2^*|Y_2Z_1}(y_2^* \mid y_2, z_1)] f_{Y_2|VZ_1}(y_2 \mid v, z_1) dy_2 \\
& + \int_{\mathcal{Y}_2} f_{Y_1|Y_2Z_1}(y_1 \mid y_2, z_1) f_{Y_2^*|Y_2Z_1}(y_2^* \mid y_2, z_1)[f_3(y_2 \mid v, z_1) - f_{Y_2|VZ_1}(y_2 \mid v, z_1)] dy_2 \bigg\}
\end{aligned}$$

and

$$\|\hat{\beta}_n - \beta_0\|_\beta \ \equiv \ \sqrt{E\left\{\left(\frac{d \ln f_{Y^*|W}(Y^* \mid W; \beta_0)}{d\beta}[\hat{\beta}_n - \beta_0]\right)^2\right\}},$$

we have that for a constant $c > 0$,

$$\begin{aligned}
& E[\|\hat{m}(W, \alpha) - m(W, \alpha)\|_E^2] \\
= \ & \frac{1}{n}\sum_{i=1}^n E\left[\left|\int \rho(y, W_i, \alpha)(\hat{f}_{Y|W}(y|W_i) - f_{Y|W}(y|W_i))dy\right|^2\right] \\
= \ & \frac{1}{n}\sum_{i=1}^n E\bigg[\bigg|\int_{\mathcal{Y}_1}\int_{\mathcal{Y}_2} \rho(y, W_i, \alpha)(\hat{f}_{Y_1|Y_2Z_1}(y_1 \mid y_2, Z_{1i})\hat{f}_{Y_2|VZ_1}(y_2 \mid V_i, Z_{1i}) \\
& \quad - f_{Y_1|Y_2Z_1}(y_1 \mid y_2, Z_{1i}) f_{Y_2|VZ_1}(y_2 \mid V_i, Z_{1i}))dy_1 dy_2\bigg|^2\bigg] \\
\leq \ & c\frac{1}{n}\sum_{i=1}^n \sup_{y_2, y_2^*, w} \int_{\mathcal{Y}_1} |\rho(y_1, y_2, w, \alpha)|^2 f_{Y^*|W}(y_1, y_2^* \mid w; \beta_0) dy_1 E\bigg[\bigg(\frac{1}{f_{Y^*|W}(Y_i^* \mid W_i; \beta_0)} \\
& \times \bigg\{\int_{\mathcal{Y}_2} [\hat{f}_1(Y_{1i} \mid y_2, Z_{1i}) - f_{Y_1|Y_2Z_1}(Y_{1i} \mid y_2, Z_{1i})] f_{Y_2^*|Y_2Z_1}(Y_{2i}^* \mid y_2, Z_{1i}) f_{Y_2|VZ_1}(y_2 \mid V_i, Z_{1i}) dy_2 \\
& \quad + \int_{\mathcal{Y}_2} f_{Y_1|Y_2Z_1}(Y_{1i} \mid y_2, Z_{1i})[\hat{f}_2(Y_{2i}^* \mid y_2, Z_{1i}) - f_{Y_2^*|Y_2Z_1}(Y_{2i}^* \mid y_2, Z_{1i})] f_{Y_2|VZ_1}(y_2 \mid V_i, Z_{1i}) dy_2 \\
& \quad + \int_{\mathcal{Y}_2} f_{Y_1|Y_2Z_1}(\scriptstyle Y_{1i}|y_2,Z_{1i}\textstyle) f_{Y_2^*|Y_2Z_1}(\scriptstyle Y_{2i}^*|y_2,Z_{1i}\textstyle)[\hat{f}_3(y_2 \mid V_i, Z_{1i}) - f_{Y_2|VZ_1}(y_2 \mid V_i, Z_{1i})] dy_2 \bigg\}\bigg)^2\bigg] \\
\leq \ & c\frac{1}{n}\sum_{i=1}^n \sup_{y_2, y_2^*, w} \int_{\mathcal{Y}_1} |\rho(y_1, y_2, w, \alpha)|^2 f_{Y^*|W}(y_1, y_2^* \mid w; \beta_0) dy_1 \|\hat{\beta}_n - \beta_0\|_\beta^2 \\
= \ & o_p(n^{-1/2}),
\end{aligned}$$

by Assumptions 4.1 and 4.6 (i), and $\|\hat{\beta}_n - \beta_0\|_\beta = o_p(n^{-1/4})$. Thus we get $\delta_{m,n} = o(n^{-1/4})$ so that $\|\hat{\alpha}_n - \alpha_0\|_\alpha = o_p(n^{-1/4})$.

H Proof of Theorem 5.2

Recall $\hat{m}(w,\alpha) \equiv \int_{\mathcal{Y}_2}[\int_{\mathcal{Y}_1} \rho(x,\alpha)\hat{f}_{Y_1|Y_2Z_1}(y_1 \mid y_2, z_1)dy_1]\hat{f}_{Y_2|VZ_1}(y_2 \mid v, z_1)dy_2$ and the projection of $m(w,\alpha)$ on the estimated densities as $\tilde{m}(w,\alpha) \equiv \int[\int m(w,\alpha)\hat{f}_{Y_1|Y_2Z_1}(y_1 \mid y_2, z_1)dy_1]\hat{f}_{Y_2|VZ_1}(y_2 \mid v, z_1)dy_2$.

Lemma H.1. *(i) Assumptions 4.1 (i)-(ii), 4.2 (i), 4.2 (iii), 4.3 (i), 4.5, 5.3 (i), 5.4 (i), 5.8 and 5.13 (i) imply that uniformly over $\tilde{\alpha} \in \mathcal{N}_{0n}$,*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{d^2\tilde{m}(W_i,\tilde{\alpha})}{d\alpha d\alpha}[b_n^*, b_n^*]\right)' \hat{\Sigma}(W_i)^{-1}\tilde{m}(W_i,\tilde{\alpha}) \quad = \quad o_p(n^{-1/4}).$$

(ii) Assumptions 4.1 (i)-(ii), 4.5, 5.5 (ii), 5.8, 5.9 (i), 5.9 (iii), 5.11 (i) and 5.13 (ii) imply that uniformly over $\tilde{\alpha} \in \mathcal{N}_{0n}$,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{d\tilde{m}(W_i,\tilde{\alpha})}{d\alpha}[b_n^*]\right)' \hat{\Sigma}(W_i)^{-1}\left(\frac{d\tilde{m}(W_i,\tilde{\alpha})}{d\alpha}[b_n^*]\right) \quad = \quad O_p(1).$$

Proof: (i) For a generic constant $c > 0$, uniformly over $\tilde{\alpha} \in \mathcal{N}_{0n}$, we have

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\left(\frac{d^2\tilde{m}(W_i,\tilde{\alpha})}{d\alpha d\alpha}[b_n^*, b_n^*]\right)' \hat{\Sigma}(W_i)^{-1}\tilde{m}(W_i,\tilde{\alpha}) \\
\leq \; & \sup_{w\in\mathcal{W}} \lambda_{min}^{-1}(\hat{\Sigma}(w))\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\frac{d^2\tilde{m}(W_i,\tilde{\alpha})}{d\alpha d\alpha}[b_n^*, b_n^*]\right\|_E^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\tilde{m}(W_i,\tilde{\alpha})\|_E^2} \\
\leq \; & c\sqrt{E[\|\tilde{m}(W,\tilde{\alpha})\|_E^2]} \\
= \; & o_p(n^{-1/4}),
\end{aligned}$$

where $\lambda_{min}(A)$ denotes the smallest eigenvalue of a matrix A and where the first inequality holds by Cauchy-Schwarz inequality, the second inequality holds by Assumptions 4.5 and 5.13 (i), and the last equality holds by Assumptions 5.4 (i), 5.8 and $\tilde{m}(w,\alpha_0) = 0$.

(ii) Uniformly over $\tilde{\alpha} \in \mathcal{N}_{0n}$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\frac{d\tilde{m}(W_i,\tilde{\alpha})}{d\alpha}[b_n^*]\right)' \hat{\Sigma}(W_i)^{-1}\left(\frac{d\tilde{m}(W_i,\tilde{\alpha})}{d\alpha}[b_n^*]\right) \\
= \;& \frac{1}{n}\sum_{i=1}^{n}\left(\frac{d\tilde{m}(W_i,\tilde{\alpha})}{d\alpha}[b_n^*] - \frac{dm(W_i,\alpha_0)}{d\alpha}[b_n^*]\right)' \hat{\Sigma}(W_i)^{-1}\left(\frac{d\tilde{m}(W_i,\tilde{\alpha})}{d\alpha}[b_n^*]\right) \\
&+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{dm(W_i,\alpha_0)}{d\alpha}[b_n^*]\right)' \hat{\Sigma}(W_i)^{-1}\left(\frac{d\tilde{m}(W_i,\tilde{\alpha})}{d\alpha}[b_n^*] - \frac{dm(W_i,\alpha_0)}{d\alpha}[b_n^*]\right) \\
&+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{dm(W_i,\alpha_0)}{d\alpha}[b_n^*]\right)' (\hat{\Sigma}(W_i)^{-1} - \Sigma(W_i)^{-1})\left(\frac{dm(W_i,\alpha_0)}{d\alpha}[b_n^*]\right) \\
&+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{dm(W_i,\alpha_0)}{d\alpha}[b_n^*]\right)' \Sigma(W_i)^{-1}\left(\frac{dm(W_i,\alpha_0)}{d\alpha}[b_n^*]\right) \\
= \;& o_p(n^{-1/2}) + O_p(1) \\
= \;& O_p(1),
\end{aligned}$$

by Assumptions 4.5, 5.5 (ii) and 5.13 (ii).

Lemma H.2. *Assumptions 4.1 (i)-(ii), 4.2 (i), 4.2 (iii), 4.3 (i), 4.5, 5.3 (i), 5.4 (i), 5.5 (i), 5.8, 5.9, 5.10 (i) and 5.11-5.13 imply that uniformly over $\tilde{\alpha} \in \mathcal{N}_{0n}$,*

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\frac{d\tilde{m}(W_i,\tilde{\alpha})}{d\alpha}[b_n^*]\right)' \hat{\Sigma}(W_i)^{-1}\tilde{m}(W_i,\tilde{\alpha}) \\
= \;& \frac{1}{n}\sum_{i=1}^{n}\left(\frac{dm(W_i,\alpha_0)}{d\alpha}[b^*]\right)' \Sigma(W_i)^{-1}\rho(X_i,\alpha_0) \\
&+E\left[\left(\frac{dm(W,\alpha_0)}{d\alpha}[b^*]\right)' \Sigma(W)^{-1}\left(\frac{dm(W,\alpha_0)}{d\alpha}[\tilde{\alpha}-\alpha_0]\right)\right] + o_p(n^{-1/2}).
\end{aligned}$$

Proof: Uniformly over $\tilde{\alpha} \in \mathcal{N}_{0n}$,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{d\tilde{m}(W_i,\tilde{\alpha})}{d\alpha}[b_n^*]\right)' \hat{\Sigma}(W_i)^{-1}\tilde{m}(W_i,\tilde{\alpha}) - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{dm(W_i,\alpha_0)}{d\alpha}[b^*]\right)' \Sigma(W_i)^{-1}\tilde{m}(W_i,\tilde{\alpha})\right| \\
\leq \;& \left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{d\tilde{m}(W_i,\tilde{\alpha})}{d\alpha}[b_n^*] - \frac{dm(W_i,\alpha_0)}{d\alpha}[b_n^*]\right)' \hat{\Sigma}(W_i)^{-1}\tilde{m}(W_i,\tilde{\alpha})\right| \\
&+\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{dm(W_i,\alpha_0)}{d\alpha}[b_n^*]\right)' (\hat{\Sigma}(W_i)^{-1} - \Sigma(W_i)^{-1})\tilde{m}(W_i,\tilde{\alpha})\right| \\
&+\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{dm(W_i,\alpha_0)}{d\alpha}[b_n^* - b^*]\right)' \Sigma(W_i)^{-1}\tilde{m}(W_i,\tilde{\alpha})\right| \\
\equiv \;& I_n + II_n + III_n.
\end{aligned}$$

Then by triangle inequality and Cauchy-Schwarz inequality, we have

$$\begin{aligned} I_n &\leq \sup_{w\in\mathcal{W}} \lambda_{min}^{-1}(\hat{\Sigma}(w))\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\tilde{m}(W_i,\tilde{\alpha})\|_E^2}\left(\sqrt{\frac{2}{n}\sum_{i=1}^{n}\left\|\frac{d\tilde{m}(W_i,\tilde{\alpha})}{d\alpha}[b_n^*]-\frac{d\tilde{m}(W_i,\alpha_0)}{d\alpha}[b_n^*]\right\|_E^2}\right.\\ &\quad \left.+\sqrt{\frac{2}{n}\sum_{i=1}^{n}\left\|\frac{d\tilde{m}(W_i,\alpha_0)}{d\alpha}[b_n^*]-\frac{dm(W_i,\alpha_0)}{d\alpha}[b_n^*]\right\|_E^2}\right)\\ &= o_p(n^{-1/4})\times(o_p(n^{-1/4})+o_p(n^{-1/4}))\\ &= o_p(n^{-1/2}), \end{aligned}$$

where the first equality holds by Assumptions 4.5 (ii), 5.11 (i) and 5.13 (ii) and $E[\|\tilde{m}(W,\tilde{\alpha})\|_E^2] = o_p(n^{-1/2})$ by Assumptions 5.4 (i), 5.8 and $\tilde{m}(w,\alpha_0)=0$. Note that by Cauchy-Schwarz inequality, we have

$$\begin{aligned} II_n &\leq \sup\nolimits_{w\in\mathcal{W}} \quad |\hat{\Sigma}(w)^{-1}-\Sigma(w)^{-1}|\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\frac{dm(W_i,\alpha_0)}{d\alpha}[b_n^*]\right\|_E^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\tilde{m}(W_i,\tilde{\alpha})\|_E^2}\\ &= \qquad o_p(n^{-1/4})\times o_p(n^{-1/4})\\ &= \qquad o_p(n^{-1/2}), \end{aligned}$$

where the first equality holds by Assumptions 5.5 (i), 5.10 (i) and $E[\|\tilde{m}(W,\tilde{\alpha})\|_E^2] = o_p(n^{-1/2})$. Also note that by Cauchy-Schwarz inequality, we have

$$\begin{aligned} III_n &\leq \sup_{w\in\mathcal{W}} \lambda_{min}^{-1}(\Sigma(w))\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\frac{dm(W_i,\alpha_0)}{d\alpha}[b_n^*-b^*]\right\|_E^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\tilde{m}(W_i,\tilde{\alpha})\|_E^2}\\ &= o_p(n^{-1/4})\times o_p(n^{-1/4})\\ &= o_p(n^{-1/2}), \end{aligned}$$

where the first equality holds by Assumptions 4.5 (iii), 5.9 (iii) and $E[\|\tilde{m}(W,\tilde{\alpha})\|_E^2] = o_p(n^{-1/2})$. As a result, we obtain

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\left(\frac{d\tilde{m}(W_i,\tilde{\alpha})}{d\alpha}[b_n^*]\right)'\hat{\Sigma}(W_i)^{-1}\tilde{m}(W_i,\tilde{\alpha})\\ =\;& \frac{1}{n}\sum_{i=1}^{n}\left(\frac{dm(W_i,\alpha_0)}{d\alpha}[b^*]\right)'\Sigma(W_i)^{-1}\tilde{m}(W_i,\tilde{\alpha})+o_p(n^{-1/2}) \qquad (22) \end{aligned}$$

Recall that $g(W,b^*) \equiv \left(\frac{dm(W,\alpha_0)}{d\alpha}[b^*]\right)'\Sigma(W)^{-1}$ and its projection onto the integral function

$\tilde{g}(W, b^*) \equiv \int[\int g(W, b^*)\hat{f}_{Y_1|Y_2Z_1}(y_1 \mid y_2, z_1)dy_1]\hat{f}_{Y_2|VZ_1}(y_2 \mid v, z_1)dy_2$ and note that

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n} g(W_i, b^*)\tilde{m}(W_i, \tilde{\alpha}) \\
= \; & \frac{1}{n}\sum_{i=1}^{n} g(W_i, b^*)(\hat{m}(W_i, \alpha_0) + \tilde{m}(W_i, \tilde{\alpha})) + o_p(n^{-1/2}) \\
= \; & \frac{1}{n}\sum_{i=1}^{n} \tilde{g}(W_i, b^*)(\rho(X_i, \alpha_0) + m(W_i, \tilde{\alpha})) + o_p(n^{-1/2}) \qquad (23) \\
= \; & \frac{1}{n}\sum_{i=1}^{n} g(W_i, b^*)(\rho(X_i, \alpha_0) + m(W_i, \tilde{\alpha})) + o_p(n^{-1/2}),
\end{aligned}$$

where the first, second, third equalities hold by a similar argument to the proof of Theorem 4.1, definitions of $\tilde{g}(W, b^*)$ and $\tilde{m}(W, \alpha)$, and Assumption 5.11 (ii), respectively.

Since $\left\{g(W, b^*)m(W, \alpha) : \alpha \in \mathcal{N}_{0n}, m \in \Lambda_c^{\gamma,\omega}(\mathcal{W})\right\}$ is a Donsker class by Assumption 5.12, we have by a first-order Taylor expansion, for $\bar{\alpha}$ between $\tilde{\alpha}$ and α_0,

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n} g(W_i, b^*)(m(W_i, \tilde{\alpha}) - m(W_i, \alpha_0)) \\
= \; & E[g(W, b^*)(m(W, \tilde{\alpha}) - m(W, \alpha_0))] + o_p(n^{-1/2}) \\
= \; & E\left[g(W, b^*)\left(\frac{dm(W, \alpha_0)}{d\alpha}[\tilde{\alpha} - \alpha_0]\right)\right] \qquad (24) \\
& + E\left[g(W, b^*)\left(\frac{dm(W, \bar{\alpha})}{d\alpha}[\tilde{\alpha} - \alpha_0] - \frac{dm(W, \alpha_0)}{d\alpha}[\tilde{\alpha} - \alpha_0]\right)\right] + o_p(n^{-1/2}) \\
= \; & E\left[g(W, b^*)\left(\frac{dm(W, \alpha_0)}{d\alpha}[\tilde{\alpha} - \alpha_0]\right)\right] + o_p(n^{-1/2}),
\end{aligned}$$

where the third equality holds by Assumption 5.13 (iii). Thus, by combining the equations (22)-(24), we obtain the result.

Proof of Theorem 5.2: We follow similar steps in the proof of Theorem 4.1 in Ai and Chen (2003). Recall

$$\hat{Q}_n(\alpha) \;\; \equiv \;\; \left\{\frac{1}{n}\sum_{i=1}^{n} \hat{m}(W_i, \alpha)'[\hat{\Sigma}(W_i)]^{-1}\hat{m}(W_i, \alpha) + \lambda_n \hat{P}_n(h)\right\}.$$

Let $\varepsilon_n = o(n^{-1/2})$ be a positive sequence and $u_n^* = \pm b_n^*$. Take a continuous path $\{\alpha(t) \in \mathcal{N}_{0n} : t \in [0, 1]\}$ such that $\alpha(0) = \hat{\alpha}_n$ and $\alpha(1) = \hat{\alpha}_n + \varepsilon_n u_n^*$. By Assumptions 5.11 (i) and 5.13 (i), $\hat{Q}_n(\alpha(t))$ is twice continuously differentiable. By definition of $\hat{\alpha}_n$ and a second-order Taylor expansion around

$t = 0$, we have

$$\begin{aligned} 0 &\leq -\hat{Q}_n(\hat{\alpha}_n) + \hat{Q}_n(\hat{\alpha}_n + \varepsilon_n u_n^*) \\ &= \hat{Q}_n(\alpha(1)) - \hat{Q}_n(\alpha(0)) \\ &= \left.\frac{d\hat{Q}_n(\alpha(t))}{dt}\right|_{t=0} + \frac{1}{2}\left.\frac{d^2\hat{Q}_n(\alpha(t))}{dt^2}\right|_{t=s} \end{aligned}$$

with $s \in [0, 1]$. Since

$$\frac{1}{n}\sum_{i=1}^{n} \|\hat{\Sigma}(W_i)^{-1/2}\hat{m}(W_i,\alpha)\|_E^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} \|\hat{\Sigma}(W_i)^{-1/2}\tilde{m}(W_i,\alpha)\|_E^2 + o_p(n^{-1})$$

by Assumption 4.5 (ii) and a similar argument to the proof of Theorem 4.1, we get for $s \in [0, 1]$

$$\begin{aligned} 0 \;\leq\; & \frac{2}{n}\sum_{i=1}^{n}\left(\frac{d\tilde{m}(W_i,\hat{\alpha}_n)}{d\alpha}[\varepsilon_n u_n^*]\right)' \hat{\Sigma}(W_i)^{-1}\tilde{m}(W_i,\hat{\alpha}_n) \\ & + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{d^2\tilde{m}(W_i,\alpha(s))}{d\alpha d\alpha}[\varepsilon_n u_n^*, \varepsilon_n u_n^*]\right)' \hat{\Sigma}(W_i)^{-1}\tilde{m}(W_i,\alpha(s)) \\ & + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{d\tilde{m}(W_i,\alpha(s))}{d\alpha}[\varepsilon_n u_n^*]\right)' \hat{\Sigma}(W_i)^{-1}\left(\frac{d\tilde{m}(W_i,\alpha(s))}{d\alpha}[\varepsilon_n u_n^*]\right) + o_p(n^{-1}) \\ \leq\; & \frac{2\varepsilon_n}{n}\sum_{i=1}^{n}\left(\frac{d\tilde{m}(W_i,\hat{\alpha}_n)}{d\alpha}[u_n^*]\right)' \hat{\Sigma}(W_i)^{-1}\tilde{m}(W_i,\hat{\alpha}_n) + O_p(\varepsilon_n^2), \end{aligned}$$

where the last inequality holds by Lemma H.1. Since $\varepsilon_n = o(n^{-1/2}) > 0$ and $u_n^* = \pm b_n^*$, we thus obtain

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{d\tilde{m}(W_i,\hat{\alpha}_n)}{d\alpha}[b_n^*]\right)' \hat{\Sigma}(W_i)^{-1}\tilde{m}(W_i,\hat{\alpha}_n) \;=\; o_p(n^{-1/2}).$$

Then by Lemma H.2 and definition of $\langle b^*, \hat{\alpha}_n - \alpha_0\rangle_\alpha$, we get

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{dm(W_i,\alpha_0)}{d\alpha}[b^*]\right)' \Sigma(W_i)^{-1}\rho(X_i,\alpha_0) \quad + \langle b^*, \hat{\alpha}_n - \alpha_0\rangle_\alpha = \quad o_p(n^{-1/2}),$$

so that

$$\sqrt{n}\langle b^*, \hat{\alpha}_n - \alpha_0\rangle_\alpha \;=\; -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{dm(W_i,\alpha_0)}{d\alpha}[b^*]\right)' \Sigma(W_i)^{-1}\rho(X_i,\alpha_0) + o_p(1).$$

Thus, we obtain the result by applying a standard central limit theorem for i.i.d. data.

References

[1] Ai, C. and X. Chen (2003), "Efficient Estimation of Models With Conditional Moment Restrictions," *Econometrica*, 71, 1795–1843.

[2] Altonji, J.G. and R.L. Matzkin (2005), "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica,* 73, 1053-1102.

[3] Blundell, R., X. Chen, and D. Kristensen (2007), "Semi-nonparametric IV Estimation of Shape-invariant Engel Curves," *Econometrica,* 75, 1613-1669.

[4] Butcher, K. and A. Case (1994): "The Effects of Sibling Sex Composition on Women's Education and Earnings," *Quarterly Journal of Economics*, 109, 531-563.

[5] Case, A., D. Lubotsky, and C. Paxson (2002), "Economic Status and Health in Childhood: The Origins of the Gradient," *American Economic Review*, 92, 1308-1334.

[6] Chen, X., H. Hong, E. Tamer (2005), "Measurement Error Models with Auxiliary Data," *Review of Economic Studies*, 72, 343–366.

[7] Chen, X. and D. Pouzo (2012), "Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Moments," *Econometrica,* 80, 277-321.

[8] Chen, X. and X. Shen (1998), "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica,* 66, 289-314.

[9] Condliffe, S. and C.R. Link (2008), "The Relationship between Economic Status and Child Health: Evidence from the United States," *American Economic Review*, 98, 1605-1618.

[10] Currie, J. and M. Stabile (2003), "Socioeconomic Status and Child Health: Why Is the Relationship Stronger for Older Children?" *American Economic Review*, 93, 1813-1823.

[11] Dawid, A.P. (1979a), "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society, Series B*, 41, 1-31.

[12] Dawid, A.P. (1979b), "Some Misleading Arguments Involving Conditional Independence," *Journal of the Royal Statistical Society, Series B*, 41, 249-252.

[13] Dawid, A.P. (1980), "Conditional Independence for Statistical Operations," *Annals of Statistics,* 8, 598-617.

[14] Dehejia, R.H. and S. Wahba (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association,* 94, 1053-1062.

[15] Gurland, J. (1948), "Inversion Formulae for the Distribution of Ratios," *Annals of Mathematical Statistics,* 19, 228-237.

[16] Hahn, J. and G. Ridder (2013), "Asymptotic Variance of Semiparametric Estimators with Generated Regressors," *Econometrica,* 81, 315-340.

[17] Heckman, J., H. Ichimura, and P. Todd (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

[18] Heckman, J. and E. Vytlacil (2005), "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica,* 73, 669-738.

[19] Hu, Y. and S.M. Schennach (2008), "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica,* 76, 195–216.

[20] Huang, J. (1998), "Projection Estimation in Multiple Regression With Application to Functional ANOVA Models," *Annals of Statistics*, 26, 242-272.

[21] Imbens, G. and W. Newey (2009), "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512.

[22] Imbens, G. and J. M. Wooldridge (2009), "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature*, 47, 5–86.

[23] Lechner, M. (2001), "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption," in M. Lechner, F. Pfeiffer (Eds.), *Econometric Evaluation of Labour Market Policies*, Heidelberg: Physica, 43-58.

[24] Li, T. and Q. Vuong (1998), "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," *Journal of Multivariate Analysis,* 65, 139-165.

[25] Matzkin, R.L. (2003), "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71, 1339–1375.

[26] Newey, W. and J. Powell (2003), "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

[27] Newey, W., J. Powell, and F. Vella (1999), "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565–603.

[28] Pearl, J. (2009), *Causality,* Cambridge University Press.

[29] Schennach, S.M. (2004), "Estimation of Nonlinear Models with Measurement Error," *Econometrica,* 72, 33-75.

[30] Shen, X. (1997), "On Methods of Sieves and Penalization," *Annals of Statistics,* 22, 2555-2591.

[31] Song, K. (2009), "Testing Conditional Independence via Rosenblatt Transforms," *Annals of Statistics,* 37, 4011-4045.

[32] Song, S. (2015), "Semiparametric Estimation of Models with Conditional Moment Restrictions in the Presence of Nonclassical Measurement Errors," *Journal of Econometrics*, 185, 95-109.

[33] Su, L. and H. White (2007), "A Consistent Characteristic Function-based Test for Conditional Independences," *Journal of Econometrics,* 141, 807-834.

[34] Van der Vaart, A. (1991), "On Differentiable Functionals," *Annals of Statistics,* 19, 178-204.

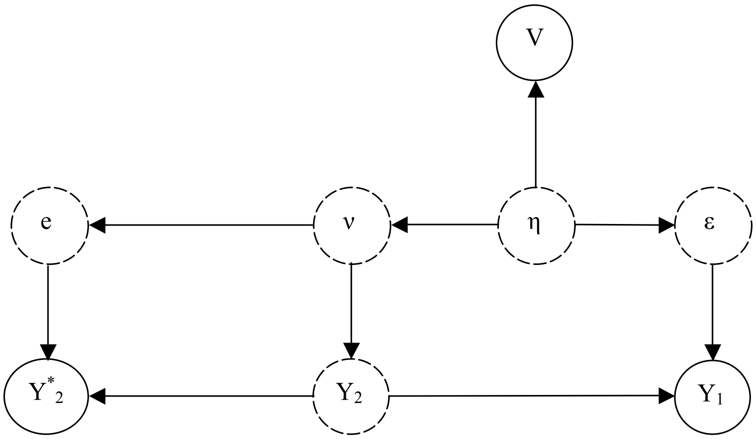

Figure 1. Causal diagram for control variable

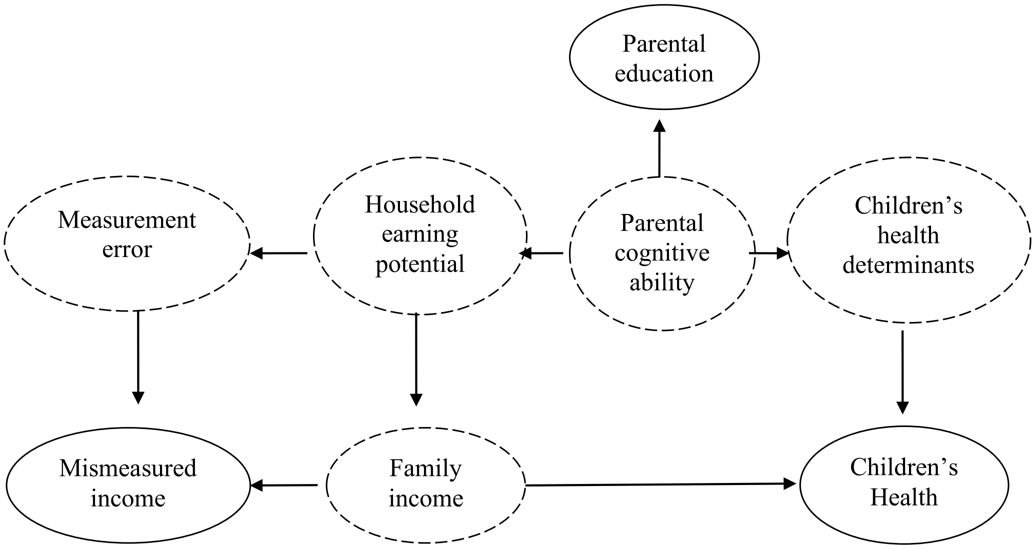
Parental education
Measurement error
Household earning potential
Parental cognitive ability
Children's health determinants
Mismeasured income
Family income
Children's Health

Figure 2. Effect of family income on children's health

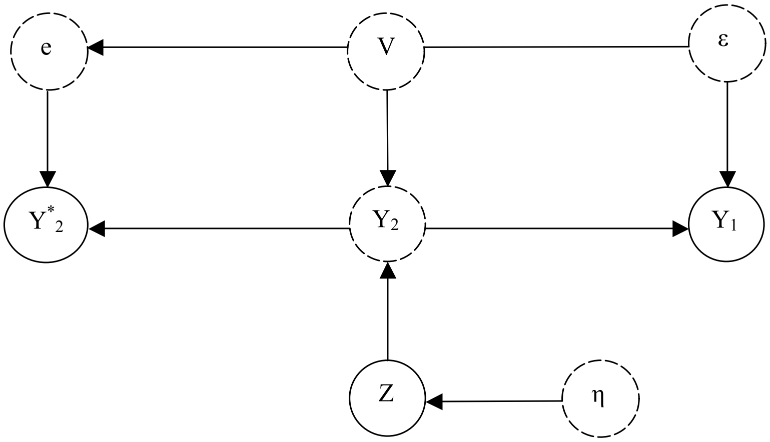

Figure 3. Causal diagram for instrumental variable

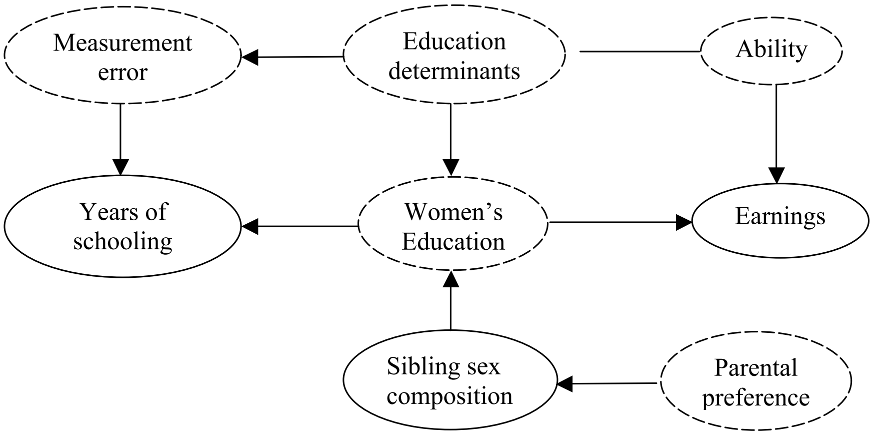

Figure 4. Effect of women's education on earnings

Figure G1: Design A, n=1,000 and s.d. = 0.5

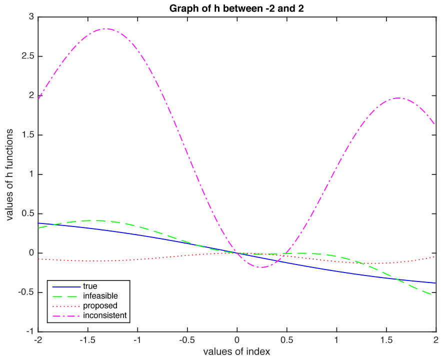

Figure G2: Design B, n=1,000 and s.d.= 0.5

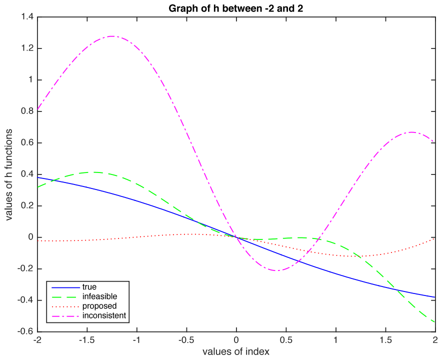

Figure G3: Design C, n=1,000 and s.d.= 0.5

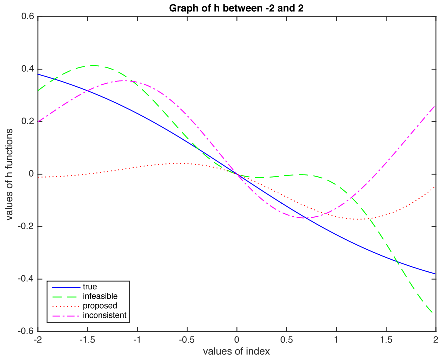

Figure G4: Design A, n=1,000 and s.d.= 1

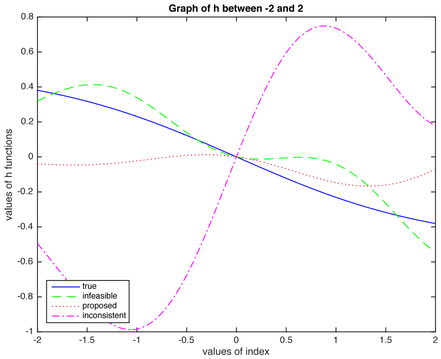

Figure G5: Design B, n=1,000 and s.d.= 1

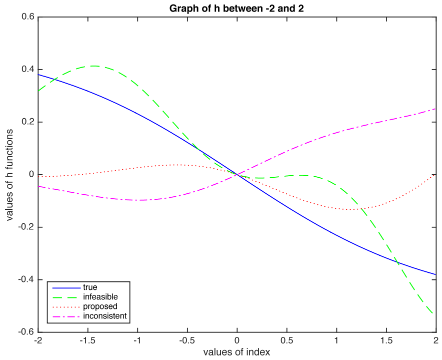

Figure G6: Design C, n=1,000 and s.d.= 1

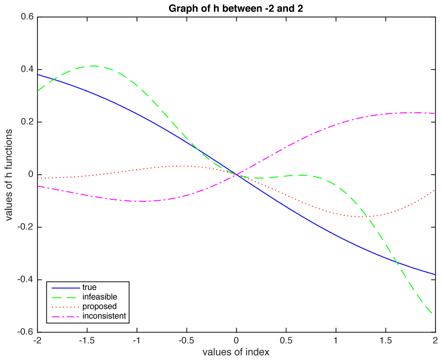

Figure G7: Design A, n=1,000 and s.d. = 1.5

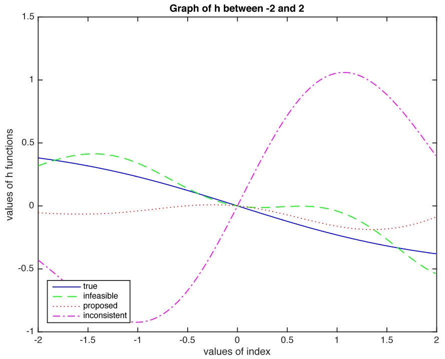

Figure G8: Design B, n=1,000 and s.d.= 1.5

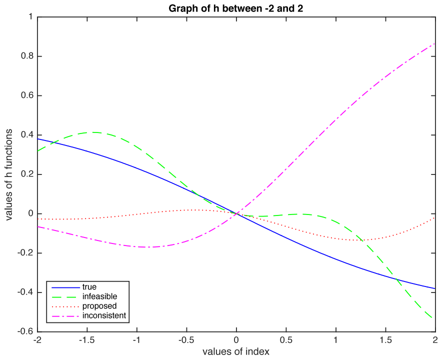

Figure G9: Design C, n=1,000 and s.d.= 1.5

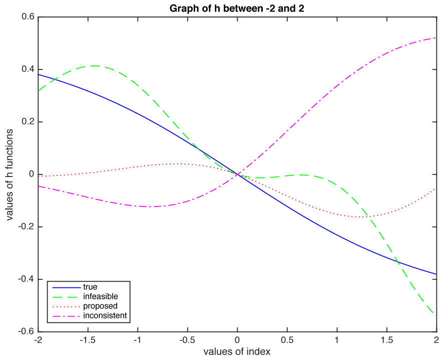

Figure G10: Design A, n=500 and s.d.= 1

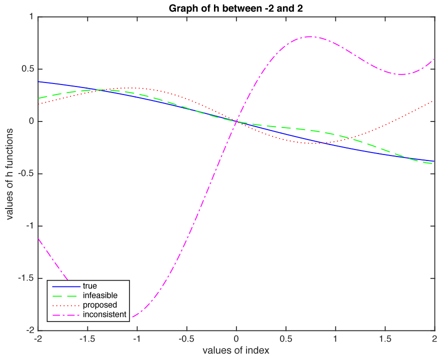

Figure G11: Design B, n=500 and s.d.= 1

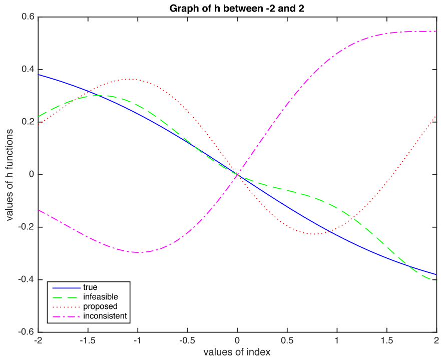

Figure G12: Design C, n=500 and s.d.= 1

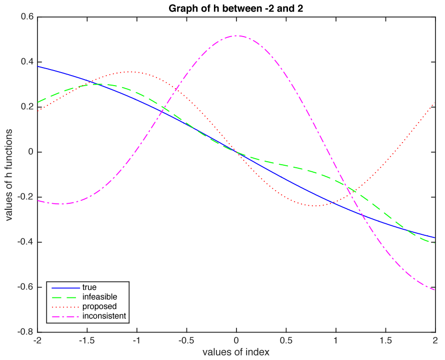

Table 1

Estimation of θ_0 in the partially linear model

Estimator	ME Structure	Design A Zero Mode	Design B Zero Mean	Design C Zero Median
Infeasible	Squared Bias	0.0034	0.0034	0.0034
	Variance	0.0055	0.0055	0.0055
	MSE	0.0089	0.0089	0.0089
Proposed	Squared Bias	0.0188	0.0262	0.0093
	Variance	0.4625	0.4499	0.4899
	MSE	0.4813	0.4761	0.4992
Inconsistent	Squared Bias	2.1440	2.1920	2.2200
	Variance	0.0017	0.0042	0.0022
	MSE	2.1457	2.1962	2.2222
s.d. of ME	0.5			

		Design A	Design B	Design C
Infeasible	Squared Bias	0.0034	0.0034	0.0034
	Variance	0.0055	0.0055	0.0055
	MSE	0.0089	0.0089	0.0089
Proposed	Squared Bias	0.0050	0.0246	0.0076
	Variance	0.4875	0.4588	0.5857
	MSE	0.4925	0.4834	0.5933
Inconsistent	Squared Bias	2.2050	2.2350	2.2460
	Variance	0.0005	0.0011	0.0006
	MSE	2.2055	2.2361	2.2466
s.d. of ME	1.0			

		Design A	Design B	Design C
Infeasible	Squared Bias	0.0034	0.0034	0.0034
	Variance	0.0055	0.0055	0.0055
	MSE	0.0089	0.0089	0.0089
Proposed	Squared Bias	0.0002	0.0234	0.0290
	Variance	0.5155	0.4878	0.5413
	MSE	0.5157	0.5112	0.5703
Inconsistent	Squared Bias	2.2220	2.2430	2.2490
	Variance	0.0002	0.0005	0.0002
	MSE	2.2222	2.2435	2.2492
s.d. of ME	1.5			

Table 2

Estimation of h_0 in the additively-separable model ($n = 500$)

Estimator	ME Structure	Design A	Design B	Design C
		Zero Mode	Zero Mean	Zero Median
Infeasible	ISB	0.0034	0.0034	0.0034
	IVAR	0.0881	0.0881	0.0881
	IMSE	0.0915	0.0915	0.0915
Proposed	ISB	0.0346	0.0434	0.0346
	IVAR	0.1499	0.1822	0.1354
	IMSE	0.1845	0.2256	0.1700
Inconsistent	ISB	2.07	0.8595	0.0732
	IVAR	53.13	10.93	6.51
	IMSE	55.20	11.79	6.58
s.d. of ME	0.5			

		Design A	Design B	Design C
Infeasible	ISB	0.0034	0.0034	0.0034
	IVAR	0.0881	0.0881	0.0881
	IMSE	0.0915	0.0915	0.0915
Proposed	ISB	0.0326	0.0355	0.0340
	IVAR	0.1690	0.1552	0.1372
	IMSE	0.2016	0.1907	0.1712
Inconsistent	ISB	1.76	0.3273	0.2002
	IVAR	112.30	12.50	9.59
	IMSE	114.06	12.83	9.79
s.d. of ME	1.0			

		Design A	Design B	Design C
Infeasible	ISB	0.0034	0.0034	0.0034
	IVAR	0.0881	0.0881	0.0881
	IMSE	0.0915	0.0915	0.0915
Proposed	ISB	0.0356	0.0369	0.0342
	IVAR	0.2160	0.1491	0.1244
	IMSE	0.2516	0.1860	0.1587
Inconsistent	ISB	4.08	0.7860	0.1974
	IVAR	146.30	22.85	13.85
	IMSE	150.38	23.64	14.05
s.d. of ME	1.5			

Table 3

Estimation of h_0 in the additively-separable model ($n = 1,000$)

Estimator	ME Structure	Design A	Design B	Design C
		Zero Mode	Zero Mean	Zero Median
Infeasible	ISB	0.0102	0.0102	0.0102
	IVAR	0.0459	0.0459	0.0459
	IMSE	0.0561	0.0561	0.0561
Proposed	ISB	0.0299	0.0300	0.0300
	IVAR	0.0903	0.0909	0.0859
	IMSE	0.1202	0.1209	0.1159
Inconsistent	ISB	2.93	0.4569	0.0441
	IVAR	44.50	6.64	2.73
	IMSE	47.43	7.10	2.77
s.d. of ME	0.5			

		Design A	Design B	Design C
Infeasible	ISB	0.0102	0.0102	0.0102
	IVAR	0.0459	0.0459	0.0459
	IMSE	0.0561	0.0561	0.0561
Proposed	ISB	0.0304	0.0288	0.0304
	IVAR	0.0890	0.0914	0.0858
	IMSE	0.1193	0.1202	0.1162
Inconsistent	ISB	0.7639	0.1311	0.1386
	IVAR	28.97	3.60	2.82
	IMSE	29.73	3.73	2.96
s.d. of ME	1.0			

		Design A	Design B	Design C
Infeasible	ISB	0.0102	0.0102	0.0102
	IVAR	0.0459	0.0459	0.0459
	IMSE	0.0561	0.0561	0.0561
Proposed	ISB	0.0299	0.0300	0.0300
	IVAR	0.0856	0.0852	0.0860
	IMSE	0.1155	0.1152	0.1160
Inconsistent	ISB	0.9545	0.3625	0.2290
	IVAR	26.86	5.85	6.29
	IMSE	27.81	6.21	6.52
s.d. of ME	1.5			