A Control Function Approach in Sieve Two-Step M-Estimation of Binary Response Models with Endogenous Explanatory Variables*

Wei Lin†

November 27th, 2017

Abstract

This paper proposes a sieve two-step M-estimation via control function approaches for a special case of triangular system–a binary response model with both continuous and dummy endogenous explanatory variables. In a first step, residuals are obtained from sieve estimation of reduced-form equation for continuous endogenous explanatory variables. In a second step, the residuals (control functions) are plugged into error terms in the binary outcome equation and a reduced-form equation for the dummy endogenous explanatory variable. Unknown functional forms are assumed for the residuals, and these two equations are jointly estimated by sieve maximum likelihood estimation. In order to identify causal effects of interest, estimators from both steps are plugged into a third-step method of moments estimation of functionals called average partial effects (APEs), defined as marginal effects of an average structural function. Under the framework of Hahn, Liao, and Ridder (2015), I establish $\sqrt{n}$ asymptotic normality for APEs and provide consistent estimators for asymptotic variances. Due to their numerical equivalence result, I show practical inference for the asymptotic variance, using a parametric model with the number of terms in basis functions fixed at a given sample size. Thus, the proposed sieve two-step M-estimation methods combined with control function approaches are flexible, robust to misspecification, easy to implement, computationally simple, and feasible for conducting practical inference. A simulation study demonstrates these advantages.

*I would like to thank Jeffrey M. Wooldridge, Kyoo il Kim, Tim Vogelsang, and Peter Schmidt for their continuing support and guidance. I have also benefited from discussions with participants at Midwest Econometrics Group meeting in 2016 and the seminars at Michigan State University. The usual disclaimer applies.

†Center for Real Estate, Massachusetts Institute of Technology, Cambridge, MA 02139 , United States. linwei1@mit.edu. The latest version is available at http://weilinmetrics.weebly.com.

1 Introduction

It is notably challenging to deal with endogeneity in limited dependent variable models, especially when the endogenous explanatory variables (EEVs, for short) can have differing attributes, with some being continuous and some being discrete. In principle, one can make enough distributional assumptions so that maximum likelihood estimation (MLE) is possible. Depending on how complicated the model is, MLE can be computationally demanding (although simulation methods of estimation can make complicated models feasible). In any case, the estimates of parameters and other quantities of interest, such as marginal effects, can be sensitive to parametric assumptions.

This paper focuses on a semiparametric specification and estimation of a special case of a triangular system. Namely, the response variable is binary, there is a single binary EEV, and potentially many continuous EEVs. The model can be applied to endogenous switching with a binary response and multiple continuous EEVs – as could happen, for example, in a production environment with two production regimes and several continuous inputs. Or, the multiple continuous EEVs can be prices, as in Petrin and Train (2010). Because of the semiparametric setting, the approach provides a more flexible and robust way to conduct estimation and inference for causal marginal effects.

This paper extends the two-step parametric approach in Lin and Wooldridge (2015a). There, joint normality and linearity assumptions are used to obtain a simple two-step control function estimator, where reduced forms for all continuous EEVs are first estimated to obtain reduced-form residuals (control functions). In a second step, a bivariate probit model is used which includes the control functions and any functions of the EEVs. The method is computationally simple and $\sqrt{n}$-asymptotically normal, provided the parametric assumptions are satisfied. In addition, average marginal effects are identified and easily estimated. Bootstrapping the full procedure provides straightforward inference; or, the delta method can be used to obtain proper standard errors.

One can extend the Lin and Wooldridge (2015a) in several directions because there are multiple unobservables and several linear functional forms. It is essentially impossible to allow a fully nonparametric analysis because one loses identification with discrete EEVs (Chesher, 2003, for example). This paper relaxes assumptions on the reduced forms for the continuous EEVs and the relationships between the structural errors and the control functions. I maintain linearity in the structural equation, and I also maintain a conditional bivariate normality assumption on the errors in the binary response equations. While one can

debate these choices, at a minimum they extend the analysis in Lin and Wooldridge (2015a) to allow for the kind of nonparametric reduced forms for the continuous EEVs in Blundell and Powell (2004). Unlike the Blundell and Powell framework, the analysis here allows for a binary EEV.

Unlike Blundell and Powell (2004), which uses a matching approach in the second step estimation, and Rothe (2009), which applies the Klein and Spady (1993) semiparametric maximum likelihood estimator, here I use sieve estimation for the nonparametric components in both steps. The method of sieves, formally introduced by Grenander (1981), allows for unknown functions that lie in infinite-dimensional parameter spaces to be approximated by finite dimensional spaces of growing dimensions. The growth in the dimensionality is linked to increase in sample size to guarantee consistency. As a practical matter, sieve estimation is appealing because, for a given sample size, it is flexible parametric estimation. The difficulty lies in theoretical derivations of the asymptotic properties of estimators with a growing parameter set. To this end, I apply general results in Chen (2007) and Chen, Linton, and Van Keilegom (2003) for consistency and rates of convergence for sieve two-step estimation, where the first step consists of sieve least squares and the second step sieve MLE. A third step estimation is used to estimate the average partial effects (APEs) based on the average structural function (ASF) of Blundell and Powell (2004). Due to the numerical equivalence result shown in Hahn, Liao, and Ridder (2015), we can treat the sieve two-step M-estimation as if it is a standard two-step parametric problem, and thus easily conduct practical inference on APEs using delta methods. It seems very likely that bootstrapping can also be justified.

In addition to full parametric approaches and the semiparametric sieve approach proposed here, there are other modeling strategies and estimation methods that have been proposed – some only for the special case of a single binary EEV and others that also allow discrete EEVs. Recently, Han and Vytlacil (2015) advance identification analyses for unknown marginal distributions in a bivariate response model using the copula approach. A restrictive feature – other than that it applies to the special case of a single binary EEV that appears additively in the structural equation – is that the copula function is required to be parametric. Here, I restrict the marginals but, in effect, allow general correlation among the reduced form errors of the continuous and binary EEVs and the structural error.

For handling dummy EEVs, Terza, Basu, and Rathouz (2008) and Wooldridge (2014) argue for plugging in a residual or generalized residual obtained from a reduced form probit. While simple, these CF methods are controversial because they rely on nonstandard parametric assumptions to achieve identification, and

therefore consistency[1]. The CF approach is much easier to justify for the continuous EEVs, and that is motivation for the current paper.

Lewbel (2000) proposes a different semiparametric approach for estimating the parameters of a binary response, and he allows both continuous and discrete EEVs. Unfortunately, Lewbel's approach is limited by its requirement of a "special regressor," which is an observed covariate assumed to appear in the structural equation but not in the reduced forms for the EEVs. Except when the special regressor is randomized, this is a strong requirement. Moreover, as argued in Lin and Wooldridge (2015b), Lewbel's special regressor approach currently does not allow one to uncover meaningful partial effects.

To summarize, the approach here builds off the parametric approach in Lin and Wooldridge (2015a), relaxing linearity assumptions in both reduced form and control function relationships. It allows the continuous and binary EEVs to appear in complicated nonlinear ways in a linear index – for example, squares and interactions are allowed. I also show how to estimate and perform inference on average partial effects, something that is often ignored in the literature on semiparametric analysis.

The remainder of the paper is organized as follows. Section 2 presents the model specifications and sufficient identification conditions, starting with the parametric case to provide background. Section 3 elaborates on the sieve two-step M-estimation problem. Section 4 presents asymptotics properties of the sieve two-step M-estimator, including consistency and asymptotic normality of nonlinear functionals. Section 5 shows a practical inference for the APEs resulting from the semiparametric two-step procedure. In section 6, I undertake a Monte Carlo study showing how the sieve approximation works compared with other choices. Section 7 concludes and proposes directions for future research.

[1]As pointed out in Wooldridge (2014), it is natural to use plugging in of generalized residual under nonstandard assumption as a means to test for endogeneity of dummy variable in binary outcome equation, rather than to estimate the coefficient.

2 Model specification

2.1 The parametric model

As a starting point, consider a simple parametric triangular system for a binary response y_1 with a continuous EEV y_2 and a dummy EEV y_3 of the following form:

$$y_1 = 1\left[\mathbf{x}_1\boldsymbol{\beta}_o + u_1 \geq 0\right], \tag{1a}$$

$$y_2 = \mathbf{z}\boldsymbol{\delta}_o + v_2, \tag{1b}$$

$$y_3 = 1\left[\mathbf{z}\boldsymbol{\gamma}_o + u_3 \geq 0\right], \tag{1c}$$

where equation $(1a)$ is the structural equation that bears economic meanings, and equations $(1b)$ and $(1c)$ are the reduced forms for EEVs y_2 and y_3, respectively. On the right-hand side, $\mathbf{x}_1$ is a $1 \times K_1$ vector of functions of $(\mathbf{z}_1, y_2, y_3)$, so that $\mathbf{x}_1 \equiv f_1(\mathbf{z}_1, y_2, y_3)$. The leading case is $\mathbf{x}_1 \equiv (1, \mathbf{z}_1, y_2, y_3)$, where the first element of $\mathbf{x}_1$ is unity. In general, $\mathbf{x}_1$ could include quadratics, interactions, logarithms of $(\mathbf{z}_1, y_2, y_3)$. Similarly, $\mathbf{z} \equiv f(\mathbf{z_1}, \mathbf{z_2})$ is a $1 \times L$ vector of functions of all exogenous variables, with $\mathbf{z} \equiv (1, \mathbf{z_1}, \mathbf{z_2})$ being the leading case, where $\mathbf{z_1}$ is the vector of the included exogenous variables and $\mathbf{z_2}$ the excluded exogenous variables. Without changing the estimation procedure, control function approaches can easily accommodate flexible functions in $\mathbf{x}_1$ and $\mathbf{z}$.

In order to apply control function approaches, a set of sufficient conditions for identification in this parametric setting is stated as follows.

ASSUMPTION 2.1 (Joint Normality)

$$D\left(\begin{array}{c|c} u_1 & \\ v_2 & \mathbf{z} \\ u_3 & \end{array}\right) = D\begin{pmatrix} u_1 \\ v_2 \\ u_3 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_o & \tau_o \\ \rho_o & \sigma_o^2 & \omega_o \\ \tau_o & \omega_o & 1 \end{pmatrix}\right],$$

where ρ_o, ω_o, and τ_o are the covariances, and σ_o^2 is the variance of v_2. For simplicity, the variances of u_1, u_3 are assumed to be at unity.

Following, for example, Blundell and Smith (1994), Heckman (1978), and Rivers and Vuong (1988), Assumption 2.1 starts with a joint normality of all unobservables. With the statistical independence of the exogenous variables $\mathbf{z}$, Assumption 2.1 takes a stronger form of an exogeneity assumption. Finally, in As-

sumption 2.1, endogeneity is assumed to come from an omitted variable problem, where the reduced-form errors are assumed to be correlated with the structural error.

As a proxy for the endogeneity of y_2, error term v_2 needs to be partialed out from the error terms in the two equations for discrete outcomes, y_1 and y_3:

$$u_1 = \frac{\rho_o}{\sigma_o} v_2 + v_1,$$
$$u_3 = \frac{\omega_o}{\sigma_o} v_2 + v_3.$$

The remaining error terms, vector (v_1, v_3), by the trivariate normality in Assumption 2.1, has a zero mean, bivariate normal distribution, and is independent of $(v_2, \mathbf{z})$, as follows:

$$D\left(\begin{array}{c} v_1 \\ v_3 \end{array} \middle| \, \mathbf{z}, v_2 \right) = D\left(\begin{array}{c} v_1 \\ v_3 \end{array} \right) \sim N\left[\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} 1-\rho_o^2 & \tau_o - \rho_o \omega_o \\ \tau_o - \rho_o \omega_o & 1-\omega_o^2 \end{array} \right) \right].$$

With this projection, the set of equations in model (1) is reduced to

$$y_1 = 1\left[\mathbf{x}_1 \boldsymbol{\beta}_o + \frac{\rho_o}{\sigma_o} v_2 + v_1 \geq 0 \right], \tag{2a}$$
$$y_3 = 1\left[\mathbf{z} \boldsymbol{\gamma}_o + \frac{\omega_o}{\sigma_o} v_2 + v_3 \geq 0 \right]. \tag{2b}$$

Because of the nonlinearity, estimating parameters in the above two equations calls for a joint MLE procedure as follows. Additionally, in practice, identifying based on nonlinearity results in poor statistical properties, so an auxiliary order condition of $L \geq K_1$ is required.

PROCEDURE 2.1 (CF under Joint Normality) *Under Assumption 2.1 and the order condition stated above, the following parametric two-step biprobit procedure is consistent for parameters of interest, namely, $\boldsymbol{\beta}_o$ and $\frac{\rho_o}{\sigma_o}$ in equation (2a), the structural equation for binary outcome y_1:*

(a) Obtain $\widehat{\delta}$ from OLS estimation of the reduced form of y_2. Obtain the residual $\widehat{v}_2$, defined as $\widehat{v}_2 = y_2 - \mathbf{z}\widehat{\boldsymbol{\delta}}$.

(b) Using $\widehat{v}_2$ in place of v_2, estimate the bivariate probit model, as specified in equations (2a) and (2b), by joint MLE.

Following these assumptions, consistency resulting from Procedure 2.1 is straightforward. In addition, implementation of Procedure 2.1 is fairly easy and routinely available in standard statistical packages. However, as mentioned before, the joint normality assumption on the error terms is sufficient but not necessary.

With the robust nature of control function approaches, Procedure 2.1 can go through under a weaker set of assumptions. In particular, when there are many continuous EEVs, say a vector of $\mathbf{y_2}$, the multivariate normality assumption on the corresponding error terms $\mathbf{v_2}$ can be dropped. Instead, the conditional distribution of (u_1, u_3) given $\mathbf{v_2}$ is assumed to follow a zero mean, bivariate normal distribution. With the weaker conditional bivariate normality assumption, rather than making the unit variance assumption for the distribution of the structural errors (u_1, u_3), the variances for remaining errors (v_1, v_3) are assumed to be at unity, for simplicity of estimation, as formally stated in the following.

ASSUMPTION 2.1′ (Conditional Normality)

(a) *The conditional expectations of* (u_1, u_3) *given* $\mathbf{v_2}$ *are linear in* $\mathbf{v_2}$*, so* (u_1, u_3) *can be decomposed as their respective conditional expectations plus error terms,*

$$u_1 = \boldsymbol{\lambda_o}\mathbf{v_2} + v_1,$$

$$u_3 = \boldsymbol{\eta_o}\mathbf{v_2} + v_3.$$

(b) *The remaining error terms* (v_1, v_3) *are independent of* $\mathbf{z}$ *and* $\mathbf{v_2}$ *and follow a zero mean, bivariate normal distribution with variances one and covariance* ρ_o*,*

$$D\left(\begin{array}{c} v_1 \\ v_3 \end{array}\middle|\, \mathbf{z},\mathbf{v_2}\right) = D\begin{pmatrix} v_1 \\ v_3 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_o \\ \rho_o & 1 \end{pmatrix}\right].$$

(c) *The reduced-form errors,* $\mathbf{v_2}$*, are mean independent of* z*, that is,* $E(\mathbf{v_2}|\mathbf{z}) = 0$*.*

More generally, maintaining the conditional bivariate normality has mainly two advantages. First, by assuming a bivariate normality for the remaining errors, identification conditions for single index models for binary responses are automatically satisfied. Specifically, the structural errors (u_1, u_3) have a large support because of the bivariate normality in their components (v_1, v_3), and they are mean zero because intercepts have been absorbed in covariates $\mathbf{x}_1$. Second, uncovering marginal effects of interest is easier because routines in common statistical software packages are available for conducting joint estimation and extrapolating to counterfactual outcomes under a bivariate normality assumption. Relaxing this bivariate normality assumption for the remaining errors is mainly theoretical, and is therefore beyond the scope of this paper.

Although weaker, Assumption 2.1′ is still restrictive along two dimensions. First, it is difficult to maintain the linearity assumption on the conditional expectations of (u_1, u_3) given $\mathbf{v_2}$. With the remaining additive errors (v_1, v_3) assumed to be independent of $\mathbf{v_2}$, the linearity assumption eliminates $\mathbf{v_2}$ in any other form or channel in the joint distribution of (u_1, u_3). Second, the mean independence assumption between $\mathbf{z}$ and $\mathbf{v_2}$, combined with linear functional forms of $\mathbf{z}$, requires the conditional means of $\mathbf{y_2}$ given $\mathbf{z}$ to be linear, which lacks flexibility.

To relax these two restrictions, a semiparametric specification is introduced in the following section. Drawing on the weaker conditional bivariate normality in Assumption 2.1′, this semiparametric approach adds flexibility and robustness by introducing arbitrary unknown functions along the above two dimensions. As we are inherently agnostic about distributions and functional forms in economic theories, it is desirable to impose as few assumptions as possible.

2.2 The semiparametric specification

The semiparametric specification in this section relaxes restrictions along the following two dimensions. First, as in Blundell and Powell (2004), I allow the reduced forms for endogenous EEVs $\mathbf{y_2}$ to be fully nonparametric, composed of unknown functions $h_o(\mathbf{z})$ and additive errors $\mathbf{v_2}$:

$$\mathbf{y_2} = h_o(\mathbf{z}) + \mathbf{v_2}, \tag{3}$$

where $h_o(\mathbf{z}) \equiv E(\mathbf{y_2}|\mathbf{z})$ are the unspecified conditional means of $\mathbf{y_2}$ given $\mathbf{z}$.

Second, I let $\mathbf{v_2}$ enter the structural errors u_1 and u_3 in a partial linear fashion:

$$u_1 = m_o(\mathbf{v_2}) + v_1, \tag{4}$$

$$u_3 = q_o(\mathbf{v_2}) + v_3, \tag{5}$$

where $m_o(\cdot)$ and $q_o(\cdot)$ are unknown functions with zero means. Partialing out from structural errors (u_1, u_3) all the endogeneity of the continuous EEVs $\mathbf{y_2}$, these two arbitrary functions of $\mathbf{v_2}$ play the role of control functions. The remaining errors (v_1, v_3) are assumed to be bivariate normal.

With relaxations along the above two dimensions, a semiparametric triangular model is obtained as

follows :

$$y_1 = 1\left[\mathbf{x}_1\boldsymbol{\beta}_o + m_o(\mathbf{v_2}) + v_1 \geq 0\right], \tag{6a}$$

$$\mathbf{y_2} = h_o(\mathbf{z}) + \mathbf{v_2}, \tag{6b}$$

$$y_3 = 1\left[\mathbf{z}\boldsymbol{\gamma}_o + q_o(\mathbf{v_2}) + v_3 \geq 0\right], \tag{6c}$$

where $(m_o(\cdot), h_o(\cdot), q_o(\cdot))$ are the infinite-dimensional unknown parameters, and $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \rho)$ are the finite-dimensional unknown parameters. As in the parametric case, a set of semiparametric identification conditions based on the weaker bivariate normality assumption for the conditional distribution of (u_1, u_3) given $\mathbf{v_2}$ is stated in Assumption 2.2.

ASSUMPTION 2.2 (Semiparametric Conditional Normality)

(a) *The conditional expectations of* (u_1, u_3) *given* $\mathbf{v_2}$ *are of arbitrary functional form, and* (u_1, u_3) *are decomposed as their respective conditional expectations plus error terms,*

$$u_1 = m_o(\mathbf{v_2}) + v_1,$$

$$u_3 = q_o(\mathbf{v_2}) + v_3.$$

(b) *The remaining error terms* (v_1, v_3) *are independent of* $\mathbf{z}$ *and* $\mathbf{v_2}$ *and follow a zero mean, bivariate normal distribution with variances one and covariance* ρ_o*,*

$$D\left(\begin{array}{c} v_1 \\ v_3 \end{array}\middle|\, \mathbf{z},\mathbf{v_2}\right) = D\begin{pmatrix} v_1 \\ v_3 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_o \\ \rho_o & 1 \end{pmatrix}\right].$$

(c) *The reduced-form errors,* $\mathbf{v_2}$*, are mean independent of* z*, that is,* $E(\mathbf{v_2}|\mathbf{z}) = 0$*.*

By allowing the conditional expectations of (u_1, u_3) given $\mathbf{v_2}$ to take arbitrary functional forms, it is easier to assume that all possible forms of dependence between (u_1, u_3) and $\mathbf{v_2}$ have been accounted for, and thus $\mathbf{v_2}$ does not enter the joint distribution of (u_1, u_3) in any other channel. Similarly, with conditional expectations of $\mathbf{y_2}$ given $\mathbf{v_2}$ taking arbitrary functional forms, it is easier to justify the mean independence of $\mathbf{v_2}$ in the reduced forms of $\mathbf{y_2}$. In the meantime, with these infinite dimensional unknown functions $(m_o(\cdot), h_o(\cdot), q_o(\cdot))$, rather than relaxing into fully nonparametric models, it is crucial to maintain some parametric structure such as single indexes $\mathbf{x}_1\boldsymbol{\beta}$ in order to reduce dimensionality and to achieve greater

estimation precision.

To conduct estimation, a semiparametric control function procedure is suggested in Procedure 2.2. Unlike the full parametric case, optimization over infinite-dimensional parameter space can be problematic. To get around this, methods of sieves are incorporated in this procedure so that the optimization can be conducted over finite dimensional spaces, called sieve spaces. The dimension of the sieve spaces grows with sample sizes so that in the limit it tends to the true parameter space.

PROCEDURE 2.2 (CF under Semiparametric Conditional Normality)

Heuristically, a sieve two-step M-estimator can be implemented as follows,

(a) Obtain the reduced form residuals $\widehat{\mathbf{v}}_{\mathbf{2}}$ *from sieve least squares estimation of the reduced forms of* $\mathbf{y_2}$*.*

(b) Using $\widehat{\mathbf{v}}_{\mathbf{2}}$ *in place of* $\mathbf{v_2}$*, estimate the semiparametric bivariate probit model, as specified in equations (6a) and (6c), by sieve maximum likelihood estimation.*

To formally establish the consistency and asymptotic normality result for Procedure 2.2, the next section turns to theoretically defining the population and sample optimization problems for the sieve two-step M-estimator.

3 Estimation

As a building block for deriving asymptotic properties in the next section, this section studies identification and estimation of the semiparametric model in (6), using sieve two-step M-estimation.

3.1 The true parameter spaces

Define the true parameter in the first step as $h_o \equiv h_o(\cdot)$, and define the parameter space in the first step as $\mathcal{H}$, where $h_o(\cdot) \in \mathcal{H}$. Let h be a generic element in the parameter space $\mathcal{H}$, formally $h \in \mathcal{H}$. Assumption 3.1 describes the sampling process, and Assumption 3.2 specifies the identification of the first-step true parameter h_o as a solution to a population optimization problem.

ASSUMPTION 3.1 (Sampling) *Let a random vector* $\mathbf{w}_{i1} = (\mathbf{y_{i2}}, \mathbf{z}_i), i = 1, ..., n$ *denote the data in the first step, and a random vector* $\mathbf{w}_{i2} = (y_{i1}, \mathbf{y_{i2}}, y_{i3}, \mathbf{z}_i), i = 1, ..., n$ *denote the data in the second step. The random vectors are assumed to be independent and identically distributed (i.i.d.).*

ASSUMPTION 3.2 (True Parameters in the First Step) *The true parameter h_o in the first step is identified as a unique solution to maximize a population least squares problem over the infinite dimensional parameter space $\mathcal{H}$, formally:*

$$h_o = \arg\sup_{h \in \mathcal{H}} - E\left[\left(\mathbf{y_{i2}} - h(\mathbf{z}_i)\right)^2\right].$$

Define the true parameters in the second step as $g_o \equiv (\boldsymbol{\theta}_o, m_o(\cdot), q_o(\cdot))$ and define the parameter space in the second step as $\mathcal{G}$, where $g_o \in \mathcal{G}$. Let g be a generic element in $\mathcal{G}$. Further, $\mathcal{G}$ is a tensor product of a finite parameter space Θ and two infinite functional spaces $\mathcal{M}$ and $\mathcal{Q}$. The finite parameter space Θ contains the true finite parameters $\boldsymbol{\theta}_o \equiv (\boldsymbol{\beta}_o, \rho_o, \boldsymbol{\gamma}_o)$, formally $\boldsymbol{\theta}_o \in \Theta$, with $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \rho, \boldsymbol{\gamma})$ being a generic point in Θ. The infinite parameter spaces $\mathcal{M}$ and $\mathcal{Q}$ contain the true infinite parameters $m_o(\cdot)$ and $q_o(\cdot)$, formally $m_o(\cdot) \in \mathcal{M}$ and $q_o(\cdot) \in \mathcal{Q}$, with m being a generic element in $\mathcal{M}$ and q a generic element in $\mathcal{Q}$. Assumption 3.3 specifies the identification of the second-step true parameter g_o as a solution to a population optimization problem.

ASSUMPTION 3.3 (True Parameters in the Second Step) *Using the true parameter h_o from the first step, the true parameter g_o in the second step is identified as a unique solution to a population maximum likelihood estimation problem over the infinite dimensional parameter space $\mathcal{G}$:*

$$g_o = \arg\sup_{g \in \Theta \times \mathcal{M} \times \mathcal{Q}} E\left[\log \Phi_{v_1, v_3}\left(d_{i1}\left[\mathbf{x}_{i1}\boldsymbol{\beta} + m(\mathbf{y_{i2}} - h_o(\mathbf{z}_i))\right], d_{i2}\left[\mathbf{z}_i\boldsymbol{\gamma} + q(\mathbf{y_{i2}} - h_o(\mathbf{z}_i))\right], d_{i1}d_{i2}\rho\right)\right],$$

where $\Phi_{v_1, v_3}(\cdot, \cdot, \cdot)$ is the cumulative distribution function (CDF) of (v_1, v_3). The CDF $\Phi_{v_1, v_3}(\cdot, \cdot, \cdot)$ takes both of the two indexes, as well as the covariance ρ, of the bivariate normal distribution of (v_1, v_3) as its arguments. Symbol $\times$ is the tensor product. Indicators $d_{i1} \equiv 2y_{i1} - 1$ and $d_{i2} \equiv 2y_{i3} - 1$ are transformed from the binary outcomes (y_{i1}, y_{i3}) to switch the direction of integration in the CDF

3.2 Sieve spaces

The theoretical difficulty of uncovering the true parameters of infinite dimension is caused by the ill-posed problem. According to a survey article by Chen (2007), with infinite parameter space, there can be no unique solution to the optimization problem. Or, there can exist confounding solutions where the values of the criterion functions are indistinguishable. In either case, resulting asymptotic properties of the estimators are poor, and the estimation of the system suffers numerical instability. In Procedure 2.2, because of the

control functions obtained in the first-stage semiparametric estimation, both the conditioning variables and arguments are of infinite dimension in the second step, leading to a greater chance of ill-posed problem and numerical instability. To mitigate this problem, one can approximate unknown functions by employing the method of sieves. With the dimensionality tending to infinity as the sample size grows, the method of sieves optimizes over a sequence of finite-dimensional sieve spaces.

The method of sieves is chosen over other common nonparametric techniques, especially the kernel method, mainly due to the following three reasons. First, unlike the kernel method, it is easier to impose shape restrictions on the functional forms with the method of sieves, leading to a higher degree of flexibility. Because of this flexibility, error terms in model (6) can be decomposed in a partial linear fashion. Second, the method of sieves is relatively easier to implement in standard softwares by manually adding sieve terms as additional regressors. In implementation, one can conduct sieve estimation for model (6) by adding some sieve terms of the first-stage residuals to the simple parametric two-step biprobit procedure. Lastly, while the asymptotics of the kernel method need to be derived on a case-by-case basis, the method of sieves has a unified theory for asymptotics. As model (6) fits into the sieve two-step M-estimation framework provided by Hahn, Liao, and Ridder (2015), I use their result by verifying their conditions.

Define a sieve space $\mathcal{H}_n$ for the true parameter h_o in the first step at sample size n as

$$\mathcal{H}_n \equiv \left\{ h(\cdot) = p_1^{k(n)}(\cdot)'\delta_h : \delta_h \in \mathbb{R}^{k(n)} \right\},$$

where the real-valued function $h(\cdot)$ is a generic element in $\mathcal{H}_n$ and

$$p_1^{k(n)}(\cdot) \equiv (p_{1,1}(\cdot), ..., p_{1,k(n)}(\cdot))' \tag{7}$$

is a $k(n) \times 1$ vector of basis functions. The number of basis functions characterizes the complexity of the sieve space. Using this set of notation, as a nondecreasing sequence of sieve spaces approximating $\mathcal{H}$ as n gets larger, a sequence of pseudo-true parameters converges to the true parameters, as formally stated in Assumption 3.4.

ASSUMPTION 3.4 (Sieve Spaces for the First Step) *A nondecreasing sequence of sieve spaces approximates $\mathcal{H}$ as n gets larger*

$$\mathcal{H}_n \subseteq \mathcal{H}_{n+1} \subseteq ... \subseteq \mathcal{H},$$

and there exists a sequence $\pi_{h,n}h_o$ such that $d_h(h_o, \pi_n h_o) \to 0$ as $n \to \infty$, where $d_h(\cdot,\cdot)$ denotes a

(pseudo) metric on $\mathcal{H}$, $\pi_{h,n}$ is a projection mapping from $\mathcal{H}$ to $\mathcal{H}_n$.

Let $\mathcal{G}_n$ denote the sieve space at sample size n for the true parameter $(\boldsymbol{\theta}_o, m_o(\cdot), q_o(\cdot))$ in the second step, which is a tensor product of finite parameter space Θ and sieve spaces $\mathcal{M}_n$ and $\mathcal{Q}_n$,

$$\begin{aligned}\mathcal{G}_n &\equiv \Theta \times \mathcal{M}_n \times \mathcal{Q}_n \\ &\equiv \{(\boldsymbol{\beta}, \rho, \boldsymbol{\gamma})\} \\ &\times \{m(\cdot) = p_2^{l(n)}(\cdot)'\delta_m : \delta_m \in \mathbb{R}^{l(n)}\} \\ &\times \{q(\cdot) = p_2^{l(n)}(\cdot)'\delta_q : \delta_q \in \mathbb{R}^{l(n)}\}\end{aligned}$$

where the real-valued function $m(\cdot)$ denotes a generic element in $\mathcal{M}_n$, the real-valued function $q(\cdot)$ denotes a generic element in $\mathcal{Q}_n$, and

$$p_2^{l(n)}(\cdot) \equiv (p_{2,1}(\cdot), ..., p_{2,l(n)}(\cdot))' \tag{8}$$

denotes a $l(n) \times 1$ vector of basis functions. Let d_θ denote the dimension of the finite component $\boldsymbol{\theta}$ in the second step. The overall dimension in the second step is $k_2(n) = \dim(\mathcal{G}_n) = d_\theta + 2l(n)$. Using this set of notation, as a nondecreasing sequence of sieve spaces approximating $\mathcal{G}$ as n gets larger, a sequence of pseudo-true parameters converges to the true parameters, as formally stated in Assumption 3.5.

ASSUMPTION 3.5 (Sieve Spaces for the Second Step) *A nondecreasing sequence of sieve spaces approximates $\mathcal{G}$ as n gets larger*

$$\mathcal{G}_n \subseteq \mathcal{G}_{n+1} \subseteq ... \subseteq \mathcal{G},$$

and there exists a sequence $\pi_{g,n} g_o$ such that $d_g(g_o, \pi_{g,n} g_o) \to 0$ as $n \to \infty$, where $d_g(\cdot, \cdot)$ denotes a (pseudo) metric on $\mathcal{G}$, $\pi_{g,n}$ is a projection mapping from $\mathcal{G}$ to $\mathcal{G}_n$.

3.3 The sieve estimators

With the sieve spaces and pseudo-true parameters defined above, the sieve two-step M-estimator is formally defined by the following finite sample optimization problems in (9) and (11).

To estimate the infinite-dimensional true parameter h_o in the first step, at a given sample size n, the sample analog of the population least squares over a sieve space $\mathcal{H}_n$ is defined by

$$\widehat{h}_n = \arg\max_{h \in \mathcal{H}_n} \frac{-1}{n} \sum_{i=1}^{n} \left(\mathbf{y_{i2}} - h(\mathbf{z}_i)\right)^2. \tag{9}$$

The resulting sieve least squares estimator $\widehat{h}_n(\cdot)$ has a closed-form solution

$$\widehat{h}_n(\cdot) = p_1^{k(n)}(\cdot)'(P_1'P_1)^{-} \sum_{i=1}^{n} p_1^{k(n)}(\mathbf{z}_i)\mathbf{y_{i2}},$$

where $P_1 \equiv \left(p_1^{k(n)}(\mathbf{z}_1), ..., p_1^{k(n)}(\mathbf{z}_n)\right)'$ denotes data matrix and $(P_1'P_1)^{-}$ denotes the Moore-Penrose generalized inverse.

The residual for this first-step estimation is obtained as

$$\begin{aligned} \widehat{\mathbf{v}}_{\mathbf{i2}} &= \mathbf{y_{i2}} - \widehat{h}_n(\mathbf{z}_i) \\ &= \mathbf{y_{i2}} - p_1^{k(n)}(\mathbf{z}_i)'(P_1'P_1)^{-} \sum_{i=1}^{n} p_1^{k(n)}(\mathbf{z}_i)\mathbf{y_{i2}}. \end{aligned} \tag{10}$$

Using this residual as the input in the second step, to estimate the infinite-dimensional true parameter parameter g_o, at a given sample size n, a sample analog of the population maximum likelihood problem over the approximating sieve spaces $\mathcal{G}_n$ is defined as,

$$\widehat{g}_n = \arg\max_{g \in \Theta \times \mathcal{M}_n \times \mathcal{Q}_n} \frac{1}{n} \sum_{i=1}^{n} \log \Phi_2 \left(d_{i1} \left[\mathbf{x}_{i1}\boldsymbol{\beta} + m(y_{i2} - \widehat{h}_n(\mathbf{z}_i)) \right], d_{i2} \left[\mathbf{z}_i\boldsymbol{\gamma} + q(y_{i2} - \widehat{h}_n(\mathbf{z}_i)) \right], d_{i1}d_{i2}\rho \right). \tag{11}$$

The resulting sieve maximum likelihood estimator $\widehat{g}_n$ is composed of the following three components $\widehat{g}_n(\cdot) = \left(\widehat{\boldsymbol{\theta}}, \widehat{m}_n(\cdot), \widehat{q}_n(\cdot)\right)$ and does not have a closed-form solution.

3.4 Functionals of causal interest

Rather than delivering the magnitudes of causal effects of interest, coefficients in nonlinear models only reflect the direction of the impact. Similarly, unknown functions (h_o, g_o) in the semiparametric model in (6) are of limited interest to empirical researchers. Alternatively, average partial effects (APEs) based on an average structural function (ASF) measure the impact of an exogenous shift in a regressor on the binary response. In this semiparametric model, such APEs provide summary statistics that can compare actual outcomes to their potential outcomes.

More specifically, in model (6), the ASF is defined by taking expectation with respect to the unobservable

u_{i1} in the structural equation (6a), with $\mathbf{x}_1$ held fixed:

$$ASF_{\mathbf{x}_1} \equiv E_{u_{i1}}\left(1\left[\mathbf{x}_1\boldsymbol{\beta} + u_{i1} \geq 0\right]\right), \tag{12}$$

where subscript i indicates a random variable. The ASF defined in equation (12) is equivalent to evaluating the cumulative distribution function (CDF) of u_{i1} at $\mathbf{x}_1\boldsymbol{\beta}$, namely, $ASF_{\mathbf{x}_1} \equiv F_{u_{i1}}\left(\mathbf{x}_1\boldsymbol{\beta}\right)$. Further, by the law of iterated expectations, the ASF can also be obtained by taking expectations sequentially, first conditioning on $\mathbf{v_{i2}}$, and then over $\mathbf{v_{i2}}$:

$$\begin{aligned} ASF_{\mathbf{x}_1} &\equiv E_{\mathbf{v_{i2}}}\left[E_{u_{i1}|\mathbf{v_{i2}}}(1\left[\mathbf{x}_1\boldsymbol{\beta} + m_o(\mathbf{v_{i2}}) + v_{i1} \geq 0\right])\right] \\ &= E_{\mathbf{v_{i2}}}\left[E_{v_{i1}}\left(1\left[\mathbf{x}_1\boldsymbol{\beta} + m_o(\mathbf{v_{i2}}) + v_{1i} \geq 0\right]\right)\right] \qquad (13) \\ &= E_{\mathbf{v_{i2}}}\left(\Phi\left[\mathbf{x}_1\boldsymbol{\beta} + m_o(\mathbf{v_{i2}})\right]\right), \qquad (14) \end{aligned}$$

where equation (13) follows from the independence assumption between v_{i1} and $\mathbf{v_{i2}}$. Alternatively, plugging in the reduced form for $\mathbf{v_{i2}}$, the ASF can also be written with the data and true parameters from the first stage,

$$ASF_{\mathbf{x}_1} \equiv E_{\mathbf{w}_{i1}}\left[\Phi\left(\mathbf{x}_1\boldsymbol{\beta} + m_o\left[\mathbf{y_{i2}} - h_o(\mathbf{z}_i)\right]\right)\right], \tag{15}$$

where, recall that, $\mathbf{w}_{i1} \equiv (\mathbf{y_{i2}}, \mathbf{z}_i)$ denotes the first-stage data and $h_o(\cdot)$ denotes the true conditional mean function. In any case, as unobservables have been removed by taking expectations, an ASF is a function solely of $\mathbf{x}_1$.

By differentiating the ASF in equation (15) with respect to variables of interest, APEs in model (6) are obtained as follows. For a continuous EEV y_2, its APE, denoted by $\rho_{y_2}\left(h_o, g_o\right)$, is defined by taking the derivative of ASF with respect to y_2, as in equation (16):

$$\rho_{y_2}\left(h_o, g_o\right) \equiv \beta_2 E_{\mathbf{w}_{i1}}\left[\Phi_i^{(1)}\right], \tag{16}$$

where β_2 is the coefficient on y_2 and $\Phi_i^{(1)}$ denotes the first order derivative of the Normal CDF evaluated at the linear index, namely,

$$\Phi_i^{(1)} \equiv \phi\left(\mathbf{x}_1\boldsymbol{\beta} + m_o\left[\mathbf{y_{i2}} - h_o(\mathbf{z}_i)\right]\right). \tag{17}$$

For a binary EEV y_3, its APE, denoted by $\rho_{y_3}\left(h_o, g_o\right)$, is defined by taking a difference of ASF at the two

different values of y_3, as in equation (18):

$$\rho_{y_3}\left(h_o, g_o\right) \equiv E_{\mathbf{w}_{i1}}\left[\Phi_{i1} - \Phi_{i0}\right], \tag{18}$$

where Φ_{i1} denotes the Normal CDF evaluated at $y_3 = 1$, and Φ_{i0} denotes the Normal CDF evaluated at $y_3 = 0$, namely,

$$\Phi_{i1} \equiv \Phi\left(\mathbf{x}_{1(3)}\boldsymbol{\beta}_{(3)} + \beta_3 + m_o\left[\mathbf{y_{i2}} - h_o(\mathbf{z}_i)\right]\right), \tag{19}$$

$$\Phi_{i0} \equiv \Phi\left(\mathbf{x}_{1(3)}\boldsymbol{\beta}_{(3)} + m_o\left[\mathbf{y_{i2}} - h_o(\mathbf{z}_i)\right]\right), \tag{20}$$

with β_3 denoting the coefficient on y_3 and $\mathbf{x}_{1(3)}$ denoting $\mathbf{x}_1$ but without y_3.

As can be seen above, given a set of covariates $\mathbf{x}_1$, both of the APEs, $\rho_{\mathbf{y_2}}\left(\cdot,\cdot\right)$ and $\rho_{y_3}\left(\cdot,\cdot\right)$, are nonlinear functionals that take the infinite parameters $(h_o, \boldsymbol{\beta}, m_o)$ from both steps as arguments. To estimate these APEs, as we cannot take the expectation with respect to the first-stage data analytically, sample analogs of the true APEs in equations (16) and (18) are used, with the sieve two-step M-estimators $\left(\widehat{h}_n, \widehat{\boldsymbol{\beta}}, \widehat{m}_n\right)$ plugged in. Formally,

$$\widehat{\rho}_{y_2}\left(\widehat{h}_n, \widehat{g}_n\right) = \widehat{\beta}_2\left[n^{-1}\sum_{i=1}^{n}\widehat{\Phi}_i^{(1)}\right], \tag{21a}$$

$$\widehat{\rho}_{y_3}\left(\widehat{h}_n, \widehat{g}_n\right) = n^{-1}\sum_{i=1}^{n}\left(\widehat{\Phi}_{i1} - \widehat{\Phi}_{i0}\right), \tag{21b}$$

where $\widehat{\Phi}_i^{(1)}$, $\widehat{\Phi}_{i1}$ and $\widehat{\Phi}_{i0}$ are the sample analogs of (17), (19), and (20), respectively, as follows

$$\widehat{\Phi}_i^{(1)} \equiv \phi\left[\mathbf{x}_1\widehat{\boldsymbol{\beta}} + \widehat{m}_n(\mathbf{y_{i2}} - \widehat{h}_n(\mathbf{z}_i))\right], \tag{22}$$

$$\widehat{\Phi}_{i1} \equiv \Phi\left(\mathbf{x}_{1(3)}\widehat{\boldsymbol{\beta}}_{(3)} + \widehat{\beta}_3 + \widehat{m}_n(\mathbf{y_{i2}} - \widehat{h}_n(\mathbf{z}_i))\right), \tag{23}$$

$$\widehat{\Phi}_{i0} \equiv \Phi\left(\mathbf{x}_{1(3)}\widehat{\boldsymbol{\beta}}_{(3)} + \widehat{m}_n(\mathbf{y_{i2}} - \widehat{h}_n(\mathbf{z}_i))\right), \tag{24}$$

with $\left(\widehat{h}_n, \widehat{\boldsymbol{\beta}}, \widehat{m}_n\right)$ obtained from equations (9) and (11).

3.5 The procedure

Given the true parameters (h_o, g_o) as defined in Assumption 3.2 and 3.3 and the true APEs as defined in equations (16) and (18), more formally, a sieve two-step M-estimation procedure goes as follows.

PROCEDURE 3.1 *(a) As defined in (10), obtain the residual $\widehat{v}_2$ from the OLS regression of y_2 on $p_1^{k(n)}(\mathbf{z})$, the sieve basis functions for the first step, at a researcher's choice of $k\,(n)$.*

(b) As defined in (11), use $p_2^{l(n)}(\widehat{v}_2)$, the sieve basis functions for the second step, as additional regressors in the joint bivariate probit estimation between equation (6a) and (6c), at a researcher's choice of $l\,(n)$.

(c) As defined in equations (21a) and (21b), for each observation, construct APEs by averaging across $\widehat{v}_{i2}$.

The following section provides the asymptotic properties for estimators of parameters, as well as for estimators of the functionals of interest resulting from Procedure 3.1.

4 Asymptotic Properties

4.1 Consistency

Theorem 4.1 provides consistency results for the sieve two-step M-estimators, $\widehat{h}_n$ and $\widehat{g}_n$. Satisfying the continuity and uniform convergence condition in Chen (2007), the consistency of the first-step estimator $\widehat{h}_n$ is established. As this consistent estimator, $\widehat{h}_n$, is plugged into the second step likelihood function, the CDF of a bivariate distribution, by the continuous mapping theorem, the consistency of the second-step estimator $\widehat{g}_n$ is also established.

Theorem 4.1 *(a)Under Assumptions 3.1, 3.2, and 3.4, let $\widehat{h}_n$ be the sieve least squares estimator in the first step, defined by equation (9). We have $d_h\left(\widehat{h}_n, h_o\right) = o_p(1)$.*

(b)Given the first-step estimator $\widehat{h}_n$ is consistent, under Assumptions 3.1, 3.3 and 3.5, let $\widehat{g}_n$ be the sieve maximum likelihood estimator in the second step, defined by equation (11). We have $d_g\left(\widehat{g}_n, g_o\right) = o_p(1)$.

Theorem 4.2 provides consistency results for method of moments estimators, $\widehat{\rho}_{y_2}(\cdot,\cdot)$ and $\widehat{\rho}_{y_3}(\cdot,\cdot)$, with the two-step M-estimators $\widehat{h}_n$ and $\widehat{g}_n$ plugged in as their arguments. The consistency of these estimators of functionals of interest not only relies on the consistency of the estimator for the functionals, $\widehat{\rho}_{y_2}(\cdot,\cdot)$ and $\widehat{\rho}_{y_3}(\cdot,\cdot)$, but also on the consistency of the two plugged-in estimators, $\widehat{h}_n$ and $\widehat{g}_n$. A theoretical framework for deriving this consistency is provided in Chen, Linton, and Van Keilegom (2003). As the random functional involved in constructing APEs is a standard normal CDF, the uniform continuity condition is satisfied.

Theorem 4.2 *Assume that the true APEs of interest are nonlinear functionals, denoted by* $\rho_{y_2}(h_o, g_o)$ *and* $\rho_{y_3}(h_o, g_o)$ *as in equations (16) and (18). Under consistency of the sieve two-step M-estimators* $\left(\widehat{h}_n, \widehat{g}_n\right)$*, a method of moments estimator of APEs* $\widehat{\rho}_{y_2}(\cdot,\cdot)$ *and* $\widehat{\rho}_{y_3}(\cdot,\cdot)$ *with the two-step M-estimators* $\left(\widehat{h}_n, \widehat{g}_n\right)$ *plugged in as their arguments, as defined by equations (21a) and (21b), are consistent. Formally,* $\widehat{\rho}_{y_2}\left(\widehat{h}_n, \widehat{g}_n\right) \xrightarrow{p} \rho_{y_2}(h_o, g_o)$ *and* $\widehat{\rho}_{y_3}\left(\widehat{h}_n, \widehat{g}_n\right) \xrightarrow{p} \rho_{y_3}(h_o, g_o)$.

4.2 Asymptotic normality

The asymptotic normality for estimators of APEs draws on the sieve inference literature for nonlinear functionals. In particular, Theorem 3.1 in Hahn, Liao, and Ridder (2015) provides a general asymptotic normality result for a known functional that takes the sieve two-step M-estimates as its arguments. Instead of being a known functional, the estimator of APEs here is obtained via method of moments. As method of moments entitles consistency, we have $\widehat{\rho}_{y_2}\left(\widehat{h}_n, \widehat{g}_n\right) \xrightarrow{p} \rho_{y_2}\left(\widehat{h}_n, \widehat{g}_n\right)$ and $\widehat{\rho}_{y_3}\left(\widehat{h}_n, \widehat{g}_n\right) \xrightarrow{p} \rho_{y_3}\left(\widehat{h}_n, \widehat{g}_n\right)$. Then, by applying a stochastic equi-continuity under regularity conditions, one can show that,

$$\sup_{h,g \in \mathcal{N}_{h,n} \times \mathcal{N}_{g,n}} \left|\sqrt{n}\left(\widehat{\rho}_{y_2}(h,g) - \rho_{y_2}(h,g)\right)\right| = o_p(1). \tag{25}$$

Hence, the asymptotic distribution of $\widehat{\rho}_{y_2}\left(\widehat{h}_n, \widehat{g}_n\right)$ is equivalent to that of $\rho_{y_2}\left(\widehat{h}_n, \widehat{g}_n\right)$. By the same reasoning, the asymptotic distribution of $\widehat{\rho}_{y_3}\left(\widehat{h}_n, \widehat{g}_n\right)$ is equivalent to that of $\rho_{y_3}\left(\widehat{h}_n, \widehat{g}_n\right)$. Drawing on Theorem 3.1 in Hahn, Liao, and Ridder (2015), the following derives the asymptotic variance of APEs, $\left\|v^*_{y_2,n}\right\|^2_{sd}$ and $\left\|v^*_{y_3,n}\right\|^2_{sd}$.

First, assume that the sieve two-step estimators, $\widehat{h}_n$ and $\widehat{g}_n$, as defined in equations (9) and (11), belong to the shrinking neighborhood $\mathcal{N}_n = \{(h,g) : h \in \mathcal{N}_{h,n} \text{ and } g \in \mathcal{N}_{g,n}\}$ w.p.a. 1, where

$$\begin{aligned}
\mathcal{N}_{h,n} &= \{h \in \mathcal{H}_n : \|h - h_o\|_{\mathcal{H}} \leq \delta_{1n}\}, \\
\mathcal{N}_{g,n} &= \{g \in \mathcal{G}_n : \|g - g_o\|_{\mathcal{G}} \leq \delta_{2n}\}, \\
\delta_{1n} &= \delta^*_{1n}(\log(\log(n))), \\
\delta_{2n} &= \delta^*_{2n}(\log(\log(n))),
\end{aligned}$$

with δ^*_{1n} denoting the convergence rate of the first-step sieve M estimator under the metric $\|\cdot\|_{\mathcal{H}}$ and δ_{2n} denoting the convergence rate for the second-step sieve M-estimator under metric $\|\cdot\|_{\mathcal{G}}$.

For all $h \in \mathcal{N}_{h,n}$, denote the first-step criterion function as $\varphi(\mathbf{w}_{i1}, h_o) \equiv [y_{i2} - h_o(\mathbf{z}_i)]^2$. Then, the dif-

ference in the criterion function $\varphi(\mathbf{w}_{i1}, h) - \varphi(\mathbf{w}_{i1}, h_o)$ can be approximated linearly by $\Delta_\varphi(\mathbf{w}_{i1}, h_o)[h - h_o]$, where

$$\Delta_\varphi(\mathbf{w}_{i1}, h_o)[v_h] \equiv \left.\frac{\partial \varphi(\mathbf{w}_{i1}, h_o + \tau v_h)}{\partial \tau}\right|_{\tau=0} \text{ for any } v_h \in \mathcal{N}_{h,n} - \{h_o\}. \tag{26}$$

For any $v_{h_1}, v_{h_2} \in \mathcal{N}_{h,n}$, define an inner-product on $\mathcal{N}_{h,n}$ as

$$\langle v_{h_1}, v_{h_2} \rangle_\varphi \equiv - \left.\frac{\partial E[\Delta_\varphi(\mathbf{w}_{i1}, h_o + \tau v_{h_2})[v_{h_1}]]}{\partial \tau}\right|_{\tau=0}.$$

Let $\mathcal{V}_1$ be the Hilbert space generated by $\mathcal{H} - \{h_o\}$ under the inner product $\langle \cdot, \cdot \rangle_\varphi$, and $\|v\|_\varphi^2 = \langle v, v \rangle_\varphi$. Due to asymptotic equivalence theorem, we will focus on $\rho_{y3}\left(\widehat{h}_n, \widehat{g}_n\right)$ to derive the asymptotics. Assume that there are linear functionals $\frac{\partial \rho_{y_2}(h_o, g_o)}{\partial h}[\cdot] : \mathcal{V}_1 \to \mathbb{R}$ such that

$$\frac{\partial \rho_{y_2}(h_o, g_o)}{\partial h}[v] \equiv \left.\frac{\partial \rho_{y_2}(h_o + \tau v, g_o)}{\partial \tau}\right|_{\tau=0} \text{ for all } v \in \mathcal{V}_1.$$

Let $h_{o,n}$ denote the projection of h_o on $\mathcal{H}_n$ under the norm $\|\cdot\|_\varphi$. Let $\mathcal{V}_{1,n}$ denote the Hilbert space generated by $\mathcal{N}_{h,n} - \{h_{o,n}\}$. Then $\dim(\mathcal{V}_{1,n}) = k(n) < \infty$. By Riesz representation theorem, there are sieve Riesz representers $v^*_{y_2,h_n} \in \mathcal{V}_{1,n}$ such that

$$\frac{\partial \rho_{y_2}(h_o, g_o)}{\partial h}[v] \equiv \left\langle v^*_{y_2,h_n}, v \right\rangle_\varphi \text{ for all } v \in \mathcal{V}_{1,n}.$$

For any $g \in \mathcal{N}_{g,n}$, denote the second-step criterion function as

$$\psi(\mathbf{w}_{i2}, g, h_o) \equiv \log \Phi_2\left(d_{i1}\left[\mathbf{x}_{i1}\boldsymbol{\beta} + m(y_{i2} - h_o(\mathbf{z}_i))\right], d_{i2}\left[\mathbf{z}_i\boldsymbol{\gamma} + q(y_{i2} - h_o(\mathbf{z}_i))\right], d_{i1}d_{i2}\rho\right).$$

Then, the difference in the criterion function $\psi(\mathbf{w}_{i2}, g, h_o) - \psi(\mathbf{w}_{i2}, g_o, h_o)$ can be approximated linearly by $\Delta_\psi(\mathbf{w}_{i2}, g_o, h_o)[g - g_o]$, where

$$\Delta_\psi(\mathbf{w}_{i2}, g_o, h_o)[v_g] \equiv \left.\frac{\partial \psi(\mathbf{w}_{i2}, g_o + \tau v_g, h_o)}{\partial \tau}\right|_{\tau=0} \text{ for any } v_g \in \mathcal{N}_{g,n} - \{g_o\}.$$

For any $v_{g_1}, v_{g_2} \in \mathcal{N}_{g,n}$, we define an inner-product on $\mathcal{N}_{g,n}$ as

$$\langle v_{g_1}, v_{g_2} \rangle_\psi \equiv - \left.\frac{\partial E[\Delta_\psi(\mathbf{w}_{i2}, g_o + \tau v_{g_2}, h_o)[v_{g_1}]]}{\partial \tau}\right|_{\tau=0}.$$

Let $\mathcal{V}_2$ be the Hilbert space generated by $\mathcal{G} - \{g_o\}$ under the inner product $\langle \cdot, \cdot \rangle_\psi$, and $\|v\|_\psi^2 = \langle v, v \rangle_\psi$.

We assume that there is a linear functional $\frac{\partial \rho_{y_2}(h_o,g_o)}{\partial g}[\cdot]: \mathcal{V}_2 \to \mathbb{R}$ such that

$$\frac{\partial \rho_{y_2}(h_o,g_o)}{\partial g}[v] \equiv \left.\frac{\partial \rho_{y_2}(h_o,g_o+\tau v)}{\partial \tau}\right|_{\tau=0}.$$

Let $g_{o,n}$ denote the projection of g_o on $\mathcal{G}_n$ under the norm $\|\cdot\|_\psi$. Let $\mathcal{V}_{2,n}$ denote the Hilbert space generated by $\mathcal{N}_{g,n} - \{g_{o,n}\}$. Then $\dim(\mathcal{V}_{2,n}) = l(n) < \infty$. By Riesz representation theorem, there are sieve Riesz representers $v^*_{y_2,g_n} \in \mathcal{V}_{2,n}$ such that

$$\frac{\partial \rho_{y_2}(h_o,g_o)}{\partial g}[v] \equiv \left\langle v^*_{y_2,g_n}, v\right\rangle_\psi \text{ for all } v \in \mathcal{V}_{2,n}. \tag{27}$$

Let $\mathcal{V} = \mathcal{V}_1 \times \mathcal{V}_2$. For any $v = (v_h, v_g) \in \mathcal{V}$, we denote $\alpha_o \equiv (h_o, g_o)$

$$\frac{\partial \rho_{y_2}(\alpha_o)}{\partial \alpha}[v] \equiv \frac{\partial \rho_{y_2}(\alpha_o)}{\partial h}[v_h] + \frac{\partial \rho_{y_2}(\alpha_o)}{\partial g}[v_g] \equiv \left.\frac{\partial \rho_{y_2}(h_o+\tau v_h, g_o+\tau v_g)}{\partial \tau}\right|_{\tau=0}$$

To evaluate the effect of the first-step estimation on the asymptotic variance of the second-step sieve M estimator, we define a functional $\Gamma(\alpha_o)[\cdot,\cdot]$ on $\mathcal{V}$ as

$$\Gamma(\alpha_o)[v_1,v_2] = \left.\frac{\partial^2 E[\psi(\mathbf{w}_{i2}, g_o+\tau_2 v_2, h_o+\tau_1 v_1)]}{\partial \tau_1 \partial \tau_2}\right|_{\tau_1=0,\tau_2=0} \text{for any } (v_1,v_2) \in \mathcal{V}.$$

We assume that $\Gamma(\alpha_o)[\cdot,\cdot]$ is a bilinear functional on $\mathcal{V}$. Given the Riesz representers in (27), define $v^*_{y_2,\Gamma_n} \in \mathcal{V}_{1,n}$

$$\Gamma(\alpha_o)\left[v_h, v^*_{y_2,g_n}\right] \equiv \left\langle v_h, v^*_{y_2,\Gamma_n}\right\rangle \text{ for any } v_h \in \mathcal{V}_{1,n}.$$

Using the sieve Riesz representer $v^*_{y_2,h_n}, v^*_{y_2,g_n}$ and $v^*_{y_2,\Gamma_n}$, we define the asymptotic variance for APE of y_2 as

$$\left\|v^*_{y_2,n}\right\|^2_{sd} \equiv Var\left[n^{-\frac{1}{2}}\sum_{i=1}^n\left(\Delta_\varphi(\mathbf{w}_{i1},h_o)\left[v^*_{y_2,h_n}+v^*_{y_2,\Gamma_n}\right] + \Delta_\psi(\mathbf{w}_{i2},g_o,h_o)\left[v^*_{y_2,g_n}\right]\right)\right]. \tag{28}$$

By the same procedure, using the sieve Riesz representer $v^*_{y_3,h_n}, v^*_{y_3,g_n}$ and $v^*_{y_3,\Gamma_n}$, we define the asymptotic variance for APE of y_3 as

$$\left\|v^*_{y_3,n}\right\|^2_{sd} \equiv Var\left[n^{-\frac{1}{2}}\sum_{i=1}^n\left(\Delta_\varphi(\mathbf{w}_{i1},h_o)\left[v^*_{y_3,h_n}+v^*_{y_3,\Gamma_n}\right] + \Delta_\psi(\mathbf{w}_{i2},g_o,h_o)\left[v^*_{y_3,g_n}\right]\right)\right]. \tag{29}$$

After verifying the assumptions in Theorem 3.1 of Hahn, Liao, and Ridder (2015) as in Appendix A.1, the following Theorem 4.3 states that a third step estimator for APEs has a root n asymptotic normality.

Theorem 4.3 *With the two-step M-estimators, $\widehat{h}_n$ and $\widehat{g}_n$, plugged in as their arguments, method of moments*

estimators of APEs, $\widehat{\rho}_{y_2}(\cdot,\cdot)$ and $\widehat{\rho}_{y_3}(\cdot,\cdot)$, have the following asymptotic distributions,

$$\frac{\sqrt{n}\left[\widehat{\rho}_{y_2}\left(\widehat{h}_n,\widehat{g}_n\right)-\rho_{y_2}\left(h_o,g_o\right)\right]}{\left\|v^*_{y_2,n}\right\|_{sd}} \xrightarrow{d} N(0,1),$$
$$\frac{\sqrt{n}\left[\widehat{\rho}_{y_3}\left(\widehat{h}_n,\widehat{g}_n\right)-\rho_{y_3}\left(h_o,g_o\right)\right]}{\left\|v^*_{y_3,n}\right\|_{sd}} \xrightarrow{d} N(0,1),$$

where $\left\|v^*_{y_2,n}\right\|_{sd}$ and $\left\|v^*_{y_3,n}\right\|_{sd}$ are defined in (28) and (29).

4.3 Consistent variance estimation

Following Hahn, Liao, and Ridder (2015), I introduce the notion of $\left\|\widehat{v}^*_{y_2,n}\right\|_{n,sd}$, a consistent estimator of the asymptotic variance $\left\|v^*_{y_2,n}\right\|_{sd}$. The notation of $\left\|\widehat{v}^*_{y_3,n}\right\|_{n,sd}$ follows the same procedure and thus is omitted.

Denote

$$\Delta_\varphi(\mathbf{w}_{i1},h)[v_{h_1}] \equiv \frac{\partial\varphi(\mathbf{w}_{i1},h+\tau v_{h_1})}{\partial\tau}$$

and

$$r_\varphi(\mathbf{w}_{i1},h)[v_{h_1},v_{h_2}] \equiv \left.\frac{\partial\Delta_\varphi(\mathbf{w}_{i1},h+\tau v_{h_1})[v_{h_2}]}{\partial\tau}\right|_{\tau=0} \text{ for any } v_{h_1},v_{h_2}\in\mathcal{V}_{1,n}.$$

Similarly, define $\Delta_\psi(\mathbf{w}_{i2},g,h)[v_{g_1}]$ and $r_\psi(\mathbf{w}_{i2},g,h)[v_{g_1},v_{g_2}]$ for $v_{g_1},v_{g_2}\in\mathcal{V}_{2,n}$.

Define the empirical Riesz representer $\widehat{v}^*_{y_2,h_n}$ by

$$\frac{\partial\rho_{y_2}(\widehat{\alpha}_n)}{\partial h}[v_h] \equiv \left\langle v_h,\widehat{v}^*_{y_2,h_n}\right\rangle_{n,\varphi} \text{ for any } v_h\in\mathcal{V}_{1,n},$$

where $\langle v_{h_1},v_{h_2}\rangle_{n,\varphi} \equiv -\frac{1}{n}\sum_{i=1}^n r_\varphi\left(\mathbf{w}_{i1},\widehat{h}_n\right)[v_{h_1},v_{h_2}]$. Similarly, we define the empirical Riesz representer $\widehat{v}^*_{y_2,g_n}$ as

$$\frac{\partial\rho_{y_2}(\widehat{\alpha}_n)}{\partial g}[v_g] \equiv \left\langle v_g,\widehat{v}^*_{y_2,g_n}\right\rangle_{n,\psi} \text{ for any } v_g\in\mathcal{V}_{2,n},$$

where $\langle v_{g_1},v_{g_2}\rangle_{n,\psi} \equiv -\frac{1}{n}\sum_{i=1}^n r_\psi\left(\mathbf{w}_{i2},\widehat{h}_n,\widehat{g}_n\right)[v_{g_1},v_{g_2}]$. Define the empirical Riesz representer $\widehat{v}^*_{y_2,\Gamma_n}$ as

$$\Gamma_n\left(\widehat{h}_n,\widehat{g}_n\right)\left[\widehat{v}^*_{y_2,g_n},v_h\right] \equiv \left\langle v_h,\widehat{v}^*_{y_2,\Gamma_n}\right\rangle_{n,\varphi} \text{ for any } v_h\in\mathcal{V}_{1,n}$$

where

$$\Gamma_n\left(\widehat{h}_n,\widehat{g}_n\right)\left[\widehat{v}^*_{y_2,g_n},v_h\right] \equiv \frac{1}{n}\sum_{i=1}^n \left.\frac{\partial\Delta_\psi\left(\mathbf{w}_{i2},\widehat{g}_n,\widehat{h}_n+\tau v_h\right)\left[\widehat{v}^*_{y_2,g_n}\right]}{\partial\tau}\right|_{\tau=0}$$

As a sample analog of (28) and (29), consistent estimators of the asymptotic variance are

$$\left\|\widehat{v}^*_{y_2,n}\right\|^2_{n,sd} = \frac{1}{n}\sum_{i=1}^{n}\left|\left(\Delta_\varphi\left(\mathbf{w}_{i1},\widehat{h}_n\right)\left[\widehat{v}^*_{y_2,h_n}+\widehat{v}^*_{y_2,\Gamma_n}\right]+\Delta_\psi\left(\mathbf{w}_{i2},\widehat{g}_n,\widehat{h}_n\right)\left[\widehat{v}^*_{y_2,g_n}\right]\right)\right|^2,$$

$$\left\|\widehat{v}^*_{y_3,n}\right\|^2_{n,sd} = \frac{1}{n}\sum_{i=1}^{n}\left|\left(\Delta_\varphi\left(\mathbf{w}_{i1},\widehat{h}_n\right)\left[\widehat{v}^*_{y_3,h_n}+\widehat{v}^*_{y_3,\Gamma_n}\right]+\Delta_\psi\left(\mathbf{w}_{i2},\widehat{g}_n,\widehat{h}_n\right)\left[\widehat{v}^*_{y_3,g_n}\right]\right)\right|^2.$$

Theorem 4.4 *Suppose that the data are i.i.d. and the conditions in Theorem 4.3 are satisfied. Then under Assumption B.4, B.5 and B.6 in Appendix A.2, we have*

$$\left|\frac{\left\|\widehat{v}^*_{y_2,n}\right\|_{n,sd}}{\left\|v^*_{y_2,n}\right\|_{sd}}-1\right| = o_p(1),$$

$$\left|\frac{\left\|\widehat{v}^*_{y_3,n}\right\|_{n,sd}}{\left\|v^*_{y_3,n}\right\|_{sd}}-1\right| = o_p(1).$$

Therefore,

$$\frac{\sqrt{n}\left[\widehat{\rho}_{y_2}\left(\widehat{h}_n,\widehat{g}_n\right)-\rho_{y_2}\left(h_o,g_o\right)\right]}{\left\|\widehat{v}^*_{y_2,n}\right\|_{sd}} \xrightarrow{d} N(0,1),$$

$$\frac{\sqrt{n}\left[\widehat{\rho}_{y_3}\left(\widehat{h}_n,\widehat{g}_n\right)-\rho_{y_3}\left(h_o,g_o\right)\right]}{\left\|\widehat{v}^*_{y_3,n}\right\|_{sd}} \xrightarrow{d} N(0,1).$$

See Appendix A.2 for proof of the theorem.

5 Practical Inference

As indicated in Procedure 3.1, a researcher can estimate the semiparametric model by adding a certain number of basis functions to the parametric two-step biprobit, projecting the unknown functions onto finite dimensional spaces. The length of the basis functions is at the researcher's choice. In practice, with this finite dimensional model, inference for parameters of interest can be conducted via a standard delta method (see, for example, Wooldridge, 2010, section 12.4.1). Numerical equivalence between these practical inferences and the consistent estimators of asymptotic variance for a sieve two-step M-estimation procedure is shown in Hahn, Liao, and Ridder (2015).

To derive the practical inference for model (6), assume a researcher believes that the model is in fact

finite dimensional with dimensions fixed at $k = k(n)$ and $l = l(n)$, respectively:

$$h_o(\cdot) = p_1^k(\cdot)'\delta_{h_o}, \tag{32a}$$

$$g_o = (\boldsymbol{\theta}, m_o(\cdot), q_o(\cdot)) = \left(\boldsymbol{\theta}, p_2^l(\cdot)'\delta_{m_o}, p_2^l(\cdot)'\delta_{q_o}\right). \tag{32b}$$

where $p_1^k(\cdot)$ is the basis function for $\mathcal{H}_n$, as defined in (7), and $p_2^l(\cdot)$ is the basis function for $\mathcal{M}_n$ and $\mathcal{Q}_n$, as defined in (8). The first-step parameters δ_{h_o} is estimated by a parametric least squares:

$$\widehat{\delta}_h = \underset{\delta_h \in \mathcal{D}_h}{\arg\max} \frac{-1}{n} \sum_{i=1}^{n} \left[y_{i2} - p_1^k(\mathbf{z}_i)'\delta_h\right]^2,$$

where the parameter space $\mathcal{D}_h$ is a compact set in $\mathbb{R}^k$. The second-step parameters $\delta_{g_o} \equiv (\boldsymbol{\theta}, \delta_{m_o}, \delta_{q_o})$ is estimated by parametric joint maximum likelihood:

$$\begin{aligned}
\widehat{\delta}_g &= (\widehat{\boldsymbol{\theta}}, \widehat{\delta}_m, \widehat{\delta}_q) \\
&= \underset{(\boldsymbol{\beta},\rho,\boldsymbol{\gamma},\delta_m,\delta_q)\in\Theta\times\mathcal{D}_m\times\mathcal{D}_q}{\arg\max} \frac{1}{n}\sum_{i=1}^{n} \log \Phi_2 \left[d_{i1}\left(\mathbf{x}_{i1}\boldsymbol{\beta} + p_2^l\left[y_{i2} - p_1^k(\mathbf{z}_i)'\widehat{\delta}_h\right]'\delta_m\right), \right. \\
&\left. d_{i2}\left(\mathbf{z}_i\boldsymbol{\gamma} + p_2^l\left[y_{i2} - p_1^k(\mathbf{z}_i)'\widehat{\delta}_h\right]'\delta_q\right), d_{i1}d_{i2}\rho \right],
\end{aligned}$$

where Θ, $\mathcal{D}_m$ and $\mathcal{D}_q$ are compact sets in $\mathbb{R}^{d_{\boldsymbol{\theta}}}$, $\mathbb{R}^l$ and $\mathbb{R}^l$.

By a change of notation, the APEs of interest at finite dimension $\rho_{y_2}(h_{o,n}, g_{o,n})$ and $\rho_{y_3}(h_{o,n}, g_{o,n})$ can be defined in terms of the coefficients on the basis functions, $\pi_{y_2}(\delta_{h_o}, \delta_{g_o})$ and $\pi_{y_3}(\delta_{h_o}, \delta_{g_o})$, formally:

$$\begin{aligned}
\rho_{y_2}(h_{o,n}, g_{o,n}) &= \rho_{y_2}\left(p_1^k(\cdot)'\delta_{h_o}, \boldsymbol{\theta}, p_2^l(\cdot)'\delta_{m_o}, p_2^l(\cdot)'\delta_{q_o}\right) \\
&\equiv \pi_{y_2}(\delta_{h_o}, \delta_{g_o}) \\
&= \beta_2 E_{\mathbf{w}_{i1}}\left[\Phi_{i,p}^{(1)}\right],
\end{aligned}$$

and

$$\begin{aligned}
\rho_{y_3}(h_{o,n}, g_{o,n}) &= \rho_{y_3}\left(p_1^k(\cdot)'\delta_{h_o}, \boldsymbol{\theta}, p_2^l(\cdot)'\delta_{m_o}, p_2^l(\cdot)'\delta_{q_o}\right) \\
&\equiv \pi_{y_3}(\delta_{h_o}, \delta_{g_o}) \\
&= E_{\mathbf{w}_{i1}}\left[\Phi_{i1,p} - \Phi_{i0,p}\right],
\end{aligned}$$

where

$$\begin{aligned}
\Phi_{i,p}^{(1)} &\equiv \left.\frac{\partial \Phi_p(t)}{\partial t}\right|_{t=\mathbf{x}_1\boldsymbol{\beta}+p_2^l\left[y_{i2}-p_1^k(\mathbf{z}_i)'\delta_{h_o}\right]'\delta_{m_o}} \\
&= \phi\left[\mathbf{x}_1\boldsymbol{\beta}+p_2^l\left[y_{i2}-p_1^k(\mathbf{z}_i)'\delta_{h_o}\right]'\delta_{m_o}\right], \\
\Phi_{i1,p} &\equiv \Phi\left(\mathbf{x}_{1(3)}\boldsymbol{\beta}_{(3)}+\beta_3+p_2^l\left[y_{i2}-p_1^k(\mathbf{z}_i)'\delta_{h_o}\right]'\delta_{m_o}\right), \\
\Phi_{i0,p} &\equiv \Phi\left(\mathbf{x}_{1(3)}\boldsymbol{\beta}_{(3)}+p_2^l\left[y_{i2}-p_1^k(\mathbf{z}_i)'\delta_{h_o}\right]'\delta_{m_o}\right).
\end{aligned}$$

Now, absorb the randomness of the first step data $\mathbf{w}_{i1}$ into subscripts, and let the partial effect for an individual i be denoted by

$$\begin{aligned}
r_{i,y_2}\left(\delta_{h_o},\delta_{g_o}\right) &\equiv \beta_2\Phi_{i,p}^{(1)}, \\
r_{i,y_3}\left(\delta_{h_o},\delta_{g_o}\right) &\equiv \Phi_{i1,p}-\Phi_{i0,p}.
\end{aligned}$$

By the same notation, the sample analogs of the functional $\pi_{y_2}\left(\delta_{h_o},\delta_{g_o}\right)$ and $\pi_{y_3}\left(\delta_{h_o},\delta_{g_o}\right)$ are as follows,

$$\begin{aligned}
\widehat{\pi}_{y_2}\left(\widehat{\delta}_h,\widehat{\delta}_g\right) &\equiv \widehat{\beta}_2\left[n^{-1}\sum_{i=1}^{n}\widehat{\Phi}_{i,p}^{(1)}\right], \\
\widehat{\pi}_{y_3}\left(\widehat{\delta}_h,\widehat{\delta}_g\right) &\equiv n^{-1}\sum_{i=1}^{n}\left[\widehat{\Phi}_{i1,p}-\widehat{\Phi}_{i0,p}\right],
\end{aligned}$$

where

$$\begin{aligned}
\widehat{\Phi}_{i,p}^{(1)} &\equiv \left.\frac{\partial \Phi_p(t)}{\partial t}\right|_{t=\mathbf{x}_1\widehat{\boldsymbol{\beta}}+p_2^l\left[y_{i2}-p_1^k(\mathbf{z}_i)'\widehat{\delta}_h\right]'\widehat{\delta}_m} \\
&= \phi\left[\mathbf{x}_1\widehat{\boldsymbol{\beta}}+p_2^l\left[y_{i2}-p_1^k(\mathbf{z}_i)'\widehat{\delta}_h\right]'\widehat{\delta}_m\right], \\
\widehat{\Phi}_{i1,p} &\equiv \Phi\left(\mathbf{x}_{1(3)}\widehat{\boldsymbol{\beta}}_{(3)}+\widehat{\beta}_3+p_2^l\left[y_{i2}-p_1^k(\mathbf{z}_i)'\widehat{\delta}_h\right]'\widehat{\delta}_m\right), \\
\widehat{\Phi}_{i0,p} &\equiv \Phi\left(\mathbf{x}_{1(3)}\widehat{\boldsymbol{\beta}}_{(3)}+p_2^l\left[y_{i2}-p_1^k(\mathbf{z}_i)'\widehat{\delta}_h\right]'\widehat{\delta}_m\right).
\end{aligned}$$

Since we cannot observe the true parameters $(\delta_{h_o},\delta_{g_o})$, we plug in the estimates of these parameters into the individual partial effects instead,

$$r_{i,y_2}\left(\widehat{\delta}_h,\widehat{\delta}_g\right) \equiv \widehat{\beta}_2\widehat{\Phi}^{(1)}_{i,p},$$
$$r_{i,y_3}\left(\widehat{\delta}_h,\widehat{\delta}_g\right) \equiv \widehat{\Phi}_{i1,p} - \widehat{\Phi}_{i0,p}.$$

Under standard regularity conditions for parametric estimation, it is easy to show that the following proposition holds for inference of the APEs in the misspecified parametric problem. Note that APEs are sample analogs of nonlinear functions that contain the parameters from both steps as arguments.

Proposition 5.1 *When a researcher believes that the parameters, $\delta_{h_o}$and δ_{g_o}, are fixed at a finite dimension, as in equations (32a) and (32b), estimates of APEs of y_2 and y_3 have asymptotic distributions as follows,*

$$\sqrt{n}\left[\widehat{\pi}_{y_2}\left(\widehat{\delta}_h,\widehat{\delta}_g\right) - \pi_{y_2}\left(\delta_{h_o},\delta_{g_o}\right)\right] \xrightarrow{d} N\left(0,V_{y_2}\right),$$
$$\sqrt{n}\left[\widehat{\pi}_{y_3}\left(\widehat{\delta}_h,\widehat{\delta}_g\right) - \pi_{y_3}\left(\delta_{h_o},\delta_{g_o}\right)\right] \xrightarrow{d} N\left(0,V_{y_3}\right),$$

where V_{y_2} and V_{y_3} are the asymptotic variance defined as

$$V_{y_2} \equiv Var\left[r_{i,y_2}\left(\delta_{h_o},\delta_{g_o}\right) - \pi_{y_2}\left(\delta_{h_o},\delta_{g_o}\right) - R^{y_2}_{o,h}H^{-1}_{o,h}S_{i,h} - R^{y_2}_{o,g}H^{-1}_{o,g}\left(S_{i,g} + F_{o,gh}H^{-1}_{o,h}S_{i,h}\right)\right],$$
$$V_{y_3} \equiv Var\left[r_{i,y_3}\left(\delta_{h_o},\delta_{g_o}\right) - \pi_{y_3}\left(\delta_{h_o},\delta_{g_o}\right) - R^{y_3}_{o,h}H^{-1}_{o,h}S_{i,h} - R^{y_3}_{o,g}H^{-1}_{o,g}\left(S_{i,g} + F_{o,gh}H^{-1}_{o,h}S_{i,h}\right)\right].$$

Proof. See Appendix A.3 for the definition of the rest of the terms and derivation. ■

Since now the parameters live in the finite dimensional space, by the usual argument, the consistent estimator of the variance V_{y_2} and V_{y_3} are the sample analogs,

$$\widehat{V}_{y_2} \equiv \frac{1}{n}\sum_{i=1}^{n}\left[r_{i,y_2}\left(\widehat{\delta}_h,\widehat{\delta}_g\right) - \widehat{\pi}_{y_2}\left(\widehat{\delta}_h,\widehat{\delta}_g\right) - \widehat{R}^{y_2}_h\widehat{H}^{-1}_h\widehat{S}_{i,h} - \widehat{R}^{y_2}_g\widehat{H}^{-1}_g\left(\widehat{S}_{i,g} + \widehat{F}_{gh}\widehat{H}^{-1}_h\widehat{S}_{i,h}\right)\right]^2,$$
$$\widehat{V}_{y_3} \equiv \frac{1}{n}\sum_{i=1}^{n}\left[r_{i,y_3}\left(\widehat{\delta}_h,\widehat{\delta}_g\right) - \widehat{\pi}_{y_3}\left(\widehat{\delta}_h,\widehat{\delta}_g\right) - \widehat{R}^{y_3}_h\widehat{H}^{-1}_h\widehat{S}_{i,h} - \widehat{R}^{y_3}_g\widehat{H}^{-1}_g\left(\widehat{S}_{i,g} + F_{gh}\widehat{H}^{-1}_h\widehat{S}_{i,h}\right)\right]^2.$$

6 Simulation Study

6.1 Designs

In three designs, this section compares the finite sample behavior of the proposed estimator to alternative methods. These three designs differ in their levels of misspecification of the error terms. Design 1 is the

baseline design, where the error terms follow a conditional bivariate normal distribution, as in Assumption 2.2. Design 2 is the case of full misspecification, where error terms follow a joint Chi-square distribution, violating Assumption 2.2. Design 3 is a case of correct specification where the error terms follow a trivariate normal distribution, as in Assumption 2.1. In each design, the performance of the sieve two-step M-estimators, two-stage least squares (2SLS) estimator, and the special regressor estimator are compared. The number of Monte Carlo replications is 1000. The simulation results show the robustness of the proposed sieve two-step M-estimators, in terms of less bias and RMSE compared with other types of estimators.

Design 1

The data generating process is as follows:

$$y_1 = 1\left[z_1 + y_2 + y_3 + m_o(v_2) + v_1 \geq 0\right], \quad v_1 = v_3 + e_1, \quad e_1 \sim \text{Normal}(0,1), \quad v_3 \sim \text{Normal}(0,1),$$

$$y_2 = h_o(z_2, z_3) + v_2, \quad v_2 \sim 0.8\text{Normal}(-1, .6) + 0.2\text{Normal}(4, 2),$$

$$y_3 = 1\left[0.1z_2 + z_3 + q_o(v_2) + v_3 \geq 0\right],$$

$$z_1 \sim \text{Normal}(0, 9), \quad z_2 = 1[e_2 > 0], \quad e_2 \sim \text{Normal}(0, 1), \quad z_3 \sim \text{Normal}(0, 1),$$

$$m_o(v_2) = \frac{1}{1 + \exp(-v_2)}, \quad h_o(z_2, z_3) = \log\left(\left|z_2 + 0.1z_3\right|/2.5\right), \quad q_o(v_2) = v_2.$$

For the sieve two-step M-estimators, I use a power series to approximate the unknown control functions: $m_o(\cdot)$, $h_o(\cdot)$, and $q_o(\cdot)$. The power series is chosen over spline series for its simplicity of application. Let $u_1 = m_o(v_2) + v_1$, where $m_o(\cdot)$ denotes the conditional mean of u_1 given v_2. This conditional mean is designed as a squashing function restricted to the unit interval. Taking the form of a log transformation, $h_o(\cdot)$ denotes the conditional mean of y_2 given z_2 and z_3. Identity function $q_o(\cdot)$ is an transformation with an unrestricted range. To demonstrate the robustness of the sieve two-step M-estimators resulting from making the more flexible conditional normality assumption, v_2 is not chosen as a standard normal distribution, but rather a mixture of two normal distributions: Normal$(-1, .6)$ with probability 0.8 and Normal$(4, 2)$ with probability 0.2. This assumption on v_2 violates the joint normality assumption but satisfies the conditional normality assumption.

To compare with the sieve two-step M-estimator, other parametric and semiparametric instrumental variables estimation methods such as 2SLS and special regressor are conducted. In order to apply the method of 2SLS, exogenous variable z_1 is the instrument for itself, while exogenous variables z_2 and z_3 are instruments for the EEVs y_2 and y_3. More specifically, the binary exogenous variable z_2 is most highly

correlated with the continuous EEV y_2, and the continuous exogenous variable z_3 is most highly correlated with the binary EEV y_3. Further, z_1 satisfies the independence, additivity, continuity, and large support requirements of a special regressor. To illustrate the role of having large support in reducing the bias of the special regressor estimator, DGPs with both large and small support are compared. In particular, when drawn from Normal$(0, 9)$, z_1 is considered as having large support; when drawn from Normal$(0, 1)$, z_1 is considered as having small support.

For all the estimators, we are primarily interested in marginal effects. Given the data generating process in this design, the true marginal effects of interest are defined through APEs as follows:

$$\rho_{y_2} = \frac{1}{\sqrt{2}} \sum_{j=1}^{n} \sum_{i=1}^{n} \phi \left[\frac{1}{\sqrt{2}} \left(z_{j1} + y_{j2} + y_{j3} + m_o(v_{i2}) \right) \right],$$
$$\rho_{y_3} = \sum_{j=1}^{n} \sum_{i=1}^{n} \Phi \left[\frac{1}{\sqrt{2}} \left(z_{j1} + y_{j2} + 1 + m_o(v_{i2}) \right) \right] - \Phi \left[\frac{1}{\sqrt{2}} \left(z_{j1} + y_{j2} + m_o(v_{i2}) \right) \right].$$

which are the partial effects first averaged across v_{i2} by holding (z_{j1}, y_{j2}, y_{j3}) fixed and then across (z_{j1}, y_{j2}, y_{j3}). While the sieve two-step M-estimators can identify the APEs defined above, the special regressor method does not recover the structural error and thus does not permit estimating APEs. Alternatively, average index functions (AIFs) are used after special regressor estimation to obtain marginal effects. AIFs are estimated as the partial effects of the conditional mean of y_2 given the linear index of all regressors, in this case, including v_{i2}. Because of the differences in defining partial effects, a relative effect, or the ratio of $\frac{\beta_3}{\beta_2}$, is also reported for comparison across different types of the estimators. By design, the true ratio of the coefficients of y_3 over y_2 is $\frac{\beta_3}{\beta_2} = 1$.

Design 2

The data generating process is as follows:

$$y_1 = 1\left[z_1 + y_2 + y_3 + u_1 \geq 0\right], \quad u_1 = v_1 + v_2 + v_3 \sim \chi^2_3 - 3, \quad v_1, v_2, v_3 \overset{iid}{\sim} \chi^2_1 - 1,$$
$$y_2 = z_2 + 0.1z_3 + v_2,$$
$$y_3 = 1\left[0.1z_2 + z_3 + u_3 \geq 0\right], \quad u_3 = v_2 + v_3 \sim \chi^2_2 - 2,$$
$$z_1 \sim \text{Normal}(0, 9), \quad z_2 = 1[e_2 > 0], \quad e_2 \sim \text{Normal}(0, 1), \quad z_3 \sim \text{Normal}(0, 1).$$

Design 2 differs from Design 1 in the distributional assumptions of the error terms and the reduced-form specifications. In Design 2, v_1, v_2, and v_3 each independently follows a demeaned Chi-square distribution

with one degree of freedom, denoted by $\chi_1^2 - 1$. Thus, the error term u_1, as the sum of v_1, v_2, and v_3, follows a demeaned Chi-square distribution with three degrees of freedom. Similarly, the error term u_3, as the sum of v_2 and v_3, follows a demeaned Chi-square distribution with two degrees of freedom. As the error terms in Design 2 have larger variances than those in Design 1, due to the scaling effect,z_1 drawn from Normal$(0, 9)$ has relatively small support. To generate the case of large support, z_1 is also drawn from Normal$(0, 16)$. To show the bias resulting solely from the violation of the conditional normality assumption, the reduced form for y_2 is now simplified to a usual linear function, rather than a nonlinear function that needs to be approximated by the method of sieves.

Let $F_{\chi_3^2}(\cdot)$ denote the CDF of χ_3^2, the chi-squared distribution with three degrees of freedom, and $f_{\chi_3^2}(\cdot)$ denote the corresponding probability density function (PDF). Since all the error terms belong to the same distributional family, APEs for y_2 and y_3 at a fixed point (z_1, y_2, y_3) can be expressed analytically as follows:

$$\rho_{y_2} = f_{\chi_3^2}\left[3 - (z_1 + y_2 + y_3)\right],$$
$$\rho_{y_3} = F_{\chi_3^2}\left[3 - (z_1 + y_2)\right] - F_{\chi_3^2}\left[3 - (z_1 + y_2 + 1)\right].$$

The ratio of coefficients of y_3 over y_2 is still designed to be $\frac{\beta_3}{\beta_2} = 1$.

Design 3

The data generating process is as follows:

$$y_1 = 1\left[z_1 + y_2 + y_3 + u_1 \geq 0\right],$$
$$y_2 = z_2 + 0.1z_3 + v_2,$$
$$y_3 = 1\left[0.1z_2 + z_3 + u_3 \geq 0\right]$$
$$\begin{pmatrix} u_1 \\ v_2 \\ u_3 \end{pmatrix} \sim \text{Normal}\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 & 0.625 \\ 0.5 & 1 & 0.5 \\ 0.625 & 0.5 & 1 \end{pmatrix}\right]$$
$$z_1 \sim \text{Normal}(0, 9), \quad z_2 = 1[e_2 > 0], \quad e_2, z_3 \overset{iid}{\sim} \text{Normal}(0, 1).$$

Contrary to Design 2, the error terms in Design 3 is designed to follow a standard trivariate normal distribution. The method of sieves does not have a particular advantage in this design. The purpose of this design is to demonstrate the consequence of adding too many sieve terms in practice, by experimenting with

sieve spaces of various dimensions. Due to this joint normality assumption, the APEs for y_2 and y_3 at a fixed point (z_1, y_2, y_3) can be expressed analytically, as follows:

$$\rho_{y_2} = \frac{1}{2}\phi\left[\frac{1}{2}(z_1 + y_2 + y_3)\right],$$
$$\rho_{y_3} = \Phi\left[\frac{1}{2}(z_1 + y_2 + 1)\right] - \Phi\left[\frac{1}{2}(z_1 + y_2)\right].$$

The ratio of coefficients of y_3 over y_2 is still fixed at $\frac{\beta_3}{\beta_2} = 1$.

6.2 Results

The six estimators in this simulation study are named as follows: CF Biprobit K=3, CF Biprobit K=2, CF Biprobit K=1, 2SLS, SR Kernel, and SR KNN. CF Biprobit K=3 is the control function estimator with unknown functions approximated by sieve spaces of dimension three. CF Biprobit K=2 is the control function estimator with unknown functions approximated by sieve spaces of dimension two. CF Biprobit K=1 is simply a parametric two-step control function estimator. 2SLS is the linear probability model estimated with the usual IV approach. SR Kernel is the special regressor method using kernal method, and SR KNN is the special regressor method using K-Nearest Neighbor method. CF Biprobit K=3, CF Biprobit K=2, SR Kernel, and SR KNN are the semiparametric estimators. CF Biprobit and 2SLS are the parametric estimators. Figure 1, Figure 2, and Figure 3 present the empirical distributions of APEs for y_2 and y_3 and the ratio of $\frac{\beta_3}{\beta_2}$, with the red vertical bar depicting the true APE (and the true ratio) in each case.

Figure 1 depicts the performance of the estimators in Design 1 when z_1 has a large support. For all three quantities of interest (APE for y_2, APE for y_3, and $\frac{\beta_3}{\beta_2}$), CF Biprobit K=3, CF Biprobit K=2, and CF Biprobit K=1 are of the highest precision and of the smallest biases. The biases decrease slightly with the increase in the dimensions of the sieve spaces. For the APE of y_2, partial effects based on AIFs from SR Kernel and SR KNN suffers from the largest downward bias. The bias of coefficient from 2SLS is between the special regressor methods and the sieve methods. For the APE of y_3, partial effects based on AIF from SR Kernel and SR KNN almost center around the true APE, but with less precision. 2SLS suffers the largest upward bias. For the relative effects (ratio of $\frac{\beta_3}{\beta_2}$), SR KNN has a smaller bias than SR Kernel, with 2SLS having the largest upward bias.

Figure 2 depicts the performance of the estimators in Design 1 except that now z_1 has a small support. With the smaller support, for all three quantities of interest (APE for y_2, APE for y_3, and $\frac{\beta_3}{\beta_2}$), special

regressor estimators are the most biased, indicating the sensitivity of special regressor methods to the support condition. However, it is unclear whether the bias is due to the nature of the special regressor method or the different way their partial effect is constructed. The sieve estimators, CF Biprobit K=3, CF Biprobit K=2, and CF Biprobit K=1, are still of the highest precision and of the smallest biases. The impact of increase in dimensions is more obvious for all three quantities. More specifically, the improvement from adding one dimension is large when K increases from two to one, but becomes very small when K goes beyond three, indicating that K=2 would be an optimal sieve length for this given sample size. 2SLS has a moderate bias for APEs of y_2 and y_3 and a huge bias for the relative effects.

Figure 3 depicts the case when the support of z_1 is relatively small and error terms are fully misspecified as joint Chi-squared distribution, violating the trivariate normality or conditional normality assumption. The seive estimators, CF Biprobit K=3, CF Biprobit K=2 and CF Biprobit K=1 still perform the best with smallest bias and highest precision, indicating robustness of the seive estimators. However, due to the misspecification, increasing dimension in the sieve spaces does not significantly improve the approximation. Further, although special regressor methods do not seem to have a large bias in terms of the partial effects for y_2 and y_3, they have huge variances in the relative effects. 2SLS has a moderate bias and a relatively high precision in all cases. Due to the misspecification, the behaviors of the estimators do not change much when the support of z_1 is larger, as in Figure 4.

Figure 5 depicts the ideal case where the error terms are jointly normally distributed with a large support of z_1. Even though special regressor methods are centered around the true value for the relative effect, they have moderate bias for the partial effects. The sieve estimators, CF Biprobit K=3, CF Biprobit K=2, and CF Biprobit K=1 overlap with one another, indicating that adding more sieve terms does not hurt. 2SLS has a smaller bias compared with the previous designs.

Table 1 reports the biases and root-mean-square errors (RMSEs) for all three quantities (APE for y_2, APE for y_3 and $\frac{\beta_3}{\beta_2}$) in Design 1, with both large and small support of z_1. For the partial effects, when the support is small, not only the special regressors estimators, but also the sieve estimators have a larger bias and RMSE. For the relative effect, when the support is small, the sieve estimators have a larger bias but smaller RMSE, and the special regressor estimators have both larger bias and RMSE. Regardless of whether the support is large, adding more dimensions in the sieve spaces has an effect in reducing bias and improving precision.

Table 2 reports the biases and RMSEs for all the three quantities in Design 2, with both large and small

support of z_1. Although all of the estimators are misspecified, adding more dimensions in sieve spaces still significantly reduces bias for the partial effect of y_2. There is no observed improvement for the partial effect of y_3 or the relative effect, $\frac{\beta_3}{\beta_2}$. For the relative effects, the RMSE for SR Kernel and SR KNN are huge, indicating a large variance. This large variance decreases with the increase of the support of z_1.

Table 3 reports the biases and RMSEs for all the three quantities in Design 3, just with a large support. Although all of the estimators are consistent, as before, there are still improvements from increasing the dimension of the sieve spaces for the partial effects of y_2, y_3 and the relative effects.

To summarize, the simulation result in Design 1 shows that the CF Biprobit, with unknown functions approximated by sieve methods, is superior to other existing estimation methods, whether parametric or semiparametric. This is because the estimates of the true APEs converge quickly as the number of basis functions increases, sufficiently close to the true APE by simply adding two or three basis functions. Design 2 demonstrates the robustness of the CF Biprobit, with unknown functions approximated by sieve methods under misspecification. Adding more sieve basis functions, although not ideal, does not impair efficiency. Design 3 illustrates that, under trivariate normality, adding more sieve terms would increase the mean of APEs, which may correct for some bias. Note that as documented in Ai and Chen (2003), the power series may have erratic tail behaviors, so the spline sieve will be used to examine approximation performance. Since the second step biprobit is a MLE estimator, in practice, software packages use penalization methods in MLE to overcome numerical instability, which creates bias too. So the simulation results may have suffered from this embedded problem.

7 Conclusion and Future Work

This paper proposes a sieve two-step M-estimation via a control function approach to account for endogeneity in a triangular system for a single index binary response model. The endogeneity comes from one dummy EEV and from potentially many continuous EEVs. By adding more flexibility and robustness, this semiparametric two-step procedure serves as an extension and improvement of the parametric two-step biprobit proposed in Lin and Wooldridge (2015a).

In particular, the sieve two-step M-estimation relaxes the trivariate joint normality assumption in the parametric case to a conditional bivariate normality assumption. The latent errors for the binary outcome and dummy EEV are decomposed in a partial linear fashion. Each of the latent errors contains an unknown

function as its conditional mean, plus an independent remainder term. The remainder terms of the latent errors are assumed to be bivariate normally distributed. The first-stage sieve least squares regression error enters the unknown function as an argument. The second step simultaneously estimates the linear indexes and the unknown control functions by using a sieve joint MLE of the binary outcome and dummy EEV. The method of sieves makes functional forms more flexible for reduced forms of the continuous EEV and imposes less restriction on the joint distribution of error terms. It is the shape restriction feature of the sieves that enables us to decompose the error terms in this partial linear fashion, and thus, for a given sample size, this two-step sieve procedure is essentially turned into a parametric procedure. A researcher could simply add a certain number of sieve basis functions to transform variables by hand in both steps.

Average partial effects (APEs) based on the average structural function (ASF) are proposed as the measure of causal effects. A third-step estimation of the APEs employs a method of moments estimator, with the previous two-step estimators plugged in. Asymptotic properties, such as consistency and asymptotic normality, for the APEs and two-step estimators can be easily derived by mapping this model to a stream of theoretical sieve literature. Moreover, I show a practical inference for the APEs using a standard delta method for a parametric multiple step estimation as in Wooldridge (2010). The numerical equivalence of the parametric inference and the semiparametric inferece, as established in Hahn, Liao, and Ridder (2015), allows researchers to turn to the practical inference for simplicity. In the Monte Carlo experiments, I illustrate how the sieve approximation performs with a variation of sieve lengths under different degrees of misspecification.

As a topic of on-going research, this sieve two-step M-estimator for the binary response model with varying nature of endogeneity embraces numerous possibilities for future extensions. First, in order to further relax the level of distributional assumptions, it is computationally easy to apply the semi-nonparametric maximum likelihood estimator by Gallant and Nychka (1987) towards approximating the multivariate density of the likelihood function in the second step. In that case, asymptotics for a two-step estimator need to be carefully derived. Another avenue for relaxing the joint distribution is to use sieve approximations for the copula function that is now required to be parametric in literature. Second, to accommodate other practical complications in economic data, interactions between the error terms and explanatory variables can be used to arrive at endogenous switching or random coefficients models. It is also interesting to add weak instruments (Staiger and Stock, 1997) into this framework. Third, it is important to discuss in theory the conditions for point identification in this model. Partial identification can be employed in the scenarios

where point identification is not enabled. Fourth, rather than computing APEs at a given observation, one could derive APEs over the whole sample. U statistics can be used to derive asymptotic properties. Distributional treatment effects, which recover treatment effects for the entire distribution of outcomes, are also a promising way to provide more causal statistics to empirical researchers.

A.1 Assumptions and Proof of Results in Section 4.2

Proof. of Theorem 4.3. The following borrows from Hahn, Liao, and Ridder (2015) Appendix B.

ASSUMPTION A.1 **(a)** $\liminf_n \left\|v^*_{y_2,n}\right\|_{sd} > 0; \liminf_n \left\|v^*_{y_3,n}\right\|_{sd} > 0$

(b) *Let* $\alpha_o \equiv (h_o, g_o)$ *be the true unknown functions, the functional* $\rho_{y_2}(\cdot,\cdot)$ *and* $\rho_{y_3}(\cdot,\cdot)$ *satisfies*

$$\begin{aligned}
\sup_{\alpha\in\mathcal{N}_n}\left|\frac{\rho_{y_2}(\alpha)-\rho_{y_2}(\alpha_o)-\frac{\partial\rho_{y_2}(\alpha_o)}{\partial h}[h-h_o]-\frac{\partial\rho_{y_2}(\alpha_o)}{\partial g}[g-g_o]}{\left\|v^*_{y_2,n}\right\|_{sd}}\right| &= o\left(n^{-\frac{1}{2}}\right),\\
\sup_{\alpha\in\mathcal{N}_n}\left|\frac{\rho_{y_3}(\alpha)-\rho_{y_3}(\alpha_o)-\frac{\partial\rho_{y_3}(\alpha_o)}{\partial h}[h-h_o]-\frac{\partial\rho_{y_3}(\alpha_o)}{\partial g}[g-g_o]}{\left\|v^*_{y_3,n}\right\|_{sd}}\right| &= o\left(n^{-\frac{1}{2}}\right);
\end{aligned}$$

(c) *There exists* $g_n \in \mathcal{G}_n$ *such that* $\|g_n - g_o\|_{\mathcal{G}} = O\left(\delta^*_{2,n}\right)$ *and for any* $v_h \in \mathcal{V}_1$ *and* $v_g \in \mathcal{V}_2$, $\|v_h\|_\varphi \leq c_\varphi \|v_h\|_{\mathcal{H}}$ *and* $\|v_g\|_\psi \leq c_\psi \|v_h\|_{\mathcal{G}}$*where* c_φ *and* c_ψ *are some generic finite positive constants;*

(d)

$$\begin{aligned}
\frac{1}{\left\|v^*_{y_2,n}\right\|_{sd}}\max\left\{\left|\frac{\partial\rho_{y_2}(\alpha_o)}{\partial h}[h_{o,n}-h_o]\right|,\frac{\partial\rho_{y_2}(\alpha_o)}{\partial g}[g_{o,n}-g_o]\right\} &= o\left(n^{-\frac{1}{2}}\right);\\
\frac{1}{\left\|v^*_{y_3,n}\right\|_{sd}}\max\left\{\left|\frac{\partial\rho_{y_3}(\alpha_o)}{\partial h}[h_{o,n}-h_o]\right|,\frac{\partial\rho_{y_3}(\alpha_o)}{\partial g}[g_{o,n}-g_o]\right\} &= o\left(n^{-\frac{1}{2}}\right)
\end{aligned}$$

Assumption A.1 (a) ensures tha the sieve variance is asymptotic nonzero. Assumption A.1 (b) implies that there is a linear approximation for $\rho_{y_2}(\alpha_o)$ and $\rho_{y_3}(\alpha_o)$ uniformly over $\alpha \in \mathcal{N}_n$ with approximation error $o\left(\left\|v^*_{y_3,n}\right\|_{sd} n^{-\frac{1}{2}}\right)$. Assumption A.1 (c) implies that $\|\cdot\|_\varphi$ and $\|\cdot\|_\psi$ may be weaker than the pseudo-metrics $\|\cdot\|_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{G}}$ respectively. By Assumption A.1 (c) and the definition of $g_{o,m}$, we have

$$\|g_o - g_{o,n}\|_\varphi \leq \|g_o - g_n\|_\varphi \leq \|g_o - g_n\|_{\mathcal{H}} = O\left(\delta^*_{2,n}\right)$$

which indicates that $g_{o,n} \in \mathcal{N}_{g,n}$. Similarly, we have $h_{o,n} \in \mathcal{N}_{h,n}$, which together with the former result implies that $(h_{o,n}, g_{o,n}) \in \mathcal{N}_n$. Assumption A.1 (d) also requires that the sieve approximation error converges to zero at a rate faster than $\left\|v^*_{y_2,n}\right\|_{sd} n^{-\frac{1}{2}}$ and $\left\|v^*_{y_3,n}\right\|_{sd} n^{-\frac{1}{2}}$, which is an under-smoothing condition to derive the zero mean asymptotic normality of the sieve plug-in estimator $\rho(\widehat{\alpha}_n)$.

Define $\left(u^*_{h_n}, u^*_{g_n}, u^*_{\Gamma_n}\right) = \|v^*_n\|^{-1}_{sd}\left(v^*_{h_n}, v^*_{g_n}, v^*_{\Gamma_n}\right)$ and $g^* = g \pm \epsilon_n u^*_{g_n}$ for any $g \in \mathcal{N}_{g,n}$, where $\epsilon_n = o\left(n^{-1/2}\right)$ is some positive sequence. Let $\mu_n[\cdot]$ be the emprical process such that

$\mu_n[\psi(Z_{2,}g,h)] \equiv \frac{1}{n}\sum_{i=1}^n \{\psi(Z_{2,i},g,h) - E[\psi(Z_{2,}g,h)]\}$.

ASSUMPTION A.2 (a) *The following stochastic equicontinuity condition hold:*

$$\begin{aligned} \sup_{\alpha\in\mathcal{N}_n} \left|\mu_n\left\{\psi(Z_2,g^*,h) - \psi(Z_2,g,h) - \Delta_\psi(Z_2,g,h)\left[\pm\varepsilon_n u^*_{g_n}\right]\right\}\right| &= O_p\left(\varepsilon_n^2\right) \\ \text{and } \sup_{\alpha\in\mathcal{N}_n} \left|\mu_n\left\{\Delta_\psi(Z_2,g,h)\left[u^*_{g_n}\right] - \Delta_\psi(Z_2,g_o,h_o)\left[u^*_{g_n}\right]\right\}\right| &= O_p(\varepsilon_n) \end{aligned}$$

(b) *let* $K_\psi(g,h) \equiv E[\psi(Z_2,g,h) - \psi(Z_2,g_o,h_o)]$, *then*

$$K_\psi(g,h) - K_\psi(g^*,h) = \mp\varepsilon_n\Gamma(\alpha_o)\left[h-h_o, u^*_{g_n}\right] + \frac{\|g^*-g_o\|_\psi^2 - \|g-g_o\|_\psi^2}{2} + O\left(\varepsilon_n^2\right) \quad \text{(A.1)}$$

uniformaly over $(h,g)\in\mathcal{N}_n$.

The stochastic equicontinuity conditions are regular assumptions in the sieve method literature, e.g., Shen (1997), Chen and Shen (1998) and Chen, Liao, and Sun (2012). Assumption A.2 (b) implies that the Kullback-Leibler type of distance has a local quadratic approximation uniformly over the shrinking neighborhood $\mathcal{N}_n$. When there is no first-step estimate $\widehat{h}_n$, i.e. $h = h_o$ in (A.1), Assumption A.2 (b) will be reduced to

$$\sup_{\alpha\in\mathcal{N}_n}\left|K_\psi(g,h) - K_\psi(g^*,h) - \frac{\|g^*-g_o\|_\psi^2 - \|g-g_o\|_\psi^2}{2}\right| = O\left(\varepsilon_n^2\right)$$

which is the condition used in Chen, Liao, and Sun (2012) to derive the asymptotic normality of one-step sieve plug-in estimate $\rho(\widehat{g}_n)$. As a result, we can view the extra term in (A.1) as the estimation effect that the first-step estimate $\widehat{h}_n$ introduces to the asymptotic distribution of the second-step sieve M-estimator $\widehat{g}_n$.

ASSUMPTION A.3 (a) *The first-step sieve M estimator* $\widehat{h}_n$ *satisfies*

$$\left|\left\langle \widehat{h}_n - h_o, u^*_{h_n} + u^*_{\Gamma_n}\right\rangle_\varphi - \mu_n\left\{\Delta_\varphi(Z_1,h_o)\left[u^*_{h_n} + u^*_{\Gamma_n}\right]\right\}\right| = O_p(\varepsilon_n)$$

(b) *the following central limit theorem (CLT) holds:*

$$n^{-\frac{1}{2}}\sum_{i=1}^n\left\{\Delta_\varphi(Z_{1,i},h_o)\left[u^*_{h_n} + u^*_{\Gamma_n}\right] + \Delta_\psi(Z_{1,i},g_o,h_o)\left[u^*_{g_n}\right]\right\} \xrightarrow{d} N(0,1)$$

where Normal $(0,1)$ *denotes a standard normal random variable;*

(c) $\epsilon_{2,n} = O(\epsilon_n)$, $\epsilon_n\delta_{2,n}^{*-1} = o(1)$ *and* $\left\|u^*_{g_n}\right\|_\psi = O(1)$.

Assumption A.3 (a) is a high level condition, which is established in Chen, Liao, and Sun (2012) under a set of sufficient conditions. Assumption A.3 (b) is implied by the triangle array CLTs. Assumption A.3 (b) implies that the optimization error $\epsilon_{2,n}$ in the second-step sieve M-estimation is of the same or larger order as ϵ_n. As $\delta_{2,n}^{*-1}$ is the convergence rate of the second-step sieve M-estimator $\widehat{g}_n$, under the stationary data assumption, it is reasonable to assume that $\delta_{2,n}^{*-1}$ converges to zero at the rate not faster than root-n, which explains the assumption $\epsilon_n \delta_{2,n}^{*-1} = o(1)$. By Assumption A.3 (c), $\epsilon_n \delta_{2,n}^{*-1} = o(1)$ and the triangle inequality, we have

$$\left\| \widehat{g}_n^* - g_o \right\|_\psi \leq \left\| \widehat{g}_n - g_o \right\| + \varepsilon_n \left\| u_{g_n}^* \right\|_\psi = O_p\left(\delta_{g,n}^*\right) \tag{A.2}$$

which implies that $\widehat{g}_n^* \in \mathcal{N}_{g,n}$ wpa1.

By the definition of $\widehat{g}_n$, we have

$$\begin{aligned} -O_p\left(\varepsilon_{2,n}^2\right) &\leq \frac{1}{n}\sum_{i=1}^n \psi\left(Z_{2,i}, \widehat{g}_n, \widehat{h}_n\right) - \frac{1}{n}\sum_{i=1}^n \left(Z_{2,i}, \widehat{g}_n^*, \widehat{h}_n\right) \\ &= \mu_n\left\{\psi\left(Z_2, \widehat{g}_n, \widehat{h}_n\right) - \psi\left(Z_2, \widehat{g}_n^*, \widehat{h}_n\right) + \Delta_\psi\left(Z_2, \widehat{g}_n, \widehat{h}_n\right)\left[\pm\varepsilon_n u_{g_n}^*\right]\right\} \\ &\quad + \mu_n\left\{\Delta_\psi\left(Z_2, g_o, h_o\right)\left[\pm\varepsilon_n u_{g_n}^*\right] - \Delta_\psi\left(Z_2, \widehat{g}_n, \widehat{h}_n\right)\left[\pm\varepsilon_n u_{g_n}^*\right]\right\} \\ &\quad - \mu_n\left\{\Delta_\psi\left(Z_2, g_o, h_o\right)\left[\pm\varepsilon_n u_{g_n}^*\right]\right\} + \left[K_\psi\left(\widehat{g}_n, \widehat{h}_n\right) - K_\psi\left(\widehat{g}_n^*, \widehat{h}_n\right)\right] \end{aligned} \tag{A.3}$$

By Assumption A.2 (a), we have

$$\mu_n\left\{\psi\left(Z_2, \widehat{g}_n, \widehat{h}_n\right) - \psi\left(Z_2, \widehat{g}_n^*, \widehat{h}_n\right) + \Delta_\psi\left(Z_2, \widehat{g}_n, \widehat{h}_n\right)\left[\pm\varepsilon_n u_{g_n}^*\right]\right\} = O_p\left(\varepsilon_n^2\right) \tag{A.4}$$

$$\text{and } \mu_n\left\{\Delta_\psi\left(Z_2, g_o, h_o\right)\left[u_{g_n}^*\right] - \Delta_\psi\left(Z_2, \widehat{g}_n, \widehat{h}_n\right)\left[u_{g_n}^*\right]\right\} = O_p\left(\varepsilon_n\right) \tag{A.5}$$

Note that $\Gamma(\alpha_o)[\cdot,\cdot]$ is a bilinear functional. Using (A.2), Assumption A.2 (b) and A.3 (c), we deduce that

$$\begin{aligned} & K_\psi\left(\widehat{g}_n, \widehat{h}_n\right) - K_\psi\left(\widehat{g}_n^*, \widehat{h}_n\right) \\ = \; & \mp\varepsilon_n \Gamma(\alpha_o)\left[\widehat{h}_n - h_o, u_{g_n}^*\right] + \frac{\left\|\widehat{g}_n^* - g_o\right\|_\psi^2 - \left\|\widehat{g}_n - g_o\right\|_\psi^2}{2} + o_p\left(\varepsilon_n^2\right) \\ = \; & \left\langle \mp\varepsilon_n u_{\Gamma_n}^*, \widehat{h}_n - h_o \right\rangle_\varphi + \frac{\varepsilon_n^2 \left\|u_{g_n}^*\right\|_\psi^2}{2} + \left\langle \pm\varepsilon_n u_{gn}^*, \widehat{g}_n - g_o \right\rangle_\psi + o_p\left(\varepsilon_n^2\right) \\ = \; & \left\langle \mp\varepsilon_n u_{\Gamma_n}^*, \widehat{h}_n - h_o \right\rangle_\varphi + \left\langle \pm\varepsilon_n u_{gn}^*, \widehat{g}_n - g_o \right\rangle_\psi + O_p\left(\varepsilon_n^2\right) \end{aligned} \tag{A.6}$$

From $\varepsilon_{2,n} = O(\varepsilon_n)$, (A.3), (A.4), (A.5) and (A.6), we get

$$-O_p\left(\varepsilon_n^2\right) \leq \mp\varepsilon_n \mu_n\left\{\Delta_\psi\left(Z_2, g_o, h_o\right)\left[u_{g_n}^*\right]\right\} \mp \varepsilon_n \left\langle u_{\Gamma_n}^*, \widehat{h}_n - h_o \right\rangle_\varphi \pm \varepsilon_n \left\langle u_{g_n}^*, \widehat{g}_n - g_o \right\rangle_\psi.$$

Dividing by ε_n, we obtain

$$\left|\left\langle u_{g_n}^*, \widehat{g}_n - g_o\right\rangle_\psi - \left\langle u_{\Gamma_n}^*, \widehat{h}_n - h_o\right\rangle_\varphi - \mu_n\left\{\Delta_\psi\left(Z_2, g_o, h_o\right)\left[u_{g_n}^*\right]\right\}\right| = O_p\left(\varepsilon_n\right) \tag{A.7}$$

By definition, $g_{o,n}$ is the projection of g_o on $\mathcal{V}_{2,n}$ under the semi-norm $\|\cdot\|_\psi$. Hence there is $\left\langle g_{o,n} - g_o, u_{g_n}^*\right\rangle_\psi = 0$ and

$$\left\langle \widehat{g}_n - g_o, u_{g_n}^*\right\rangle_\psi = \left\langle \widehat{g}_n - g_{o,n}, u_{g_n}^*\right\rangle_\psi \tag{A.8}$$

From (A.7), (A.8) and $\varepsilon_n = o(n^{-\frac{1}{2}})$, we get

$$\left|\left\langle \widehat{g}_n - g_{o,n}, u_{g_n}^*\right\rangle_\psi - \left\langle u_{\Gamma_n}^*, \widehat{h}_n - h_o\right\rangle_\varphi - \mu_n\left\{\Delta_\psi\left(Z_2, g_o, h_o\right)\left[u_{g_n}^*\right]\right\}\right| = o_p\left(n^{-\frac{1}{2}}\right). \tag{A.9}$$

By Assumption A.1 (a)-(c) and the Riesz representation theorem,

$$\begin{aligned}
&\sqrt{n}\frac{\rho_{y_2}\left(\widehat{h}_n, \widehat{g}_n\right) - \rho_{y_2}\left(h_{o,n}, g_{o,n}\right)}{\left\|v_{y_2,n}^*\right\|_{sd}} = \sqrt{n}\frac{\frac{\partial\rho_{y_2}(\alpha_o)}{\partial h}\left[\widehat{h}_n - h_{o,n}\right] + \frac{\partial\rho_{y_2}(\alpha_o)}{\partial h}\left[\widehat{g}_n - g_{o,n}\right]}{\left\|v_{y_2,n}^*\right\|_{sd}} \\
&+\sqrt{n}\frac{\rho_{y_2}\left(\widehat{h}_n, \widehat{g}_n\right) - \rho_{y_2}\left(h_o, g_o\right) - \frac{\partial\rho_{y_2}(\alpha_o)}{\partial h}\left[\widehat{h}_n - h_o\right] - \frac{\partial\rho_{y_2}(\alpha_o)}{\partial g}\left[\widehat{g}_n - g_o\right]}{\left\|v_{y_2,n}^*\right\|_{sd}} \\
&-\sqrt{n}\frac{\rho_{y_2}\left(h_{o,n}, g_{o,n}\right) - \rho_{y_2}\left(h_o, g_o\right) - \frac{\partial\rho_{y_2}(\alpha_o)}{\partial h}\left[h_{o,n} - h_o\right] - \frac{\partial\rho_{y_2}(\alpha_o)}{\partial g}\left[g_{o,n} - g_o\right]}{\left\|v_{y_2,n}^*\right\|_{sd}} \\
= \;& \sqrt{n}\left[\left\langle \widehat{h}_n - h_{o,n}, u_{h_n}^*\right\rangle_\varphi + \left\langle \widehat{g}_n - g_{o,n}, u_{g_n}^*\right\rangle_\psi\right] + o_p(1)
\end{aligned} \tag{A.10}$$

By definition, $h_{o,n}$ is the projection of h_o on $\mathcal{V}_{1,n}$ under the semi-norm $\|\cdot\|_\varphi$. Hence there is $\left\langle h_{o,n} - h_o, u_{h_n}^*\right\rangle_\varphi = 0$ and

$$\left\langle \widehat{h}_n - h_{o,n}, u_{h_n}^*\right\rangle_\varphi = \left\langle \widehat{h}_n - h_o, u_{h_n}^*\right\rangle_\varphi \tag{A.11}$$

From the results in (A.10) and (A.11), we get

$$\sqrt{n}\frac{\rho_{y_2}\left(\widehat{h}_n, \widehat{g}_n\right) - \rho_{y_2}\left(h_{o,n}, g_{o,n}\right)}{\left\|v_{y_2,n}^*\right\|_{sd}} = \sqrt{n}\left[\left\langle \widehat{h}_n - h_o, u_{h_n}^*\right\rangle_\varphi + \left\langle \widehat{g}_n - g_{o,n}, u_{g_n}^*\right\rangle_\psi\right] + o_p(1)$$

which, together with (A.9) and Assumption A.3 (a), implies that

$$\begin{aligned}
&\sqrt{n}\frac{\rho_{y_2}\left(\widehat{h}_n,\widehat{g}_n\right)-\rho_{y_2}\left(h_{o,n},g_{o,n}\right)}{\left\|v^*_{y_2,n}\right\|_{sd}} \\
= \;& \sqrt{n}\left[\left\langle \widehat{h}_n-h_o,u^*_{h_n}+u^*_{\Gamma_n}\right\rangle_{\varphi}+\mu_n\left\{\Delta_\psi\left(Z_2,g_o,h_o\right)\left[u^*_{g_n}\right]\right\}\right]+o_p(1) \\
= \;& n^{-\frac{1}{2}}\sum_{i=1}^{n}\left\{\Delta_\varphi\left(Z_{1,i},h_o\right)\left[u^*_{h_n}+u^*_{\Gamma_n}\right]+\Delta_\psi\left(Z_{2,i},g_o,h_o\right)\left[u^*_{g_n}\right]\right\}+o_p(1) \qquad \text{(A.12)}
\end{aligned}$$

Furthermore, from Assumption A.1, we get

$$\begin{aligned}
&\left|\frac{\rho_{y_2}\left(h_{o,n},g_{o,n}\right)-\rho_{y_2}\left(h_o,g_o\right)}{\left\|v^*_{y_2,n}\right\|_{sd}}\right| \\
\leq \;& \left|\frac{\rho_{y_2}\left(h_{o,n},g_{o,n}\right)-\rho_{y_2}\left(h_o,g_o\right)-\frac{\partial\rho_{y_2}(\alpha_o)}{\partial h}\left[h_{o,n}-h_o\right]-\frac{\partial\rho_{y_2}(\alpha_o)}{\partial g}\left[g_{o,n}-g_o\right]}{\left\|v^*_{y_2,n}\right\|_{sd}}\right| \\
&+\frac{1}{\left\|v^*_{y_2,n}\right\|_{sd}}\left[\left|\frac{\partial\rho_{y_2}\left(\alpha_o\right)}{\partial h}\left[h_{o,n}-h_o\right]\right|+\left|\frac{\partial\rho_{y_2}\left(\alpha_o\right)}{\partial h}\left[g_{o,n}-g_o\right]\right|\right] \\
= \;& o\left(n^{-\frac{1}{2}}\right) \qquad \text{(A.13)}
\end{aligned}$$

which combined with (A.12) implies

$$\frac{\sqrt{n}\left[\widehat{\rho}_{y_2}\left(\widehat{h}_n,\widehat{g}_n\right)-\rho_{y_2}\left(h_o,g_o\right)\right]}{\left\|v^*_{y_2,n}\right\|_{sd}}\xrightarrow{d}N\left(0,1\right).$$

Same thing goes for

$$\frac{\sqrt{n}\left[\widehat{\rho}_{y_3}\left(\widehat{h}_n,\widehat{g}_n\right)-\rho_{y_3}\left(h_o,g_o\right)\right]}{\left\|v^*_{y_3,n}\right\|_{sd}}\xrightarrow{d}N\left(0,1\right).$$

■

A.2 Assumptions and Proof of Results in Section 4.3

Proof. of Theorem 4.4

The following borrows from Hahn, Liao, and Ridder (2015) Appendix C.

ASSUMPTION B.4 *Let* $\mathcal{W}_{2,n}\equiv\left\{g\in\mathcal{V}_{2,n}:\|g\|_\psi\leq 1\right\}$*, then there is*

(a) $\langle v_{g_1},v_{g_2}\rangle_\psi=E\left\{-r_\psi\left(Z_2,\alpha_o\right)\left[v_{g_1},v_{g_2}\right]\right\}$ *for any* $v_{g_1},v_{g_2}\in\mathcal{V}_2$;

(b) $\sup_{\alpha\in\mathcal{N}_n,v_{g_1},v_{g_2}\in\mathcal{W}_{2,n}}\left|E\left\{r_\psi\left(Z_2,\alpha\right)\left[v_{g_1},v_{g_2}\right]-r_\psi\left(Z_2,\alpha_o\right)\left[v_{g_1},v_{g_2}\right]\right\}\right|=o(1)$

(c) $\sup_{\alpha\in\mathcal{N}_n,v_{g_1},v_{g_2}\in\mathcal{W}_{2,n}}\mu_n\left\{r_\psi\left(Z_2,\alpha\right)\left[v_{g_1},v_{g_2}\right]\right\}=o_p(1)$

(d) $\sup_{\alpha\in\mathcal{N}_n,v_g\in\mathcal{W}_{2,n}}\left|\frac{\partial\rho(\alpha)}{\partial g}\left[v_g\right]-\frac{\partial\rho(\alpha_o)}{\partial g}\left[v_g\right]\right|=o(1)$

Under Assumption B.4, the empirical Riesz representer $\widehat{v}_{g_n}^*$ satisfies:

$$\frac{\left\|\widehat{v}_{g_n}^*-v_{g_n}^*\right\|_\psi}{\left\|v_{g_n}^*\right\|_\psi}=o_p\left(1\right). \tag{A.14}$$

ASSUMPTION B.5 *Let* $\mathcal{W}_{1,n}\equiv\left\{h\in\mathcal{V}_{1,n}:\|h\|_\varphi\leq 1\right\}$, *and*
$\mathcal{B}_{2,n}^*\equiv\left\{v\in\mathcal{V}_{2,n}:\left\|v-v_{g_n}^*\right\|_\psi\left\|v_{g_n}^*\right\|_\psi^{-1}\leq\delta_{v_g,n}\right\}$, *where* $\delta_{v_g,n}=o(1)$ *is some positive sequence. Then*

(a) $\left\langle v_{h_1},v_{h_2}\right\rangle_\varphi=E\left\{-r_\varphi\left(Z_1,h\right)\left[v_{h_1},v_{h_2}\right]\right\}$ *for any* $v_{h_1},v_{h_2}\in\mathcal{V}_1$;

(b) $\sup_{h\in\mathcal{N}_{h,n},v_{h_1},v_{h_2}\in\mathcal{W}_{1,n}}\left|E\left\{r_\varphi\left(Z_1,h\right)\left[v_{h_1},v_{h_2}\right]-r_\varphi\left(Z_1,h\right)\left[v_{h_1},v_{h_2}\right]\right\}\right|=o(1)$

(c) $\sup_{h\in\mathcal{N}_{h,n},v_{h_1},v_{h_2}\in\mathcal{W}_{1,n}}\mu_n\left\{r_\varphi\left(Z_1,h\right)\left[v_{h_1},v_{h_2}\right]\right\}=o_p(1)$

(d) $\sup_{\alpha\in\mathcal{N}_n,v_h\in\mathcal{W}_{1,n}}\left|\frac{\partial\rho(\alpha)}{\partial h}\left[v_h\right]-\frac{\partial\rho(\alpha_o)}{\partial h}\left[v_h\right]\right|=o(1)$

(e) $\sup_{\alpha\in\mathcal{N}_n,v_g\in\mathcal{B}_{2,n}^*,v_h\in\mathcal{W}_{1,n}}\left|\Gamma_n\left(\alpha\right)\left[v_h,v_g\right]-\Gamma\left(\alpha_o\right)\left[v_h,v_{g_n}\right]\right|=o_p(1)$

Under Assumption B.5, the empirical Riesz representers $\widehat{v}_{h_n}^*$ and $\widehat{v}_{\Gamma_n}^*$

$$\frac{\left\|\widehat{v}_{h_n}^*-v_{h_n}^*\right\|_\varphi}{\left\|v_{h_n}^*\right\|_\varphi}=o_p\left(1\right)\text{ and }\frac{\left\|\widehat{v}_{\Gamma_n}^*-v_{\Gamma_n}^*\right\|_\varphi}{\left\|v_{\Gamma_n}^*\right\|_\varphi}=o_p\left(1\right) \tag{A.15}$$

ASSUMPTION B.6 *Let* $\|\cdot\|_2$ *denote the* $L_2\left(dF_Z\right)$*-norm, then:*

(a) *the functional* $\Delta_\varphi\left(Z_1,h\right)\left[v_h\right]$ *satisfies*

$$\begin{aligned}\sup_{h\in\mathcal{N}_{h,n},v_h\in\mathcal{W}_{1,n}}\left\|\Delta_\varphi\left(Z_1,h\right)\left[v_h\right]-\Delta_\varphi\left(Z_1,h_o\right)\left[v_h\right]\right\|_2 &= o(1)\\ \textit{and}\sup_{h\in\mathcal{N}_{h,n},v_h\in\mathcal{W}_{1,n}}\left|\mu_n\left\{\Delta_\varphi^2\left(Z_1,h\right)\left[v_h\right]\right\}\right| &= o_p(1)\end{aligned}$$

(b) *the functional* $\Delta_\psi(Z_2,\alpha_o)[v_g]$ *satisfies*

$$\begin{aligned}\sup_{\alpha\in\mathcal{N}_n,v_g\in\mathcal{W}_{2,n}}\|\Delta_\psi(Z_2,\alpha)[v_g]-\Delta_\psi(Z_2,\alpha_o)[v_g]\|_2 &= o(1)\\ \textit{and}\ \sup_{\alpha\in\mathcal{N}_n,v_g\in\mathcal{W}_{2,n}}\left|\mu_n\left\{\Delta_\psi^2(Z_2,\alpha)[v_g]\right\}\right| &= o_p(1)\end{aligned}$$

(c) *the following ULLN holds*

$$\sup_{\alpha\in\mathcal{N}_n,v_h\in\mathcal{W}_{1,n},v_g\in\mathcal{W}_{2,n}}\left|\mu_n\left\{\Delta_\varphi(Z_1,h)[v_h]\,\Delta_\psi(Z_2,\alpha)[v_g]\right\}\right|=o_p(1)$$

(d) $\sup_{v_h\in\mathcal{W}_{1,n}}\|\Delta_\varphi(Z_1,h_o)[v_h]\|_2=O(1)$ *and* $\sup_{v_g\in\mathcal{W}_{2,n}}\|\Delta_\psi(Z_2,\alpha_o)[v_g]\|_2=O(1)$*; moreover*

$$\frac{\left\|v_{h_n}^*\right\|_\varphi+\left\|v_{\Gamma_n}^*\right\|_\varphi+\left\|v_{g_n}^*\right\|_\psi}{\left\|\Delta_\varphi(Z_1,h_o)\left[v_{h_n}^*+v_{\Gamma_n}^*\right]+\Delta_\psi(Z_2,\alpha_o)\left[v_{g_n}^*\right]\right\|_2}=O(1)$$

By the triangle inequality, Holder inequality, Assumption B.6 (i) and (iv) we have

$$\begin{aligned}&\sup_{h\in\mathcal{N}_{h,n},v_h\in\mathcal{W}_{1,n}}\left|\frac{1}{n}\sum_{i=1}^n\Delta_\varphi^2(Z_{1,i},h)[v_h]-E\left[\Delta_\varphi^2(Z_{1,i},h_o)[v_h]\right]\right|\\ \leq\ &\sup_{h\in\mathcal{N}_{h,n},v_h\in\mathcal{W}_{1,n}}\left|\mu\left\{\Delta_\varphi^2(Z_{1,i},h)\right\}[v_h]\right|\\ +\ &\sup_{h\in\mathcal{N}_{h,n},v_h\in\mathcal{W}_{1,n}}\left|E\left[\Delta_\varphi^2(Z_{1,i},h)[v_h]\right]-\Delta_\varphi^2(Z_{1,i},h)[v_h]\right|\\ \leq\ &o_p(1)+\sup_{h\in\mathcal{N}_{h,n},v_h\in\mathcal{W}_{1,n}}\|\Delta_\varphi(Z_{1,i},h)[v_h]-\Delta_\varphi(Z_{1,i},h_o)[v_h]\|_2^2\\ &+2\sup_{h\in\mathcal{N}_{h,n},v_h\in\mathcal{W}_{1,n}}\|\Delta_\varphi(Z_{1,i},h)[v_h]-\Delta_\varphi(Z_{1,i},h_o)[v_h]\|_2\,\|\Delta_\varphi(Z_{1,i},h_o)[v_h]\|_2\\ =\ &o_p(1)\end{aligned}\tag{A.16}$$

Similarly, applying triangle inequaltiy and Holder inequality to Assumption B.6 (ii) and (iv),

$$\sup_{\alpha\in\mathcal{N}_n,v_g\in\mathcal{W}_{2,n}}\left|\frac{1}{n}\sum_{i=1}^n\Delta_\psi^2(Z_{2,i},\alpha)[v_g]-E\left[\Delta_\psi^2(Z_{2,i},\alpha_o)[v_g]\right]\right|=o_p(1).$$

By the triangle inequality, we have

$$\begin{aligned}&|E[\Delta_\varphi(Z_1,h)[v_h]\,\Delta_\varphi(Z_2,\alpha)[v_g]-\Delta_\varphi(Z_1,h_o)[v_h]\,\Delta_\psi(Z_2,\alpha_o)[v_g]]|\\ \leq\ &|E[(\Delta_\varphi(Z_1,h)[v_h]-\Delta_\varphi(Z_1,h_o)[v_h])(\Delta_\psi(Z_2,\alpha)[v_g]-\Delta_\psi(Z_2,\alpha_o)[v_g])]|\\ &+|E[(\Delta_\varphi(Z_1,h)[v_h]-\Delta_\varphi(Z_1,h_o)[v_h])\,\Delta_\psi(Z_2,\alpha_o)[v_g]]|\\ &+|E[\Delta_\varphi(Z_1,h_o)[v_h](\Delta_\psi(Z_2,\alpha)[v_g]-\Delta_\psi(Z_2,\alpha_o)[v_g])]|\end{aligned}$$

for any $\alpha \in \mathcal{N}_n$, $v_h \in \mathcal{W}_{1,n}$ and $v_g \in \mathcal{W}_{2,n}$. Using Holder inequality, Assumption B.6 (i) and (ii), we have for any $\alpha \in \mathcal{N}_n$, $v_h \in \mathcal{W}_{1,n}$ and $v_g \in \mathcal{W}_{2,n}$

$$\begin{aligned}
&\sup \left| E\left[\left(\Delta_\varphi\left(Z_1, h\right)[v_h] - \Delta_\varphi\left(Z_1, h_o\right)[v_h] \right) \left(\Delta_\varphi\left(Z_2, \alpha\right)[v_g] - \Delta_\psi\left(Z_2, \alpha_o\right)[v_g] \right) \right] \right| \\
&\leq \sup \left\| \Delta_\varphi\left(Z_1, h\right)[v_h] - \Delta_\varphi\left(Z_1, h_o\right)[v_h] \right\|_2 \left\| \Delta_\varphi\left(Z_2, \alpha\right)[v_g] - \Delta_\psi\left(Z_2, \alpha_o\right)[v_g] \right\|_2 \\
&= o\left(1\right)
\end{aligned}$$

Similarly by Assumption B.6 (i), (ii) and (iv), we have

$$\sup_{h \in \mathcal{N}_{h,n}, v_h \in \mathcal{W}_{1,n}, v_g \in \mathcal{W}_{2,n}} \left\| \Delta_\varphi\left(Z_1, h\right)[v_h] - \Delta_\varphi\left(Z_1, h_o\right)[v_h] \right\|_2 \left\| \Delta_\psi\left(Z_2, \alpha_o\right)[v_g] \right\|_2 = o\left(1\right) \tag{A.17}$$

and

$$\sup_{\alpha \in \mathcal{N}_n, v_h \in \mathcal{W}_{1,n}, v_g \in \mathcal{W}_{2,n}} \left\| \Delta_\psi\left(Z_2, \alpha\right)[v_h] - \Delta_\psi\left(Z_2, \alpha_o\right)[v_h] \right\|_2 \left\| \Delta_\varphi\left(Z_1, h_o\right)[v_h] \right\|_2 = o\left(1\right)$$

Thus we have

$$\begin{aligned}
&\sup_{\alpha \in \mathcal{N}_n, v_h \in \mathcal{W}_{1,n}, v_g \in \mathcal{W}_{2,n}} \left| E\left[\Delta_\varphi\left(Z_1, h\right)[v_h] \Delta_\varphi\left(Z_2, \alpha\right)[v_g] - \Delta_\varphi\left(Z_1, h_o\right)[v_h] \Delta_\psi\left(Z_2, \alpha_o\right)[v_g] \right] \right| \\
= \;& o_p\left(1\right)
\end{aligned} \tag{A.18}$$

Let c denote some generic finite positive constant. As the data is i.i.d, by definition, we have

$$\begin{aligned}
\left\| v_n^* \right\|_{sd}^2 &= Var\left[n^{-\frac{1}{2}} \sum_{i=1}^n \left(\Delta_\varphi\left(Z_{1,i}, h_o\right)\left[v_{h_n}^* + v_{\Gamma_n}^* \right] + \Delta_\psi\left(Z_{2,i}, \alpha_o\right)\left[v_{g_n}^* \right] \right) \right] \\
&= E\left[\left| \Delta_\varphi\left(Z_{1,i}, h_o\right)\left[v_{h_n}^* + v_{\Gamma_n}^* \right] + \Delta_\psi\left(Z_{2,i}, \alpha_o\right)\left[v_{g_n}^* \right] \right|^2 \right] \\
&= \left\| \Delta_\varphi\left(Z_{1,i}, h_o\right)\left[v_{h_n}^* + v_{\Gamma_n}^* \right] + \Delta_\psi\left(Z_{2,i}, \alpha_o\right)\left[v_{g_n}^* \right] \right\|_2^2
\end{aligned} \tag{A.19}$$

By the triangle inequality,

$$\begin{aligned}
\frac{\left\| \widehat{v}_n^* - v_n^* \right\|_{sd}}{\left\| v_n^* \right\|_{sd}} &\leq \frac{\left\| \Delta_\varphi\left(Z_1, h_o\right)\left[\widehat{v}_{h_n}^* - v_{h_n}^* \right] \right\|_2 + \left\| \Delta_\varphi\left(Z_1, h_o\right)\left[\widehat{v}_{\Gamma_n}^* - v_{\Gamma_n}^* \right] \right\|_2}{\left\| v_n^* \right\|_{sd}} \\
&+ \frac{\left\| \Delta_\psi\left(Z_2, \alpha_o\right)\left[\widehat{v}_{g_n}^* - v_{g_n}^* \right] \right\|_2}{\left\| v_n^* \right\|_{sd}}
\end{aligned} \tag{A.20}$$

Using (A.19), Assumpion B.6 (iv) and the result in (A.14), we have

$$\frac{\left\| \Delta_\psi\left(Z_2, \alpha_o\right)\left[\widehat{v}_{g_n}^* - v_{g_n}^* \right] \right\|_2}{\left\| v_n^* \right\|_{sd}} = \frac{\left\| v_{g_n}^* \right\|_\psi}{\left\| v_n^* \right\|_{sd}} \frac{\left\| \widehat{v}_{g_n}^* - v_{g_n}^* \right\|_\psi}{\left\| v_{g_n}^* \right\|_\psi} \left\| \Delta_\psi\left(Z_2, \alpha_o\right) \left[\frac{\widehat{v}_{g_n}^* - v_{g_n}^*}{\left\| \widehat{v}_{g_n}^* - v_{g_n}^* \right\|_\psi} \right] \right\|_2$$

and because $\left(\widehat{v}_{g_n}^* - v_{g_n}^*\right) / \left\|\widehat{v}_{g_n}^* - v_{g_n}^*\right\|_\psi \in \mathcal{W}_{g_n}$, we have

$$\frac{\left\|\Delta_\psi\left(Z_2, \alpha_o\right)\left[\widehat{v}_{g_n}^* - v_{g_n}^*\right]\right\|_2}{\left\|v_n^*\right\|_{sd}} = \frac{\left\|v_{g_n}^*\right\|_\psi}{\left\|v_n^*\right\|_{sd}} \frac{\left\|\widehat{v}_{g_n}^* - v_{g_n}^*\right\|_\psi}{\left\|v_{g_n}^*\right\|_\psi} \sup_{v_g \in \mathcal{W}_{g_n}} \left\|\Delta_\psi\left(Z_2, \alpha_o\right)[v_g]\right\|_2 = o_p(1) \tag{A.21}$$

Similarly, we have

$$\frac{\left\|\Delta_\varphi\left(Z_1, h_o\right)\left[\widehat{v}_{h_n}^* - v_{h_n}^*\right]\right\|_2 + \left\|\Delta_\varphi\left(Z_1, h_o\right)\left[\widehat{v}_{\Gamma_n}^* - v_{\Gamma_n}^*\right]\right\|_2}{\left\|v_n^*\right\|_{sd}} = o_p(1)$$

which together with (A.20) and (A.21) implies that

$$\frac{\left\|\widehat{v}_n^* - v_n^*\right\|_{sd}}{\left\|v_n^*\right\|_{sd}} = o_p(1). \tag{A.22}$$

Using (A.22) and the triangle inequality, we get

$$\begin{aligned} o_p(1) &= \frac{\left\|\widehat{v}_n^* - v_n^*\right\|_{sd}}{\left\|v_n^*\right\|_{sd}} \geq \left|\frac{\left\|\widehat{v}_n^*\right\|_{sd}}{\left\|v_n^*\right\|_{sd}} - 1\right| = \left|\frac{\left\|\widehat{v}_n^*\right\|_{sd}}{\left\|v_n^*\right\|_{n,sd}} \frac{\left\|v_n^*\right\|_{n,sd}}{\left\|\widehat{v}_n^*\right\|_{sd}} - 1\right| \\ &= \left|\frac{\left\|\widehat{v}_n^*\right\|_{sd}}{\left\|v_n^*\right\|_{n,sd}}\left(\frac{\left\|v_n^*\right\|_{n,sd}}{\left\|\widehat{v}_n^*\right\|_{sd}} - 1\right) + \left(\frac{\left\|\widehat{v}_n^*\right\|_{sd}}{\left\|v_n^*\right\|_{n,sd}} - 1\right)\right| \end{aligned} \tag{A.23}$$

We next show that $\frac{\left\|\widehat{v}_n^*\right\|_{sd}}{\left\|v_n^*\right\|_{n,sd}} - 1 = o_p(1)$. For this purpose, we first note that

$$\begin{aligned} \frac{\left\|\widehat{v}_{h_n}^* + \widehat{v}_{\Gamma_n}^*\right\|_\varphi}{\left\|\widehat{v}_n^*\right\|_{sd}} &\leq \frac{\left\|v_{h_n}^* + v_{\Gamma_n}^*\right\|_\varphi + \left\|\widehat{v}_{h_n}^* - v_{h_n}^*\right\|_\varphi + \left\|\widehat{v}_{\Gamma_n}^* - v_{\Gamma_n}^*\right\|_\varphi}{\left\|\widehat{v}_n^*\right\|_{sd}} \\ &\leq \frac{\left\|v_{h_n}^*\right\|_\varphi + \left\|v_{\Gamma_n}^*\right\|_\varphi + \left\|v_{h_n}^*\right\|_\varphi \frac{\left\|\widehat{v}_{h_n}^* - v_{h_n}^*\right\|_\varphi}{\left\|v_{h_n}^*\right\|_\varphi} + \left\|v_{\Gamma_n}^*\right\|_\varphi \frac{\left\|\widehat{v}_{\Gamma_n}^* - v_{\Gamma_n}^*\right\|_\varphi}{\left\|v_{\Gamma_n}^*\right\|_\varphi}}{\left\|\widehat{v}_n^*\right\|_{sd}} \\ &= \frac{\left\|v_{h_n}^*\right\|_\varphi}{\left\|\widehat{v}_n^*\right\|_{sd}}\left(1 + \frac{\left\|\widehat{v}_{h_n}^* - v_{h_n}^*\right\|_\varphi}{\left\|v_{h_n}^*\right\|_\varphi}\right) + \frac{\left\|v_{\Gamma_n}^*\right\|_\varphi}{\left\|\widehat{v}_n^*\right\|_{sd}}\left(1 + \frac{\left\|\widehat{v}_{\Gamma_n}^* - v_{\Gamma_n}^*\right\|_\varphi}{\left\|v_{\Gamma_n}^*\right\|_\varphi}\right) \\ &= \frac{\left\|v_n^*\right\|_{sd}}{\left\|\widehat{v}_n^*\right\|_{sd}}\left[\frac{\left\|v_{h_n}^*\right\|_\varphi}{\left\|\widehat{v}_n^*\right\|_{sd}}\left(1 + o_p(1)\right) + \frac{\left\|v_{\Gamma_n}^*\right\|_\varphi}{\left\|\widehat{v}_n^*\right\|_{sd}}\left(1 + o_p(1)\right)\right] \\ &= \frac{\left\|v_n^*\right\|_{sd}}{\left\|\widehat{v}_n^*\right\|_{sd}} O_p(1) = \left(\frac{1}{\left\|\widehat{v}_n^*\right\|_{sd} / \left\|v_n^*\right\|_{sd}} - 1\right) O_p(1) + O_p(1) = O_p(1) \end{aligned} \tag{A.24}$$

where the first two inequalities are by the triangle inequality, the second equality is by (A.15), the third equality is by (A.19) and Assumption B.6(iv), and the last equality is by the first inequality in (A.23). Similarly, we can show that

$$\left\|v_{g_n}^*\right\|_\psi \left\|\widehat{v}_n^*\right\|_{sd}^{-1} = O_p(1). \tag{A.25}$$

By the triangle inequality, we get

$$\begin{aligned}\left|\frac{\|\widehat{v}_n^*\|_{n,sd}^2-\|\widehat{v}_n^*\|_{sd}^2}{\|\widehat{v}_n^*\|_{sd}^2}\right| &= \left|\frac{\frac{1}{n}\sum_{i=1}^n\left[\Delta_\varphi\left(Z_{1,i},\widehat{h}_n\right)\left[\widehat{v}_{h_n}^*+\widehat{v}_{\Gamma_n}^*\right]+\Delta_\psi\left(Z_{2,i},\widehat{\alpha}_n\right)\left[v_{g_n}^*\right]\right]^2}{\|\widehat{v}_n^*\|_{sd}^2}\right. \\ &\quad \left.-\frac{E_Z\left[\Delta_\varphi\left(Z_{1,i},h_o\right)\left[\widehat{v}_{h_n}^*+\widehat{v}_{\Gamma_n}^*\right]+\Delta_\psi\left(Z_{2,i},\alpha_o\right)\left[v_{g_n}^*\right]\right]}{\|\widehat{v}_n^*\|_{sd}^2}\right| \\ &\leq |I_{1,n}|+|I_{2,n}|+2\,|I_{3,n}| \end{aligned} \tag{A.26}$$

where $E_Z[\cdot]$ denotes the expectation taking with respect to the distribution of Z ($E_{Z_1}[\cdot]$ and $E_{Z_2}[\cdot]$ are similarly defined),

$$\begin{aligned} I_{1,n} &= \frac{\frac{1}{n}\sum_{i=1}^n\Delta_\varphi^2\left(Z_{1,i},\widehat{h}_n\right)\left[\widehat{v}_{h_n}^*+\widehat{v}_{\Gamma_n}^*\right]-E_{Z_1}\left[\Delta_\varphi^2\left(Z_{1,i},h_o\right)\left[\widehat{v}_{h_n}^*+\widehat{v}_{\Gamma_n}^*\right]\right]}{\|\widehat{v}_n^*\|_{sd}^2}, \\ I_{2,n} &= \frac{\frac{1}{n}\sum_{i=1}^n\Delta_\psi^2\left(Z_{2,i},\widehat{\alpha}_n\right)\left[\widehat{v}_{g_n}^*\right]-E_{Z_1}\left[\Delta_\psi^2\left(Z_{2,i},\alpha_o\right)\left[\widehat{v}_{g_n}^*\right]\right]}{\|\widehat{v}_n^*\|_{sd}^2} \\ I_{3,n} &= \frac{\frac{1}{n}\sum_{i=1}^n\Delta_\varphi\left(Z_{1,i},\widehat{h}_n\right)\left[\widehat{v}_{h_n}^*+\widehat{v}_{\Gamma_n}^*\right]\Delta_\psi\left(Z_{2,i},\widehat{\alpha}_n\right)\left[\widehat{v}_{g_n}^*\right]}{\|\widehat{v}_n^*\|_{sd}^2} \\ &\quad -\frac{E_Z\left[\Delta_\varphi\left(Z_{1,i},\widehat{h}_n\right)\left[\widehat{v}_{h_n}^*+\widehat{v}_{\Gamma_n}^*\right]\Delta_\psi\left(Z_{2,i},\widehat{\alpha}_n\right)\left[\widehat{v}_{g_n}^*\right]\right]}{\|\widehat{v}_n^*\|_{sd}^2} \end{aligned}$$

By (A.16) and (A.24), we have

$$\begin{aligned} |I_{1,n}| &= \frac{\left\|\widehat{v}_{h_n}^*+\widehat{v}_{\Gamma_n}^*\right\|_\varphi^2}{\|\widehat{v}_n^*\|_{sd}^2}\frac{\left|\frac{1}{n}\sum_{i=1}^n\Delta_\varphi^2\left(Z_{1,i},\widehat{h}_n\right)\left[\widehat{v}_{h_n}^*+\widehat{v}_{\Gamma_n}^*\right]-E_{Z_1}\left[\Delta_\varphi^2\left(Z_{1,i},h_o\right)\left[\widehat{v}_{h_n}^*+\widehat{v}_{\Gamma_n}^*\right]\right]\right|}{\left\|\widehat{v}_{h_n}^*+\widehat{v}_{\Gamma_n}^*\right\|_\varphi^2} \\ &\leq O_p(1)\sup_{h\in\mathcal{N}_{h,n},v_h\in\mathcal{W}_{1,n}}\left|\frac{1}{n}\sum_{i=1}^n\Delta_\varphi^2\left(Z_{1,i},h\right)[v_h]-E_{Z_1}\left[\Delta_\varphi^2\left(Z_{1,i},h_o\right)[v_h]\right]\right|=o_p(1) \end{aligned} \tag{A.27}$$

Similarly, by (A.17) and (A.25), we have

$$|I_{2,n}|=\frac{\left\|\widehat{v}_{g_n}^*\right\|_\psi^2}{\|\widehat{v}_n^*\|_{sd}^2}\frac{\left|\frac{1}{n}\sum_{i=1}^n\Delta_\psi^2\left(Z_{2,i},\widehat{\alpha}_n\right)\left[\widehat{v}_{g_n}^*\right]-E_{z_2}\left[\Delta_\varphi^2\left(Z_{2,i},\alpha_o\right)\left[\widehat{v}_{g_n}^*\right]\right]\right|}{\left\|\widehat{v}_{g_n}^*\right\|_\psi^2}=o_p(1) \tag{A.28}$$

For the last term $I_{3,n}$, note that for any $\alpha \in \mathcal{N}_n$, $v_h \in \mathcal{W}_{1,n}$ and $v_g \in \mathcal{W}_{2,n}$

$$\begin{aligned}
|I_{3,n}| &\leq \frac{\left|\mu_n\left\{\Delta_\varphi\left(Z_1, \widehat{h}_n\right)\left[\widehat{v}^*_{h_n} + \widehat{v}^*_{\Gamma_n}\right]\Delta_\psi\left(Z_2, \widehat{\alpha}_n\right)\left[\widehat{v}^*_{g_n}\right]\right\}\right|}{\|\widehat{v}^*_n\|^2_{sd}} \\
&\quad + \frac{\left|E_Z\left[\Delta_\varphi\left(Z_1, \widehat{h}_n\right)\left[\widehat{v}^*_{h_n} + \widehat{v}^*_{\Gamma_n}\right]\Delta_\psi\left(Z_2, \widehat{\alpha}_n\right)\left[\widehat{v}^*_{g_n}\right]\right.\right.}{\|\widehat{v}^*_n\|^2_{sd}} \\
&\quad \frac{\left.\left.-\Delta_\varphi\left(Z_1, h_o\right)\left[\widehat{v}^*_{h_n} + \widehat{v}^*_{\Gamma_n}\right]\Delta_\psi\left(Z_2, \alpha_o\right)\left[\widehat{v}^*_{g_n}\right]\right]\right|}{\|\widehat{v}^*_n\|^2_{sd}} \\
&\leq \frac{\left\|\widehat{v}^*_{h_n} + \widehat{v}^*_{\Gamma_n}\right\|_\varphi \left\|\widehat{v}^*_{g_n}\right\|_\psi}{\|\widehat{v}^*_n\|^2_{sd}} \left[\sup\left|\mu_n\left\{\Delta_\varphi\left(Z_1, h\right)\left[v_h\right]\Delta_\psi\left(Z_2, \alpha_o\right)\left[v_g\right]\right\}\right|\right. \\
&\quad \left.+\sup\left|E\left[\Delta_\varphi\left(Z_1, h\right)\left[v_h\right]\Delta_\psi\left(Z_2, \alpha\right)\left[v_g\right] - \Delta_\varphi\left(Z_1, h_o\right)\left[v_h\right]\Delta_\psi\left(Z_2, \alpha_o\right)\left[v_g\right]\right]\right|\right] \\
&= \frac{\left\|\widehat{v}^*_{h_n} + \widehat{v}^*_{\Gamma_n}\right\|_\varphi \left\|\widehat{v}^*_{g_n}\right\|_\psi}{\|\widehat{v}^*_n\|^2_{sd}} o_p\left(1\right) = o_p\left(1\right) \qquad \text{(A.29)}
\end{aligned}$$

where the first inequality is by the triangle inequality, the first equality is by Assumption B.6 (iii) and (A.18), the last equality is by (C18) and (C19). From the results in (A.26), (A.27), (A.28) and (A.29), we deduce that

$$\left|\frac{\|\widehat{v}^*_n\|^2_{n,sd} - \|\widehat{v}^*_n\|^2_{sd}}{\|\widehat{v}^*_n\|^2_{sd}}\right| = o_p\left(1\right) \qquad \text{(A.30)}$$

It is clear that (A.23) and (C.24) imply that $\left|\|\widehat{v}^*_n\|_{n,sd} / \|\widehat{v}^*_n\|^2_{sd} - 1\right| = o_p\left(1\right)$, which finishes the proof. ■

A.3 Proof for Results in Section 5

Proof. of Proposition 5.1.

The following aims to derive the asymptotics for $\sqrt{n}\left[\widehat{\pi}_{y_2}\left(\widehat{\delta}_h, \widehat{\delta}_g\right) - \pi_{y_2}\left(\delta_{h_o}, \delta_{g_o}\right)\right]$ and $\sqrt{n}\left[\widehat{\pi}_{y_3}\left(\widehat{\delta}_h, \widehat{\delta}_g\right) - \pi_{y_3}\left(\delta_{h_o}, \delta_{g_o}\right)\right]$. Notice that we have

$$\begin{aligned}
\pi_{y_2}\left(\delta_{h_o}, \delta_{g_o}\right) &= E_{\mathbf{w}_{i1}}\left[r_{y_2}\left(\mathbf{w}_{i1}; \delta_{h_o}, \delta_{g_o}\right)\right], \\
\pi_{y_3}\left(\delta_{h_o}, \delta_{g_o}\right) &= E_{\mathbf{w}_{i1}}\left[r_{y_3}\left(\mathbf{w}_{i1}; \delta_{h_o}, \delta_{g_o}\right)\right].
\end{aligned}$$

and

$$\widehat{\pi}_{y_2}\left(\widehat{\delta}_h,\widehat{\delta}_g\right)=n^{-1}\sum_{i=1}^{n}r_{i,y_2}\left(\widehat{\delta}_h,\widehat{\delta}_g\right),$$
$$\widehat{\pi}_{y_3}\left(\widehat{\delta}_h,\widehat{\delta}_g\right)=n^{-1}\sum_{i=1}^{n}r_{i,y_3}\left(\widehat{\delta}_h,\widehat{\delta}_g\right).$$

Now we could focus on the property of $r_{i,y_2}\left(\widehat{\delta}_h,\widehat{\delta}_g\right)$ instead to derive the asymptotics for $\widehat{\pi}_{y_2}\left(\widehat{\delta}_h,\widehat{\delta}_g\right)$.

By the mean value expansion,

$n^{-\frac{1}{2}}\sum_{i=1}^{n}r_{i,y_2}\left(\widehat{\delta}_h,\widehat{\delta}_g\right)=$

$n^{-\frac{1}{2}}\sum_{i=1}^{n}r_{i,y_2}\left(\delta_{h_o},\delta_{g_o}\right)+n^{-1}\sum_{i=1}^{n}\ddot{R}_h\sqrt{n}\left(\widehat{\delta}_h-\delta_{h_o}\right)+n^{-1}\sum_{i=1}^{n}\ddot{R}_g\sqrt{n}\left(\widehat{\delta}_g-\delta_{g_o}\right)$, where

$$\ddot{R}_{i,h}^{y2}\equiv\bigtriangledown_{\delta_h}r_{i,y_2}\left(\ddot{\delta}_h,\ddot{\delta}_g\right),$$
$$\ddot{R}_{i,g}^{y2}\equiv\bigtriangledown_{\delta_g}r_{i,y_3}\left(\ddot{\delta}_h,\ddot{\delta}_g\right).$$

We know $\left(\ddot{\delta}_h,\ddot{\delta}_g\right)$ is "trapped" between $\left(\hat{\delta}_h,\hat{\delta}_g\right)$ and $(\delta_{h_o},\delta_{g_o})$. As we have $\left(\hat{\delta}_h,\hat{\delta}_g\right)\xrightarrow{p}(\delta_{h_o},\delta_{g_o})$, it is easy to see $\left(\ddot{\delta}_h,\ddot{\delta}_g\right)$ converges in probability to $(\delta_{h_o},\delta_{g_o})$, too.

After verifying some regularity conditions, we have

$$n^{-1}\sum_{i=1}^{n}\ddot{R}_{i,h}^{y_2}\xrightarrow{p}R_{o,h}^{y_2}\equiv E\left[\bigtriangledown_{\delta_h}r_{i,y_2}\left(\delta_{h_o},\delta_{g_o}\right)\right],$$
$$n^{-1}\sum_{i=1}^{n}\ddot{R}_{i,g}^{y_2}\xrightarrow{p}R_{o,g}^{y_2}\equiv E\left[\bigtriangledown_{\delta_g}r_{i,y_2}\left(\delta_{h_o},\delta_{g_o}\right)\right].$$

The first step $\widehat{\delta}_h$ is a least squares estimator, by standard arguments, which has an influence function taking the form,

$$\sqrt{n}\left(\widehat{\delta}_h-\delta_{h_o}\right)=-n^{-\frac{1}{2}}\sum_{i=1}^{n}H_{o,h}^{-1}S_{i,h}+o_p(1)$$

where $H_{o,h}\equiv E\left(H_{i,h}\right)\equiv E\left[p_1^{k(n)}(\mathbf{z}_i)p_1^{k(n)}(\mathbf{z}_i)'\right]$ and $S_{i,h}\equiv p_1^{k(n)}(\mathbf{z}_i)\left(y_{i2}-p_1^{k(n)}(\mathbf{z}_i)'\delta_{h_o}\right)$.

Its asymptotic distribution for $\sqrt{n}\left(\widehat{\delta}_h-\delta_{h_o}\right)$ is

$$\sqrt{n}\left(\widehat{\delta}_h-\delta_{h_o}\right)\xrightarrow{d}N(0,V_h).$$

where $V_h\equiv H_{o,h}^{-1}Var\left(S_{i,h}\right)H_{o,h}^{-1}$ has the heteroskedastic robust sandwich form.

The second step $\hat{\delta}_g$ is a maximum likelihood estimator of the bivariate probit, with the first step $\widehat{\delta}_h$

plugged in. Absorb the randomness of all the data $\mathbf{w}_i \equiv (\mathbf{w}_{i1}, \mathbf{w}_{i2})$ into the subscript, use a standard delta method for a function contains two-step sieve M-estimation (Wooldridge, 2010, Chapter 12, for example), the influence function is

$$\sqrt{n}\left(\widehat{\delta}_g - \delta_{g_o}\right) = -n^{-\frac{1}{2}} \sum_{i=1}^{n} H_{o,g}^{-1}\left(S_{i,g} + F_{o,gh} H_{o,h}^{-1} S_{i,h}\right) + o_p(1).$$

$\Phi_{i,2}^{(j)}$ is the partial derivative of bivariate CDF

$$\Phi_{i,2}\left[d_{i1}\left(\mathbf{x}_{i1}\boldsymbol{\beta} + m_i'\delta_{m_o}\right), d_{i2}\left(\mathbf{z}_i\boldsymbol{\gamma} + m_i'\delta_{q_o}\right), d_{i1}d_{i2}\rho\right]$$

with respect to its j-th argument, where $j = 1, 2, 3$, and m_i is the shorthand for

$$m_i \equiv p_2^{l(n)}\left[y_{i2} - p_1^{k(n)}(\mathbf{z}_i)'\delta_{h_o}\right].$$

$\Phi_{i,2}^{(j,k)}$ is the cross derivative with respect to its j-th and k-th arguments, $j, k = 1, 2, 3$,

$$S_{i,g} \equiv \frac{\nabla_{\delta_g}\Phi_{i,2}}{\Phi_{i,2}} = \frac{1}{\Phi_{i,2}} \begin{pmatrix} \frac{\partial \Phi_{i,2}}{\partial \beta} \\ \frac{\partial \Phi_{i,2}}{\partial \gamma} \\ \frac{\partial \Phi_{i,2}}{\partial \rho} \\ \frac{\partial \Phi_{i,2}}{\partial \delta_{m_o}} \\ \frac{\partial \Phi_{i,2}}{\partial \delta_{q_o}} \end{pmatrix} = \frac{1}{\Phi_{i,2}} \begin{pmatrix} \mathbf{x}_{i1}' d_{i1} \Phi_{i,2}^{(1)} \\ \mathbf{z}_i' d_{i2} \Phi_{i,2}^{(2)} \\ d_{i1}d_{i2}\Phi_{i,2}^{(3)} \\ m_i d_{i1} \Phi_{i,2}^{(1)} \\ m_i d_{i2} \Phi_{i,2}^{(2)} \end{pmatrix},$$

$H_{o,g} \equiv E\left(H_{i,g}\right) \equiv E\left(\nabla_{\delta_g} S_{i,g}\right)$,

$\nabla_{\delta_g} S_{i,g} = \frac{1}{\Phi_{i,2}}$

$$\begin{pmatrix} \mathbf{x}_{i1}'\mathbf{x}_{i1} d_{i1} \Phi_{i,2}^{(11)} & \mathbf{x}_{i1}'\mathbf{z}_i d_{i1}d_{i2}\Phi_{i,2}^{(12)} & \mathbf{x}_{i1}' d_{i2} d_{i1} \Phi_{i,2}^{(13)} & \mathbf{x}_{i1}' m_i \prime d_{i1} \Phi_{i,2}^{(11)} & \mathbf{x}_{i1}' m_i \prime d_{i1} d_{i2} \Phi_{i,2}^{(12)} \\ \mathbf{z}_i'\mathbf{x}_{i1} d_{i1}d_{i2}\Phi_{i,2}^{(21)} & \mathbf{z}_i'\mathbf{z}_i d_{i2}\Phi_{i,2}^{(22)} & \mathbf{z}_i' d_{i2} d_{i1} \Phi_{i,2}^{(23)} & \mathbf{z}_i' m_i \prime d_{i1} d_{i2} \Phi_{i,2}^{(21)} & \mathbf{z}_i' m_i' d_{i2} \Phi_{i,2}^{(22)} \\ \mathbf{x}_{i1} d_{i1}d_{i2}\Phi_{i,2}^{(31)} & \mathbf{z}_i d_{i1}d_{i2}\Phi_{i,2}^{(32)} & d_{i1}d_{i2}\Phi_{i,2}^{(33)} & m_i' d_{i1} d_{i2} \Phi_{i,2}^{(31)} & m_i' d_{i1} d_{i2} \Phi_{i,2}^{(32)} \\ m_i\mathbf{x}_{i1} d_{i1} \Phi_{i,2}^{(11)} & m_i\mathbf{z}_i d_{i1}d_{i2}\Phi_{i,2}^{(12)} & m_i d_{i1} d_{i2} \Phi_{i,2}^{(13)} & m_i m_i' d_{i1} \Phi_{i,2}^{(11)} & m_i m_i' d_{i1} d_{i2} \Phi_{i,2}^{(12)} \\ m_i\mathbf{x}_{i1} d_{i2} \Phi_{i,2}^{(21)} & m_i\mathbf{z} d_{i2}\Phi_{i,2}^{(22)} & m_i d_{i1} d_{i2} \Phi_{i,2}^{(23)} & m_i m_i' d_{i2} d_{i1} \Phi_{i,2}^{(21)} & m_i m_i' d_{i2} \Phi_{i,2}^{(22)} \end{pmatrix},$$

$F_{o,gh} \equiv E\left(F_{i,gh}\right) \equiv E\left(\nabla_{\delta_h} S_{i,g}\right)$,

$p_2^{l(n)(1)}(\cdot) = (p_{2,1}^{(1)}(\cdot)', ..., p_{2,l(n)}^{(1)}(\cdot)')'$ is the first order derivative for each element of the basis function, shorthand $m_i^{(1)} \equiv p_2^{l(n)(1)}\left[y_{i2} - p_1^{k(n)}(\mathbf{z}_i)'\delta_{h_o}\right]$,

$\nabla_{\delta_h} S_{i,g} = \frac{1}{\Phi_{i,2}}$

$$\begin{pmatrix} -\mathbf{x}_{i1}'\delta_{m_o}' m_i^{(1)} p_1^{k(n)}(\mathbf{z}_i)' d_{i1}\Phi_{i,2}^{(11)} - \mathbf{x}_{i1}'\delta_{q_o}' m_i^{(1)} p_1^{k(n)}(\mathbf{z}_i)' d_{i1}d_{i2}\Phi_{i,2}^{(12)} \\ -\mathbf{z}_i'\delta_{m_o}' m_i^{(1)} p_1^{k(n)}(\mathbf{z}_i)' d_{i1}d_{i2}\Phi_{i,2}^{(21)} - \mathbf{z}_i'\delta_{q_o}' m_i^{(1)} p_1^{k(n)}(\mathbf{z}_i)' d_{i2}\Phi_{i,2}^{(22)} \\ -\delta_{m_o}' m_i^{(1)} p_1^{k(n)}(\mathbf{z}_i)' d_{i1}d_{i2}\Phi_{i,2}^{(31)} - \delta_{m_o}' m_i^{(1)} p_1^{k(n)}(\mathbf{z}_i)' d_{i1}d_{i2}\Phi_{i,2}^{(32)} \\ -m_i^{(1)} p_1^{k(n)}(\mathbf{z}_i)' d_{i1}\Phi_{i,2}^{(1)} - m_i\delta_{m_o}' m_i^{(1)} p_1^{k(n)}(\mathbf{z}_i)' d_{i1}\Phi_{i,2}^{(11)} - m_i\delta_{q_o}' m_i^{(1)} p_1^{k(n)}(\mathbf{z}_i)' d_{i1}d_{i2}\Phi_{i,2}^{(12)} \\ -m_i^{(1)} p_1^{k(n)}(\mathbf{z}_i)' d_{i2}\Phi_{i,2}^{(2)} - m_i\delta_{m_o}' m_i^{(1)} p_1^{k(n)}(\mathbf{z}_i)' d_{i1}d_{i2}\Phi_{i,2}^{(21)} - m_i\delta_{q_o}' m_i^{(1)} p_1^{k(n)}(\mathbf{z}_i)' d_{i2}\Phi_{i,2}^{(22)} \end{pmatrix}.$$

The asymptotic distribution for $\sqrt{n}\left(\hat{\delta}_g - \delta_{g_o}\right)$ then is

$$\sqrt{n}\left(\hat{\delta}_g - \delta_{g_o}\right) \xrightarrow{d} N(0, V_g),$$

where $V_g \equiv H_{o,g}^{-1} D_o H_{o,g}^{-1}$ and $D_o \equiv Var\left(S_{i,g} + F_{o,gh}H_{o,h}^{-1}S_{i,h}\right)$.

The influence function for $\sqrt{n}\left[\widehat{\pi}_{y_2}\left(\widehat{\delta}_h, \widehat{\delta}_g\right) - \pi_{y_2}\left(\delta_{h_o}, \delta_{g_o}\right)\right]$ then has three components

$$\begin{aligned} &\sqrt{n}\left[\widehat{\pi}_{y_2}\left(\widehat{\delta}_h, \widehat{\delta}_g\right) - \pi_{y_2}\left(\delta_{h_o}, \delta_{g_o}\right)\right] \\ &= n^{-\frac{1}{2}}\sum_{i=1}^{n}\left[r_{i,y_2}\left(\widehat{\delta}_h, \widehat{\delta}_g\right) - \pi_{y_2}\left(\delta_{h_o}, \delta_{g_o}\right)\right] + o_p(1) \\ &= n^{-\frac{1}{2}}\sum_{i=1}^{n}\left[r_{i,y_2}\left(\delta_{h_o}, \delta_{g_o}\right) - \pi_{y_2}\left(\delta_{h_o}, \delta_{g_o}\right) - R_{o,h}^{y_2}H_{o,h}^{-1}S_{i,h} - R_{o,g}^{y_2}H_{o,g}^{-1}\left(S_{i,g} + F_{o,gh}H_{o,h}^{-1}S_{i,h}\right)\right] \\ &+ o_p(1). \end{aligned}$$

Therefore,

$$\sqrt{n}\left[\widehat{\pi}_{y_2}\left(\widehat{\delta}_h, \widehat{\delta}_g\right) - \pi_{y_2}\left(\delta_{h_o}, \delta_{g_o}\right)\right] \xrightarrow{d} N(0, V_{y_2}),$$

where

$$V_{y_2} = Var\left[r_{i,y_2}\left(\delta_{h_o}, \delta_{g_o}\right) - \pi_{y_2}\left(\delta_{h_o}, \delta_{g_o}\right) - R_{o,h}^{y_2}H_{o,h}^{-1}S_{i,h} - R_{o,g}^{y_2}H_{o,g}^{-1}\left(S_{i,g} + F_{o,gh}H_{o,h}^{-1}S_{i,h}\right)\right].$$

Similarly, follow exactly the same procedure, but with a slight change of notation, we have

$$\sqrt{n}\left[\widehat{\pi}_{y_3}\left(\widehat{\delta}_h, \widehat{\delta}_g\right) - \pi_{y_3}\left(\delta_{h_o}, \delta_{g_o}\right)\right] \xrightarrow{d} N(0, V_{y_3}),$$

where

$$V_{y_3} = Var\left[r_{i,y_3}\left(\delta_{h_o}, \delta_{g_o}\right) - \pi_{y_3}\left(\delta_{h_o}, \delta_{g_o}\right) - R_{o,h}^{y_3}H_{o,h}^{-1}S_{i,h} - R_{o,g}^{y_3}H_{o,g}^{-1}\left(S_{i,g} + F_{o,gh}H_{o,h}^{-1}S_{i,h}\right)\right]. \blacksquare$$

A.4 Figures and Tables for Section 6

Figure 1: Empirical Distribution For Design 1 with $z_1 \sim \text{Normal}(0, 9)$

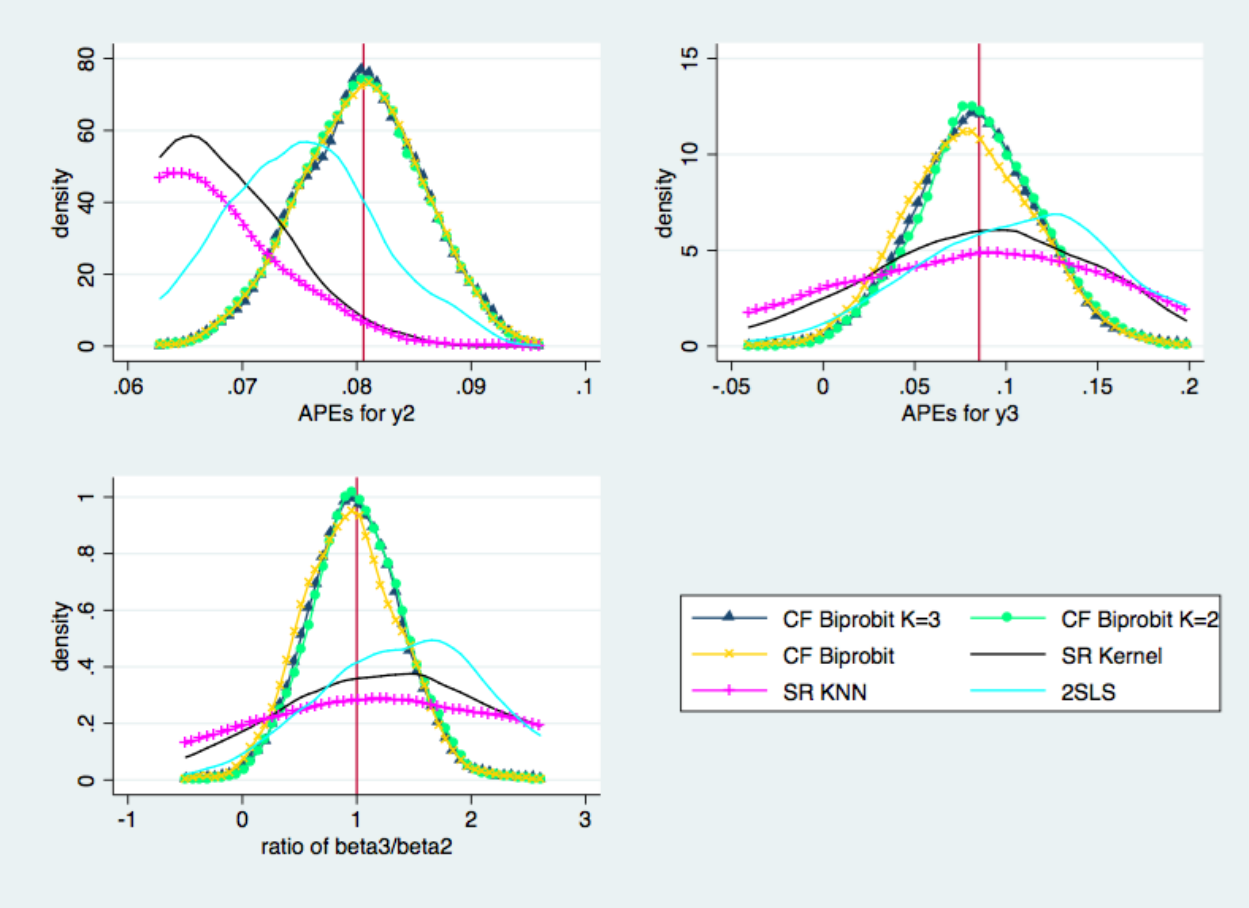

Figure 2: Empirical Distribution For Design 1 with $z_1 \sim \text{Normal}(0, 1)$

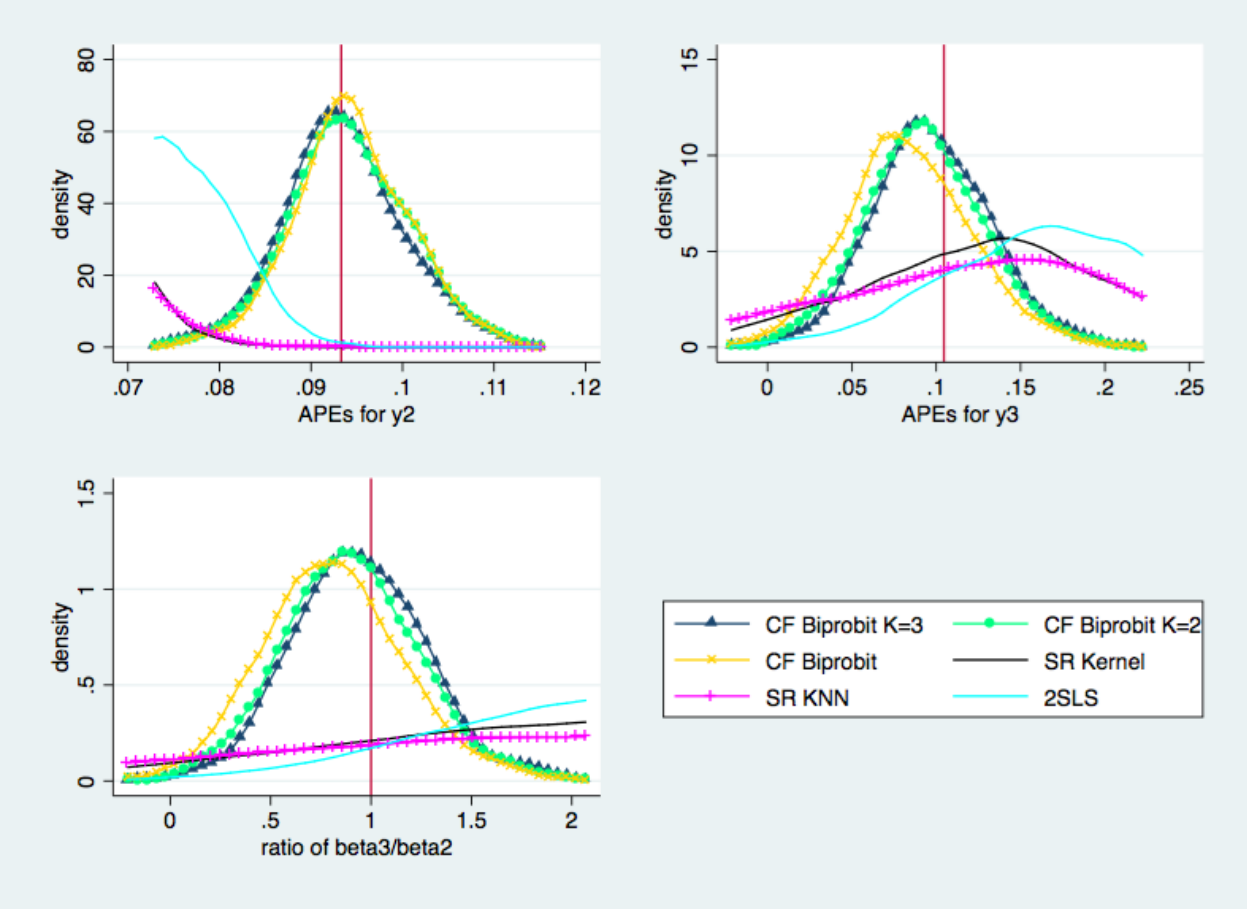

Figure 3: Empirical Distribution For Design 2 with $z_1 \sim \text{Normal}(0, 9)$

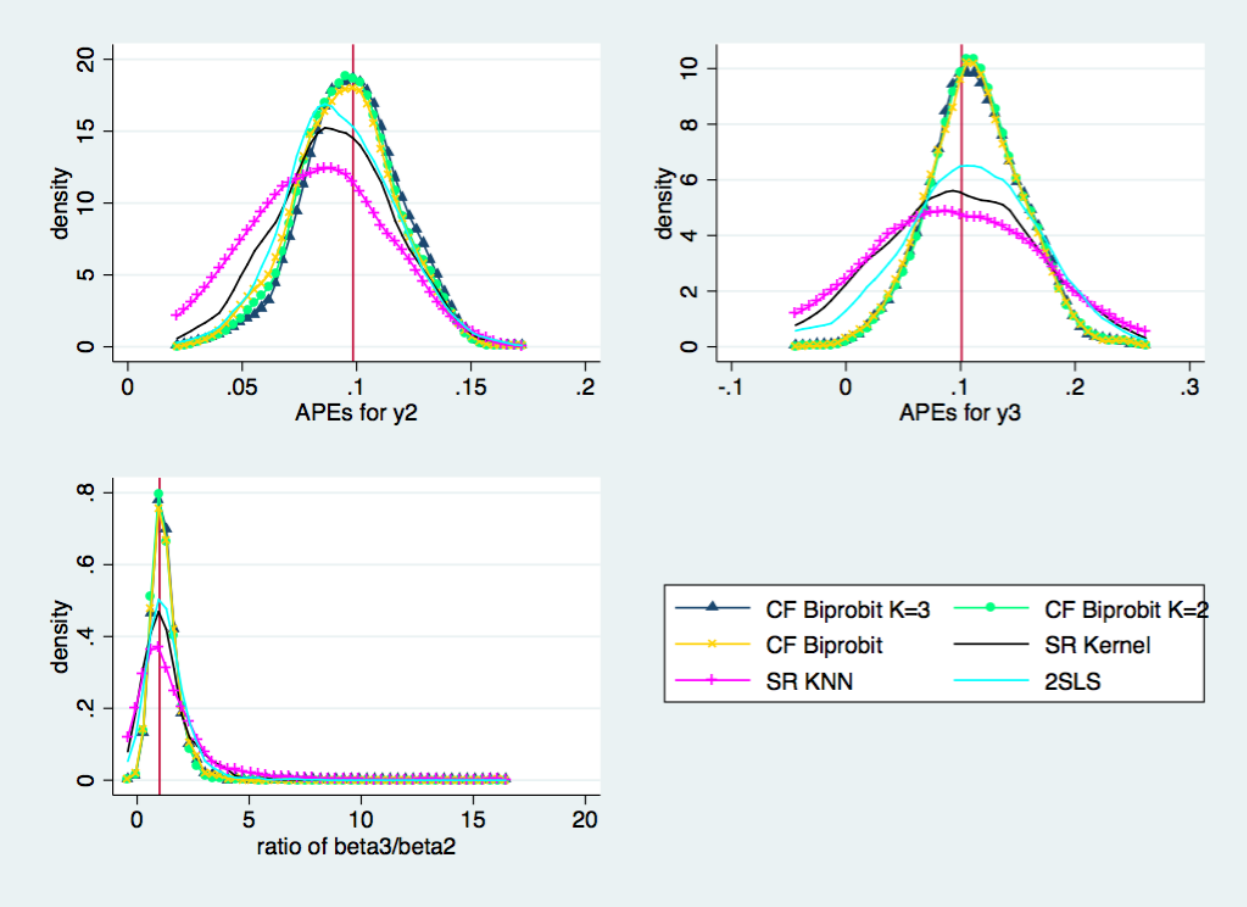

Figure 4: Empirical Distribution For Design 2 with $z_1 \sim \text{Normal}(0, 16)$

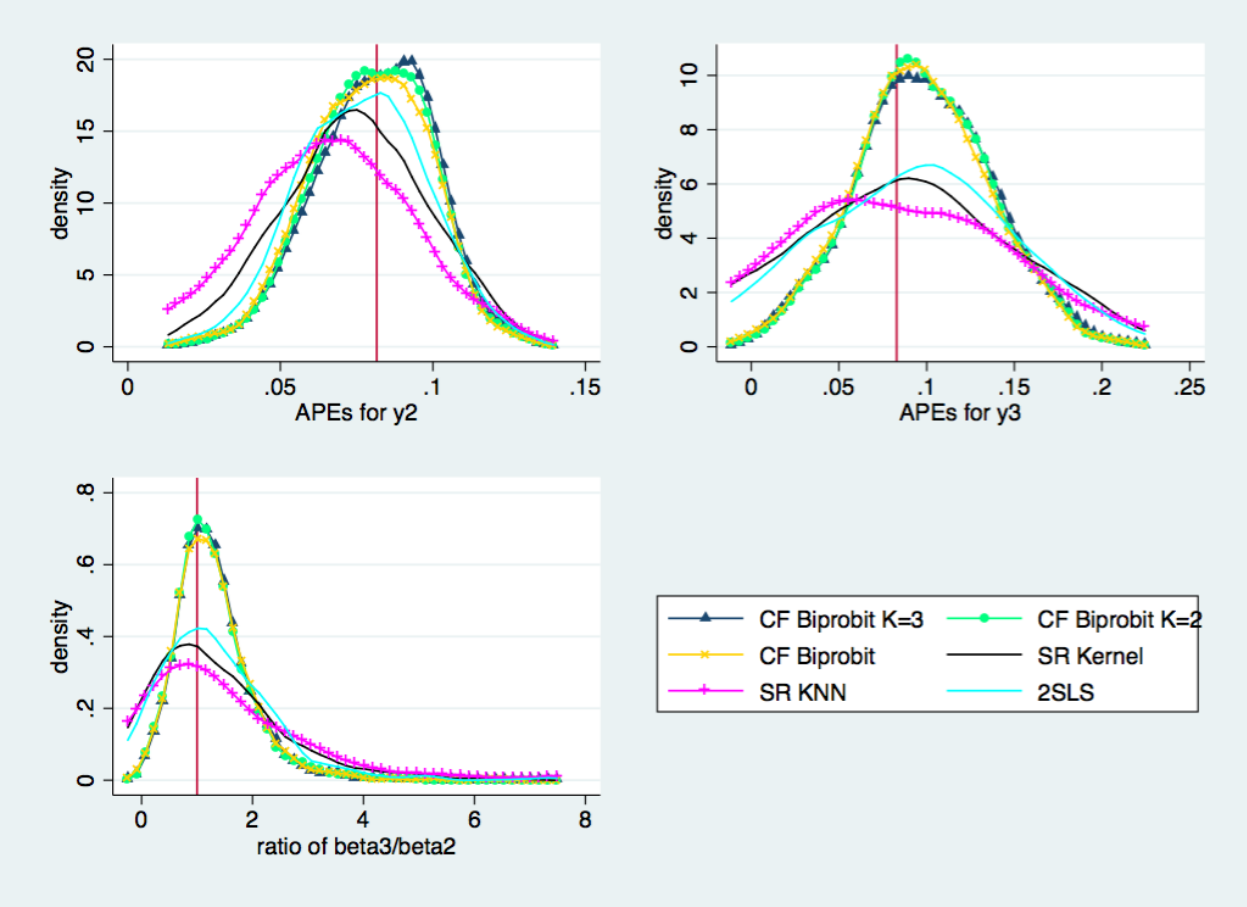

Figure 5: Empirical Distribution For Design 3 with $z_1 \sim \text{Normal}(0, 9)$

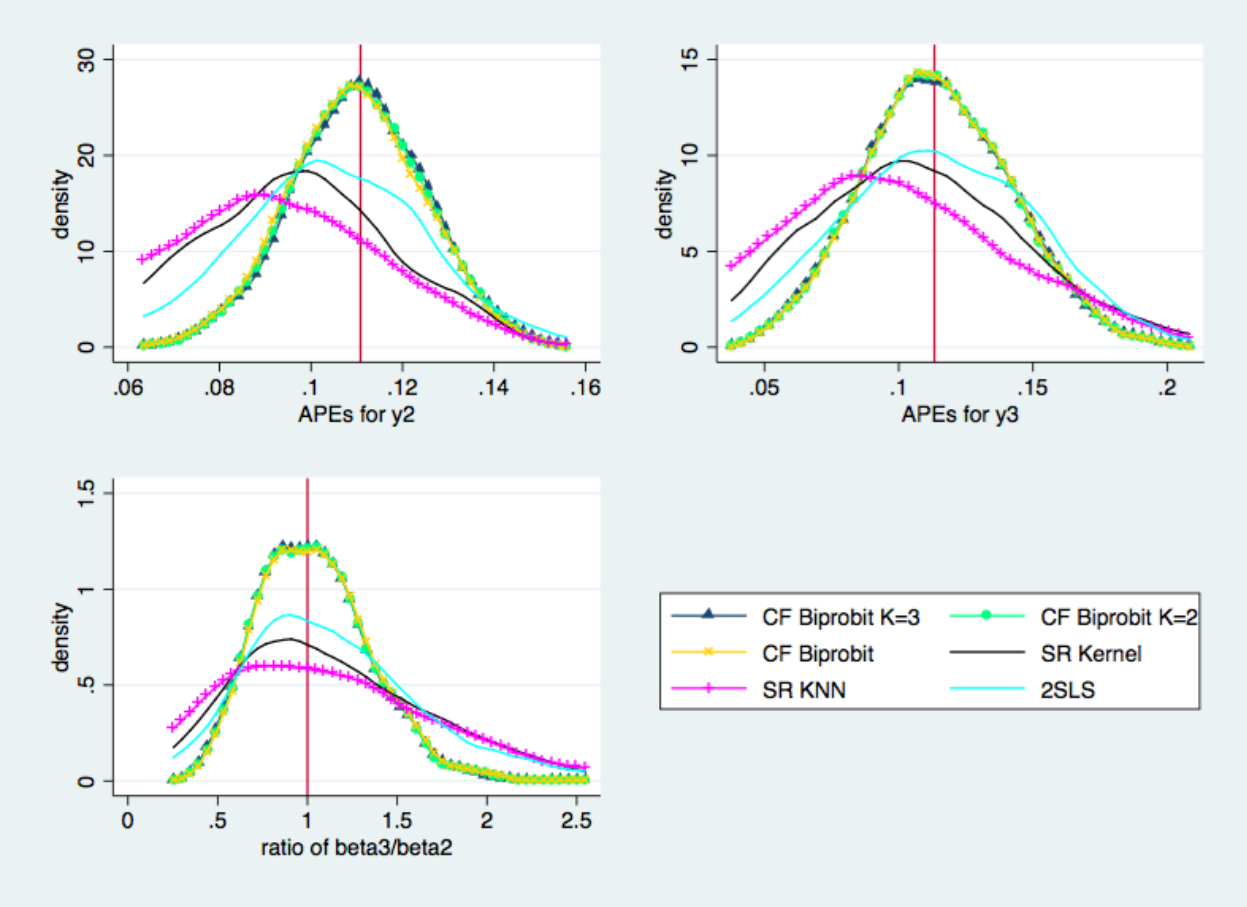

Table 1: Simulation Results for Design 1

	(1) CF Biprobit K=3	(2) CF Biprobit K=2	(3) CF Biprobit	(4) 2SLS	(5) SR Kernel	(6) SR KNN
	$z_1 \sim$Normal$(0,9)$, APE$_{y_2}$=.0806					
Bias	-.0001	-.0002	-.0002	-.0057	-.0143	-.0174
RMSE	.0053	.0053	.0054	.0090	.0160	.0195
	$z_1 \sim$Normal$(0,9)$, APE$_{y_3}$=.0851					
Bias	-.0014	.0005	-.0049	.0255	-.0009	-.0066
RMSE	.0334	.0331	.0355	.0649	.0638	.0791
	$z_1 \sim$Normal$(0,9)$, β_3/β_2=1					
Bias	-.0098	.0122	-.0494	.4993	.3227	.3258
RMSE	.4021	.3985	.4245	.9743	1.098	1.429
	$z_1 \sim$Normal$(0,1)$, APE$_{y_2}$=.0932					
Bias	.0003	.0010	.0012	-.0188	-.0328	-.0350
RMSE	.0065	.0066	.0064	.0200	.0339	.0365
	$z_1 \sim$Normal$(0,1)$, APE$_{y_3}$=.1046					
Bias	-.0071	-.0114	-.0208	.0640	.0203	.0123
RMSE	.0353	.0371	.0426	.0895	.0749	.0908
	$z_1 \sim$Normal$(0,1)$, β_3/β_2=1					
Bias	-.0603	-.1031	-.1904	1.296	1.201	1.163
RMSE	.3371	.3548	.4036	1.582	1.913	2.352

[a] Sequential averaging of the control function term v_2 and $\mathbf{x}$ is applied to compute estimates of APEs.

[b] The bias is defined as the difference between the true APEs (the true ratio one) and the estimates. RMSE is the root mean squared error.

[c] Estimator (1) is the CF approach with unknown functions approximated by sieve spaces of dimension three. Estimator (2) is the CF approach with unknown functions approximated by sieve spaces of dimension two. Estimator (3) is the standard CF approach with no unknown functions. Estimator (4) is the linear probability model estimated by usual two-stage least squares. Estimator (5) is the semiparametric special regressor method with density estimated by kernel methods. Estimator (6) is the semiparametric special regressor method with density estimated by K-Nearest-Neighbor method.

Table 2: Simulation Results for Design 2

	(1) CF Biprobit K=3	(2) CF Biprobit K=2	(3) CF Biprobit	(4) 2SLS	(5) SR Kernel	(6) SR KNN
	$z_1 \sim$Normal$(0, 9)$, APE$_{y_2}$=.0806					
Bias	-.0004	-.0023	-.0038	-.0056	-.0143	-.0165
RMSE	.0213	.0213	.0224	.0244	.0160	.0345
	$z_1 \sim$Normal$(0, 9)$, APE$_{y_3}$=.0851					
Bias	.0129	.0134	.0117	.0070	-.0009	-.0096
RMSE	.0430	.0427	.0429	.0613	.0638	.0787
	$z_1 \sim$Normal$(0, 9)$, β_3/β_2=1					
Bias	.2539	.2148	.2623	.3152	.8192	1.085
RMSE	.8013	.6549	.8199	1.036	21.975	14.040
	$z_1 \sim$Normal$(0, 16)$, APE$_{y_2}$=.0815					
Bias	-.0001	-.0013	-.0025	-.0045	-.0074	-.0152
RMSE	.0188	.0187	.0193	.0216	.0258	.0315
	$z_1 \sim$Normal$(0, 16)$, APE$_{y_3}$=.0828					
Bias	.0163	.0157	.0138	.0067	.0021	-.0058
RMSE	.0417	.0407	.0408	.0604	.0642	.0720
	$z_1 \sim$Normal$(0, 16)$, β_3/β_2=1					
Bias	.3069	.2947	.3121	.3314	.2678	.3837
RMSE	.7561	.7430	.8033	1.1816	2.924	6.094

[a] Sequential averaging of the control function term v_2 and $\mathbf{x}$ is applied to compute estimates of APEs.

[b] The bias is defined as the difference between the true APEs (the true ratio one) and the estimates. RMSE is the root mean squared error.

[c] Estimator (1) is the CF approach with unknown functions approximated by sieve spaces of dimension three. Estimator (2) is the CF approach with unknown functions approximated by sieve spaces of dimension two. Estimator (3) is the standard CF approach with no unknown functions. Estimator (4) is the linear probability model estimated by usual two-stage least squares. Estimator (5) is the semiparametric special regressor method with density estimated by kernel methods. Estimator (6) is the semiparametric special regressor method with density estimated by K-Nearest-Neighbor method.

Table 3: Simulation Results for Design 3

	(1) CF Biprobit K=3	(2) CF Biprobit K=2	(3) CF Biprobit	(4) 2SLS	(5) SR Kernel	(6) SR KNN
			APE_{y_2}=.1108			
Bias	-.0001	-.0006	-.0011	-.0065	-.0156	-.0220
RMSE	.0144	.0144	.0146	.0212	.0271	.0337
			APE_{y_3}=.1133			
Bias	.0013	.0013	.0015	.0006	-.0080	-.0187
RMSE	.0275	.0273	.0272	.0372	.0420	.0494
			β_3/β_2=1			
Bias	.0356	.0400	.0484	.1585	.2187	.3097
RMSE	.3077	.3096	.3141	.5311	.7540	1.850

[a] Sequential averaging of the control function term v_2 and $\mathbf{x}$ is applied to compute estimates of APEs.

[b] The bias is defined as the difference between the true APEs (the true ratio one) and the estimates. RMSE is the root mean squared error.

[c] Estimator (1) is the CF approach with unknown functions approximated by sieve spaces of dimension three. Estimator (2) is the CF approach with unknown functions approximated by sieve spaces of dimension two. Estimator (3) is the standard CF approach with no unknown functions. Estimator (4) is the linear probability model estimated by usual two-stage least squares. Estimator (5) is the semiparametric special regressor method with density estimated by kernel methods. Estimator (6) is the semiparametric special regressor method with density estimated by K-Nearest-Neighbor method.

References

Ai, C., Chen, X., 2003. Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions. Econometrica 71, 1795–1843.

Ai, C., Chen, X., 2007. Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. Journal of Econometrics, Semiparametric methods in econometrics 141, 5–43. doi:10.1016/j.jeconom.2007.01.013

Ackerberg, D., Chen, X., Hahn, J., 2012. A practical asymptotic variance estimator for two-step semiparametric estimators. Review of Economics and Statistics 94, 481–498. doi:10.1162/REST_a_00251

Blundell, R., Powell, J.L., 2003. Endogeneity in nonparametric and semiparametric regression models. In: Dewatripont, M., Hansen, L., Turnsovsky, S.J. (Eds.), Advances in Economics and Econometrics. Cambridge University Press, Cambridge, pp. 312–357. (Chapter 8).

Blundell, R.W., Powell, J.L., 2004. Endogeneity in semiparametric binary response models. Review of Economic Studies 71, 655–679.

Blundell, R., Smith, R.J., 1994. Coherency and estimation in simultaneous models with censored or qualitative dependent variables. Journal of Econometrics 64, 355–373. doi:10.1016/0304-4076(94)90069-8

Chen, X., 2007. Chapter 76 Large sample sieve estimation of semi-nonparametric models, in: Handbook of Econometrics. Elsevier, pp. 5549–5632.

Chen, X., Shen X., 1998. Sieve Extremum Estimates for Weakly Dependent Data, Econometrica, 66, 289-314.

Chen, X., Liao Z., Sun Y., 2012. Sieve Inference for Weakly Dependent Data, Cowles Foundation Discussion Paper 1849. Yale University.

Chen, X., Linton, O., Van Keilegom, I., 2003. Estimation of semiparametric models when the criterion function is not smooth. Econometrica 1591–1608.

Chesher, A., 2003. Identification in nonseparable models. Econometrica 71, 1405–1441. doi:10.1111/1468-0262.00454

Grenander, U., 1981. Abstract inference. Wiley Series, New York.

Gallant, A.R., Nychka, D.W., 1987. Semi-nonparametric maximum likelihood estimation. Econometrica 55, 363–390. doi:10.2307/1913241

Hahn, J., Ridder, G., 2013. Asymptotic variance of semiparametric estimators with generated regressors. Econometrica 81, 315-340.

Hahn, J., Liao, Z., Ridder, G., 2015. Nonparametric two-step sieve M estimation and inference. Mimeo.

Han, S., Vytlacil, E., 2015. Identification in a generalization of bivariate probit models with dummy endogenous regressors. Mimeo.

Heckman, J.J., 1978. Dummy endogenous variables in dummy endogenous variables in a simultaneous equation system. Econometrica 46, 931–959.

Klein, R., Spady, R., 1993. An efficient semiparametric estimator for binary response models. Econometrica 61 (2), 387–421.

Lewbel, A., 2000. Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. Journal of Econometrics 97, 145–177. doi:10.1016/S0304-4076(00)00015-4

Lin, W., Wooldridge, J.M., 2015a. Estimating binary response models with endogenous explanatory variables, using control functions in Quasi-LIML. Mimeo.

Lin, W., Wooldridge, J.M., 2015b. On different approaches to obtaining partial effects in binary response models with endogenous regressors. Economics Letters 134, 58–61. doi:10.1016/j.econlet.2015.05.019

Petrin, A., Train, K., 2010. A Control Function Approach to Endogeneity in Consumer Choice Models. Journal of Marketing Research (JMR) 47, 3–13. doi:10.1509/jmkr.47.1.3

Powell, J.L., 1994. Estimation of semiparametric models. Handbook of econometrics 4, 2443–2521.

Rivers, D., Vuong, Q.H., 1988. Limited information estimators and exogeneity tests for simultaneous probit models. Journal of Econometrics 39, 347–366. doi:10.1016/0304-4076(88)90063-2

Rothe, C., 2009. Semiparametric estimation of binary response models with endogenous regressors. Journal of Econometrics 153, 51–64. doi:10.1016/j.jeconom.2009.04.005

Shen, X., 1997. On Methods of Sieves and Penalization. Annals of Statistics, 25, 2555-2591.

Staiger, D., Stock, J.H., 1997. Instrumental variables regression with weak instruments. Econometrica 65.

Terza, J.V., Basu, A., Rathouz, P.J., 2008. Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. Journal of Health Economics 27, 531–543. doi:10.1016/j.jhealeco.2007.09.009

Wooldridge, J.M., 2010. Econometric analysis of cross section and panel data. MIT Press.

Wooldridge, J.M., 2014. Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables. Journal of Econometrics, Causality, Prediction, and Specification Analysis: Recent Advances and Future Directions 182, 226–234. doi:10.1016/j.jeconom.2014.04.020

Wooldridge, J.M., 2015. Control function methods in applied econometrics. Journal of Human Resources 50, 420–445.