Nonparametric Two-Step Sieve M Estimation and Inference*

Jinyong Hahn† Zhipeng Liao‡ Geert Ridder§
UCLA UCLA USC

First Version: April 2012; This version: March 2016

Abstract

This paper studies the two-step sieve M estimation of general semi/nonparametric models, where the second step involves sieve estimation of unknown functions that may use the nonparametric estimates from the first step as inputs, and the parameters of interest are functionals of unknown functions estimated in both steps. We establish the asymptotic normality of the plug-in two-step sieve M estimate of a functional that could be root-n estimable. The asymptotic variance may not have a closed form expression, but can be approximated by a sieve variance that characterizes the effect of the first-step estimation on the second-step estimates. We provide a simple consistent estimate of the sieve variance and hence a Wald type inference based on the Gaussian approximation. The finite sample performance of the two-step estimator and the proposed inference procedure are investigated in a simulation study.

JEL Classification: C14, C31, C32

Keywords: Two-Step Sieve Estimation; Nonparametric Generated Regressors; Asymptotic Normality; Sieve Variance Estimation

1 Introduction

Many recently introduced empirical methodologies adopt semiparametric two-step estimation approaches, where certain functions are estimated nonparametrically in the first step, and some Euclidean parameters are estimated parametrically in the second step using the nonparametric estimates from the first stage as

*We gratefully acknowledge insightful comments from Xiaohong Chen, who was a co-author of the initial version. We appreciate useful suggestions from Liangjun Su, the coeditor and three anonymous referees. All errors are the responsibility of the authors.

†Department of Economics, UCLA, Los Angeles, CA 90095-1477 USA. Email: hahn@econ.ucla.edu

‡Department of Economics, UCLA, Los Angeles, CA 90095-1477 USA. Email: zhipeng.liao@econ.ucla.edu

§Department of Economics, University of Southern California, Los Angeles, CA 90089. Email: ridder@usc.edu.

inputs. Large sample properties such as the root-n asymptotic normality of the second step parametric estimators are well established in the literature. See, e.g., Newey (1994), Chen, Linton and van Keilegom (2003), and Ichimura and Lee (2010). Despite the mature nature of the literature, the mathematical framework employed by the existing general theory papers is not easily applicable to situations where some of the regressors need to be estimated. Many estimators that utilize so-called control variables may be incorrectly analyzed if this mathematical framework is adopted without proper modification, because the control variables need to be estimated in many applications. See Hahn and Ridder (2013).

Estimators using control variables are often based on the intuition that certain endogeneity problems may be overcome by conditioning on such variables.[1] The control variables need to be estimated in practice, often by semiparametric or nonparametric methods. The second step estimation usually takes the form of a semiparametric or nonparametric estimation, as well. See, e.g., Olley and Pakes (1996), Newey, Powell and Vella (1999), Lee (2007), and Newey (2009). Such estimators are becoming increasingly important in the literature, yet their asymptotic properties have been established only on a case-by-case basis. In this paper, we establish statistical properties of the two step estimators in an extended mathematical framework that can address the generated regressor problem.

We present a mathematical framework general enough to nest both the previous two-step estimation literature and the literature on control variables. This goal is achieved by investigating statistical properties of nonparametric two-step sieve M estimation, where the second step may involve estimation of some infinite dimensional parameters, and the nonparametric components in the second step may use some estimated value from the first step as input. The parameters of interest are functionals of unknown functions estimated in both steps. The estimation procedure considered in the paper formally consists of the following steps. In the first step, we maximize $\sum_{i=1}^{n} \varphi\left(Z_{1,i}, h\right)$ with respect to a (possibly infinite dimensional) parameter h. Letting $\widehat{h}_n$ denote the maximizer in the first step, we then maximize $\sum_{i=1}^{n} \psi(Z_{2,i}, g, \widehat{h}_n)$ with respect to yet another (possibly infinite dimensional) parameter g and get the second-step estimator $\widehat{g}_n$. Let $\rho\left(h_o, g_o\right)$ denote the parameter of interest, where the functional $\rho(\cdot)$ is known. We finally estimate $\rho\left(h_o, g_o\right)$ by a simple plug-in two-step sieve M estimator $\rho(\widehat{h}_n, \widehat{g}_n)$. We allow the first step function h to enter into the second step function g as an argument, as is common in many procedures using estimated control variables. By establishing statistical properties of such sieve two-step estimators, we contribute toward a general theory of semi/nonparametric two-step estimation.[2]

Our primary concern are the practical aspects of inference. We show that the numerical equivalence results in the literature, i.e., Newey (1994, Section 6) and Ackerberg, Chen and Hahn (2012), continue to hold in the general framework that nests estimators that use estimated control variables. Newey (1994,

Section 6) and Ackerberg, Chen and Hahn (2012) consider a framework that rules out the estimated control variable problem, and showed that in the context of sieve/series estimation, practitioners may assume that his/her model is parametric when a consistent estimator of the asymptotic variance is desired. In other words, practitioners can ignore the infinite dimensional nature of the problem, and proceed with the standard parametric procedure proposed by Newey (1984) or Murphy and Topel (1985). We show that practitioners may continue to adopt such a convenient (yet incorrectly specified) parametric model even in more general estimators that require dealing with estimated control variables. To be more precise, our consistent estimator of the asymptotic variance of $\rho(\widehat{h}_n, \widehat{g}_n)$ turns out to be identical to a consistent estimator of the asymptotic variance derived under such a parametric interpretation. Because the asymptotic variance of $\rho(\widehat{h}_n, \widehat{g}_n)$ often takes a very complicated form, such numerical equivalence is expected to facilitate inference in practice.

The rest of this paper is organized as follows. Section 2 introduces the nonparametric two-step sieve M estimation and presents several examples. Section 3 establishes the asymptotic normality of $\rho(\widehat{h}_n, \widehat{g}_n)$. Section 4 proposes a consistent sieve estimator of the asymptotic variance of $\rho(\widehat{h}_n, \widehat{g}_n)$. Section 5 studies an nonparametric two-step regression example to illustrate the high-level conditions in Section 3 and Section 4. Section 6 presents simple numerically equivalent ways of calculating the sieve variance estimator, which is the main result of the paper. Section 7 reports the results of a simulation studies of the finite sample performance of the two-step nonparametric estimator and the proposed inference procedure. Section 8 concludes. A simple cross-validation method for selecting the tuning parameters in the two-step nonparametric regression is provided in Appendix A. Appendix B contains the proof of the normality theorem. The proofs of the results in Sections 4 and 6 are gathered in Appendix C. Sufficient conditions of the main theorem of the two-step nonparametric regression example are available in Appendix D. This theorem is proved by several lemmas which are included in the Supplemental Appendix of the paper. The consistency and the convergence rate of the second-step sieve M estimator $\widehat{g}_n$, which are not the main focus of this paper, are also presented in the Supplemental Appendix.

2 Two-step Sieve M Estimation

In this section we introduce some notation for the two-step nonparametric sieve M estimation and present several examples which are covered by our general setup. In the first step, we use the random vector $Z_{1,i}$ $i = 1, \ldots, n$ to identify and estimate a potentially infinite dimensional parameter h_o. In the second step, we use the random vector $Z_{2,i}$ $i = 1, \ldots, n$ to identify and estimate a potentially infinite dimensional

parameter g_o. Let Z_i be the union of the distinct elements of $Z_{1,i}$ and $Z_{2,i}$, and assume that Z_i has the same distribution F_Z. We will assume that $\{Z_i\}_{i=1}^n$ is a sequence of strictly stationary weakly dependent data.

The two-step estimation is based on the identifying assumption that $h_o \in \mathcal{H}$ is the unique solution to $\sup_{h\in\mathcal{H}} E[\varphi(Z_1, h)]$ and $g_o \in \mathcal{G}$ is the unique solution to $\sup_{g\in\mathcal{G}} E[\psi(Z_2, g, h_o)]$, where $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ and $(\mathcal{G}, \|\cdot\|_{\mathcal{G}})$ are some infinite dimensional separable complete metric spaces, and $\|\cdot\|_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{G}}$ could be the sup-norm, the $L_2(dF_Z)$-norm, the Sobolev norm, or the Hölder norm.

The two-step (approximate) sieve M estimation utilizes the sample analog of the identifying assumption. We restrict our estimators of h_o and g_o to be in some finite dimensional sieve spaces $\mathcal{H}_n$ and $\mathcal{G}_n$ respectively, where the dimension of each space grows to infinity as a function of the sample size n. In the first step, we estimate $h_o \in \mathcal{H}$ by $\widehat{h}_n \in \mathcal{H}_n$ defined as

$$\frac{1}{n}\sum_{i=1}^n \varphi\left(Z_{1,i}, \widehat{h}_n\right) \geq \sup_{h\in\mathcal{H}_n} \frac{1}{n}\sum_{i=1}^n \varphi(Z_{1,i}, h) - O_p(\varepsilon_{1,n}^2); \tag{1}$$

in the second step, we estimate $g_o \in \mathcal{G}$ by $\widehat{g}_n \in \mathcal{G}_n$ defined as

$$\frac{1}{n}\sum_{i=1}^n \psi\left(Z_{2,i}, \widehat{g}_n, \widehat{h}_n\right) \geq \sup_{g\in\mathcal{G}_n} \frac{1}{n}\sum_{i=1}^n \psi\left(Z_{2,i}, g, \widehat{h}_n\right) - O_p(\varepsilon_{2,n}^2), \tag{2}$$

where $\mathcal{H}_n \equiv \mathcal{H}_{K(n)}$ and $\mathcal{G}_n \equiv \mathcal{G}_{L(n)}$ are the sieve spaces that become dense in $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ and $(\mathcal{G}, \|\cdot\|_{\mathcal{G}})$ as $L(n) \equiv \dim(\mathcal{H}_n) \to \infty$ and $K(n) \equiv \dim(\mathcal{G}_n) \to \infty$ respectively. The magnitudes of the optimization errors $\varepsilon_{1,n}^2$ and $\varepsilon_{2,n}^2$ are small positive numbers that go to zero as $n \to \infty$. See, e.g., Chen (2007) for many examples of sieve spaces and criterion functions for sieve M estimation.

Below are a few examples in the literature that adopt such a two-step sieve M estimation strategy.

Example 2.1 (Nonparametric Triangular Simultaneous Equation Model) *Newey, Powell and Vella (1999) study a nonparametric triangular simultaneous equation model. Let $\{(y_i, x_i, z_{1,i})\}_{i=1}^n$ be a random sample generated according to the model*

$$y = m_o(x, z_2) + \epsilon, \quad E[\epsilon|u, z_1] = E[\epsilon|u],$$
$$x = h_o(z_1) + u, \quad E[u|z_1] = 0,$$

where $h_o \in \mathcal{H}$ and $m_o \in \mathcal{M}$, $z_2 \subset z_1$. This yields the identifying relation $E[y|x, z_1] = m_o(x, z_2) + \lambda_o(u)$, where $\lambda_o(u) \equiv E[\epsilon|u] \in \Lambda = \{\lambda : E[\lambda(u)^2] < \infty, \lambda(\overline{u}) = \overline{\lambda}\}$, here $\lambda(\overline{u}) = \overline{\lambda}$ (for known constants $\overline{u}$, $\overline{\lambda}$) is

a location normalization. Their model specification implies that the h_o is identified as a solution to the infinite dimensional least squares problem

$$\sup_{h\in\mathcal{H}} E\left[-\left(x-h\left(z_1\right)\right)^2\right].$$

Given h_o, the $g_o=(m_o,\lambda_o)$ is a solution to another infinite dimensional least squares problem

$$\sup_{m\in\mathcal{M},\lambda\in\Lambda} E\left[-\left(y-m\left(x,z_2\right)-\lambda\left(x-h_o\left(z_1\right)\right)\right)^2\right].$$

Newey, Powell and Vella (1999) estimate $h_o\in\mathcal{H}$ and $g_o=(m_o,\lambda_o)\in\mathcal{G}$ by a two-step series Least Squares regression. In the first step, they compute

$$\widehat{h}_n=\arg\max_{h\in\mathcal{H}_n}-\frac{1}{n}\sum_{i=1}^{n}\left(x_i-h\left(z_{1,i}\right)\right)^2;$$

in the second step, they compute

$$\widehat{g}_n=(\widehat{m}_n,\widehat{\lambda}_n)=\arg\max_{m\in\mathcal{M}_n,\lambda\in\Lambda_n}-\frac{1}{n}\sum_{i=1}^{n}\left(y_i-m\left(x_i,z_{2,i}\right)-\lambda\left(x_i-\widehat{h}_n\left(z_{1,i}\right)\right)\right)^2,$$

where $\mathcal{H}_n$ and $\mathcal{G}_n=\mathcal{M}_n\times\Lambda_n$ are finite dimensional linear sieves (or series). Let $\rho(\cdot)$ be a linear functional of g. Newey, Powell and Vella (1999) establish the asymptotic normality of plug-in sieve estimate $\rho(\widehat{g}_n)$ of their parameter of interest $\rho\left(g_o\right)$.

Example 2.2 (Nonparametric sample selection with endogeneity) *Das, Newey and Vella (2003) studies a nonparametric sample selection model with endogeneity. Let $\{(y_i,d_i,x_i,z_{1i})\}_{i=1}^n$ be a random sample generated according to the model*

$$\begin{aligned}
&y=d\cdot y^*,\quad y^*=m_o(x,z_2)+\epsilon,\\
&x=h_{o1}(z_1)+u,\quad E\left[u\middle|\,z_1\right]=0,\quad \Pr(d=1|z_1)=h_{o2}(z_1),
\end{aligned}$$

where $h_{o1}\in\mathcal{H}_1$, $h_{o2}\in\mathcal{H}_2$ and $m_o\in\mathcal{M}$, $z_2\subset z_1$, y^* is not directly observable. Das, Newey and Vella (2003) impose the restriction $E\left[\epsilon\middle|\,u,z_1,d=1\right]=\lambda_o\left(h_{o2},u\right)$ for some $\lambda_o\in\Lambda$. This gives us the identifying relation $E\left[y\middle|\,x,z_1,d=1\right]=m_o\left(x,z_2\right)+\lambda_o(h_{o2},u)\in\mathcal{G}=\mathcal{M}+\Lambda$. Their model specification implies that

the $h_o = (h_{o1}, h_{o2})$ is identified as a solution to

$$\sup_{h_1 \in \mathcal{H}_1} E[-(x - h_1(z_1))^2] \quad \textit{and} \quad \sup_{h_2 \in \mathcal{H}_2} E[-(d - h_2(z_1))^2].$$

Given $h_o = (h_{o1}, h_{o2})$, the $g_o = m_o + \lambda_o$ is a solution to the infinite dimensional least squares problem

$$\sup_{g=m+\lambda \in \mathcal{G}} E\left[-\left(y - m(x, z_2) - \lambda\left(h_{o2}(z_1), x - h_{o1}(z_1)\right)\right)^2\right].$$

Das, Newey and Vella (2003) estimate $h_o = (h_{o1}, h_{o2})$ and $g_o = m_o + \lambda_o$ by a two-step series Least Squares regression, which exactly fits into our two-step sieve M estimation procedure. Let $\rho(\cdot)$ be a linear functional of g. Das, Newey and Vella (2003) establish the asymptotic normality of plug-in sieve estimate $\rho(\widehat{g}_n)$ of their parameter of interest $\rho(g_o)$.

Example 2.3 (Production Function) *Olley and Pakes (1996) consider estimation of the Cobb-Douglas production function*

$$y_{it} = \beta_0 + \beta_k k_{it} + \beta_l l_{it} + \omega_{it} + \eta_{it}, \quad t = 1, 2$$

where y_{it}, k_{it}, l_{it} denote the (logs of) output, capital and labor inputs, respectively. The ω_{it} denotes an unobserved productivity index that follows a first-order Markov process, and η_{it} can be viewed as a measurement error. Assuming away the potential invertibility problem of the investment demand function, they show that the β_l is identified as the solution to the semiparametric problem

$$\min_{\beta_l, \phi_1} E\left[(y_{i1} - \beta_l l_{i1} - \phi_1(i_{i1}, k_{i1}))^2\right]$$

where ϕ_1 is an infinite dimensional parameter. Given (β_l, ϕ_1), they show that β_k is identified as the solution to the semiparametric problem

$$\min_{\beta_k, \lambda} E\left[\left(y_{i2} - \beta_l l_{i2} - \beta_k k_{i2} - \lambda\left(\phi_1(i_{i1}, k_{i1}) - \beta_k k_{i1}\right)\right)^2\right]$$

where λ is an infinite dimensional parameter.

3 Asymptotic Normality

In this section, we establish the asymptotic normality of the plug-in two-step sieve M estimators. We characterize the asymptotic variance of the estimator, and we propose an estimator of the asymptotic

variance that is its sample analog in the next section. As is typical in the literature on sieve estimation, our asymptotic normality result is predicated on certain consistency and rate of convergence results.[3]

Let $\widehat{\alpha}_n = (\widehat{h}_n, \widehat{g}_n)$ and $\alpha_o = (h_o, g_o)$. We assume that the convergence rates of $\widehat{h}_n$ and $\widehat{g}_n$ are $\delta^*_{1,n}$ and $\delta^*_{2,n}$ under $\|\cdot\|_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{G}}$ respectively, where $\{\delta^*_{j,n}\}$ is a positive sequence such that $\delta^*_{j,n} = o(1)$ for $j = 1, 2$. Let $\delta_{j,n} = \delta^*_{j,n} \log(\log(n))$ $(j = 1, 2)$. Then $\widehat{\alpha}_n$ belongs to the shrinking neighborhood $\mathcal{N}_n = \{(h, g) : h \in \mathcal{N}_{h,n} \text{ and } g \in \mathcal{N}_{g,n}\}$ with probability approaching 1 (wpa1), where $\mathcal{N}_{h,n} = \{h \in \mathcal{H}_n : \|h - h_o\|_{\mathcal{H}} \leq \delta_{1,n}\}$ and $\mathcal{N}_{g,n} = \{g \in \mathcal{G}_n : \|g - g_o\|_{\mathcal{G}} \leq \delta_{2,n}\}$.

We suppose that for all $h \in \mathcal{N}_{h,n}$, $\varphi(Z_1, h) - \varphi(Z_1, h_o)$ can be approximated by $\Delta_\varphi(Z_1, h_o)[h - h_o]$ such that $\Delta_\varphi(Z_1, h_o)[h - h_o]$ is linear in $h - h_o$. As h_o is the unique maximizer of $E[\varphi(Z_1, h)]$ on $\mathcal{H}$, we can let

$$-\frac{\partial E\left[\Delta_\varphi(Z_1, h_o + \tau[h - h_o])[h - h_o]\right]}{\partial \tau}\bigg|_{\tau=0} \equiv \|h - h_o\|^2_\varphi,$$

which defines a norm on $\mathcal{N}_{h,n}$. Let $\mathcal{V}_1$ be the closed linear span of $\mathcal{N}_{h,n} - \{h_o\}$ under $\|\cdot\|_\varphi$, which is a Hilbert space under $\|\cdot\|_\varphi$, with the corresponding inner product $\langle\cdot,\cdot\rangle_\varphi$ defined as

$$\langle v_{h_1}, v_{h_2}\rangle_\varphi = -\frac{\partial E\left[\Delta_\varphi(Z_1, h_o + \tau v_{h_2})[v_{h_1}]\right]}{\partial \tau}\bigg|_{\tau=0} \tag{3}$$

for any $v_{h_1}, v_{h_2} \in \mathcal{V}_1$. In many examples, we have

$$\Delta_\varphi(Z_1, h_o)[v_{h_1}] = \frac{\partial \varphi(Z_1, h_o + \tau v_{h_1})}{\partial \tau}\bigg|_{\tau=0} \text{ and } \langle v_{h_1}, v_{h_2}\rangle_\varphi = -E\left[\frac{\partial \Delta_\varphi(Z_1, h_o + \tau v_{h_2})[v_{h_1}]}{\partial \tau}\bigg|_{\tau=0}\right],$$

if the derivative exists and we can interchange differentiation and expectation. We assume that there is a linear functional $\partial_h \rho(\alpha_o)[\cdot] : \mathcal{V}_1 \to \mathbb{R}$ such that

$$\partial_h \rho(\alpha_o)[v] = \frac{\partial \rho(h_o + \tau v, g_o)}{\partial \tau}\bigg|_{\tau=0} \text{ for all } v \in \mathcal{V}_1.$$

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 $\dim(\mathcal{V}_{1,n}) = L(n) < \infty$. By Riesz representation theorem, there is a sieve Riesz representer $v^*_{h_n} \in \mathcal{V}_{1,n}$ such that

$$\partial_h \rho(\alpha_o)[v] = \left\langle v^*_{h_n}, v\right\rangle_\varphi \text{ for all } v \in \mathcal{V}_{1,n}, \quad \text{and} \quad \left\|v^*_{h_n}\right\|^2_\varphi = \sup_{0 \neq v \in \mathcal{V}_{1,n}} \frac{\left|\partial_h \rho(\alpha_o)[v]\right|^2}{\|v\|^2_\varphi}. \tag{4}$$

Moreover, $\partial_h \rho(\alpha_o)[\cdot] : \mathcal{V}_1 \to \mathbb{R}$ is a bounded (or regular) functional if and only if $\lim_{L(n)\to\infty} \left\|v^*_{h_n}\right\|_\varphi < \infty$.

Similarly, we use $\Delta_\psi(Z_2, g_o, h_o)[g-g_o]$ to denote the linear approximation of $\psi(Z_2, g, h_o)-\psi(Z_2, g_o, h_o)$ for any $g \in \mathcal{N}_{g,n}$. As g_o is the unique maximizer of $E[\psi(Z_2, g, h_o)]$ on $\mathcal{G}$, we can let

$$-\frac{\partial E\left[\Delta_\psi(Z_2, g_o+\tau[g-g_o], h_o)[g-g_o]\right]}{\partial \tau}\bigg|_{\tau=0} \equiv \|g-g_o\|_\psi^2,$$

which defines a norm on $\mathcal{N}_{g,n}$. Let $\mathcal{V}_2$ be the closed linear span of $\mathcal{N}_{g,n}-\{g_o\}$ under $\|\cdot\|_\psi$, which is a Hilbert space under $\|\cdot\|_\psi$, with the corresponding inner product $\langle\cdot,\cdot\rangle_\psi$ defined as

$$\langle v_{g_1}, v_{g_2}\rangle_\psi = -\frac{\partial E\left[\Delta_\psi(Z_2, g_o+\tau v_{g_2}, h_o)[v_{g_1}]\right]}{\partial \tau}\bigg|_{\tau=0}. \tag{5}$$

for any $v_{g_1}, v_{g_2} \in \mathcal{V}_2$. In many cases, we have

$$\Delta_\psi(Z_2, \alpha_o)[v_{g_1}] = \frac{\partial \psi(Z_2, g_o+\tau v_{g_1}, h_o)}{\partial \tau}\bigg|_{\tau=0} \text{ and } \langle v_{g_1}, v_{g_2}\rangle_\psi = -E\left[\frac{\partial \Delta_\psi(Z_2, g_o+\tau v_{g_2}, h_o)[v_{g_1}]}{\partial \tau}\right].$$

We assume that there is a linear functional $\partial_g \rho(\alpha_o)[\cdot]: \mathcal{V}_2 \to \mathbb{R}$ such that

$$\partial_g \rho(\alpha_o)[v] = \frac{\partial \rho(h_o, g_o+\tau v)}{\partial \tau}\bigg|_{\tau=0} \text{ for all } v \in \mathcal{V}_2. \tag{6}$$

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 $\dim(\mathcal{V}_{2,n}) = K(n) < \infty$. By Riesz representation theorem, there is a sieve Riesz representer $v_{g_n}^* \in \mathcal{V}_{2,n}$ such that

$$\partial_g \rho(\alpha_o)[v_g] = \left\langle v_{g_n}^*, v_g\right\rangle_\psi \text{ for all } v \in \mathcal{V}_{2,n}, \quad \text{and} \quad \left\|v_{g_n}^*\right\|_\psi^2 = \sup_{0\neq v\in\mathcal{V}_{2,n}} \frac{|\partial_g \rho(\alpha_o)[v]|^2}{\|v\|_\psi^2}. \tag{7}$$

Moreover, $\partial_g \rho(\alpha_o)[\cdot]: \mathcal{V}_2 \to \mathbb{R}$ is a bounded functional if and only if $\lim_{K(n)\to\infty}\left\|v_{g_n}^*\right\|_\psi < \infty$.

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

$$\partial_\alpha \rho(\alpha_o)[v] = \partial_h \rho(\alpha_o)[v_h] + \partial_g \rho(\alpha_o)[v_g].$$

Then $\partial_\alpha \rho(\alpha_o)[\cdot]: \mathcal{V} \to \mathbb{R}$ is a bounded functional if and only if $\lim_{L(n)\to\infty}\left\|v_{h_n}^*\right\|_\varphi < \infty$ and $\lim_{K(n)\to\infty}\left\|v_{g_n}^*\right\|_\psi < \infty$.

To evaluate the effect of the first-step estimation on the asymptotic variance of the second-step sieve

M estimator, we define

$$\Gamma_g(\alpha_o)[v_g] = \left.\frac{\partial E[\psi(Z_2, g_o + \tau v_g, h_o)]}{\partial \tau}\right|_{\tau=0} \text{ for any } v_g \in \mathcal{V}_2$$

and

$$\Gamma(\alpha_o)[v_h, v_g] = \left.\frac{\partial \Gamma_g(g_o, h_o + \tau v_h)[v_g]}{\partial \tau}\right|_{\tau=0} \text{ for any } v_h \in \mathcal{V}_1.$$

We assume that $\Gamma(\alpha_o)[\cdot,\cdot]$ is a bilinear functional on $\mathcal{V}$. Given the Riesz representer $v^*_{g_n}$ in (7), we define $v^*_{\Gamma_n} \in \mathcal{V}_{1,n}$ as

$$\Gamma(\alpha_o)\left[v_h, v^*_{g_n}\right] = \left\langle v_h, v^*_{\Gamma_n}\right\rangle_\varphi \text{ for any } v_h \in \mathcal{V}_{1,n}. \tag{8}$$

Using the sieve Riesz representers $v^*_{h_n}$, $v^*_{g_n}$ and $v^*_{\Gamma_n}$, we define

$$\|v^*_n\|^2_{sd} = \text{Var}\left[n^{-\frac{1}{2}} \sum\nolimits_{i=1}^n \left(\Delta_\varphi(Z_{1,i}, h_o)[v^*_{h_n}] + \Delta_\varphi(Z_{1,i}, h_o)[v^*_{\Gamma_n}] + \Delta_\psi(Z_{2,i}, \alpha_o)[v^*_{g_n}]\right)\right]. \tag{9}$$

As we can see from (9), the sieve variance of the plug-in estimator $\rho(\widehat{\alpha}_n)$ is determined by three components.[4] To explain the sources of these components, we assume that the criterion function in the second-step estimation is smooth. The first component $n^{-\frac{1}{2}} \sum_{i=1}^n \Delta_\varphi(Z_{1,i}, h_o)[v^*_{h_n}]$ is from the estimation error of h_o which enters $\|v^*_n\|^2_{sd}$ because $\rho(\alpha_o)$ depends on h_o. When $\rho(\alpha_o)$ only depends on g_o, we have $v^*_{h_n} = 0$ and then this component will not show up in $\|v^*_n\|^2_{sd}$. The second component $n^{-\frac{1}{2}} \sum_{i=1}^n \Delta_\varphi(Z_{1,i}, h_o)[v^*_{\Gamma_n}]$ is also from the estimation error of h_o which enters $\|v^*_n\|^2_{sd}$ because the first derivative of the criterion function in the second-step estimation depends on the first-step M estimator $\widehat{h}_n$. The last component $n^{-\frac{1}{2}} \sum_{i=1}^n \Delta_\varphi(Z_{1,i}, h_o)[v^*_{g_n}]$ is from the estimation error of g_o in the second-step estimation assuming that the unknown parameter h_o is known. It enters $\|v^*_n\|^2_{sd}$ because $\rho(\alpha_o)$ depends on g_o. When $\rho(\alpha_o)$ only depends on g_o and the parameter h_o is known, $\|v^*_n\|^2_{sd}$ is identical to the sieve variance of the one-step plug-in sieve M estimator found in Chen, Liao and Sun (2014) and Chen and Liao (2014).

In this paper, we restrict our attention to the class of functionals $\rho(\alpha_o)$ such that $\|v^*_n\|^2_{sd} \leq C$ for any n. We next list the assumptions needed for showing the asymptotic normality of $\rho(\widehat{\alpha}_n)$.

Assumption 3.1 *(i)* $\liminf_n \|v^*_n\|_{sd} > 0$*; (ii) the functional* $\rho(\cdot)$ *satisfies*

$$\sup_{\alpha \in \mathcal{N}_n} \left|\frac{\rho(\alpha) - \rho(\alpha_o) - \partial_h \rho(\alpha_o)[h - h_o] - \partial_g \rho(\alpha_o)[g - g_o]}{\|v^*_n\|_{sd}}\right| = o(n^{-\frac{1}{2}});$$

(iii) there exists $g_n \in \mathcal{G}_n$ *such that* $\|g_n - g_o\|_{\mathcal{G}} = O(\delta^*_{2,n})$ *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* $\|v_g\|_\psi \leq c_\psi \|v_g\|_{\mathcal{G}}$ *where* c_φ *and* c_ψ *are some generic finite positive constants; (iv)*

$$\frac{1}{\|v_n^*\|_{sd}} \max\left\{\left|\partial_h \rho(\alpha_o)[h_{o,n} - h_o]\right|, \left|\partial_g \rho(\alpha_o)[g_{o,n} - g_o]\right|\right\} = o(n^{-\frac{1}{2}}).$$

Assumption 3.1.(i) ensures that the sieve variance is asymptotically nonzero. When $\rho(\alpha)$ is a nonlinear functional in α, Assumption 3.1.(ii) implies that there is a linear approximation $\partial_h \rho(\alpha_o)[h - h_o] + \partial_g \rho(\alpha_o)[g - g_o]$ for $\rho(\alpha)$ uniformly over $\alpha \in \mathcal{N}_n$ with approximation error $o(\|v_n^*\|_{sd} n^{-\frac{1}{2}})$. Assumption 3.1.(iii) implies that $\|\cdot\|_\varphi$ and $\|\cdot\|_\psi$ may be weaker than the pseudo-metrics $\|\cdot\|_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{G}}$ respectively. By Assumption 3.1.(iii) and the definition of $g_{o,n}$, we have

$$\|g_o - g_{o,n}\|_\psi \leq \|g_o - g_n\|_\psi \leq \|g_o - g_n\|_{\mathcal{G}} = O(\delta_{2,n}^*),$$

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 3.1.(iv) also requires that the sieve approximation error converges to zero at a rate faster than $n^{-\frac{1}{2}} \|v_n^*\|_{sd}$, which is an under-smoothing condition to derive the zero mean asymptotic normality of the sieve plug-in estimator $\rho(\widehat{\alpha}_n)$.

Define $\left(u_{h_n}^*, u_{g_n}^*, u_{\Gamma_n}^*\right) = \|v_n^*\|_{sd}^{-1} \left(v_{h_n}^*, v_{g_n}^*, v_{\Gamma_n}^*\right)$ and $g^* = g \pm \kappa_n u_{g_n}^*$ for any $g \in \mathcal{N}_{g,n}$, where $\kappa_n = o\left(n^{-1/2}\right)$ is some positive sequence. For any function f, $\mu_n(f) = n^{-1} \sum_{i=1}^n \left[f(Z_i) - E\left[f(Z_i)\right]\right]$ denotes the empirical process indexed by f.

Assumption 3.2 *(i) The following stochastic equicontinuity conditions hold:*

$$\sup_{\alpha \in \mathcal{N}_n} \left|\mu_n\left\{\psi(Z_2, g^*, h) - \psi(Z_2, g, h) - \Delta_\psi(Z_2, g, h)[\pm \kappa_n u_{g_n}^*]\right\}\right| = O_p(\kappa_n^2) \tag{10}$$

$$\text{and } \sup_{\alpha \in \mathcal{N}_n} \left|\mu_n\left\{\Delta_\psi(Z_2, g, h)[u_{g_n}^*] - \Delta_\psi(Z_2, g_o, h_o)[u_{g_n}^*]\right\}\right| = O_p(\kappa_n); \tag{11}$$

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

$$K_\psi(g, h) - K_\psi(g^*, h) = \mp \kappa_n \Gamma(\alpha_o)\left[h - h_o, u_{g_n}^*\right] + \frac{||g^* - g_o||_\psi^2 - ||g - g_o||_\psi^2}{2} + O(\kappa_n^2) \tag{12}$$

uniformly 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 (2014). Assumption 3.2.(ii) 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 (12), Assumption 3.2.(ii) is reduced to

$$\sup_{g \in \mathcal{N}_{g,n}} \left| K_\psi(g, h_o) - K_\psi(g^*, h_o) - \frac{||g^* - g_o||_\psi^2 - ||g - g_o||_\psi^2}{2} \right| = O(\kappa_n^2)$$

which is the condition used in Chen, Liao and Sun (2014) to derive the asymptotic normality of the one-step sieve plug-in estimator. As a result, we can view the extra term in (12) as the effect of the first-step estimator $\widehat{h}_n$ on the asymptotic distribution of the second-step sieve M estimator $\widehat{g}_n$.

Assumption 3.3 *(i) The first-step sieve M estimator $\widehat{h}_n$ satisfies*

$$\left| \langle \widehat{h}_n - h_o, u^*_{h_n} + u^*_{\Gamma_n} \rangle_\varphi - \mu_n \left\{ \Delta_\varphi(Z_1, h_o)[u^*_{h_n} + u^*_{\Gamma_n}] \right\} \right| = O_p(\kappa_n); \tag{13}$$

(ii) the following central limit theorem (CLT) holds

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

(iii) $\varepsilon_{2,n} = O(\kappa_n)$, $\kappa_n \delta_{2,n}^{*-1} = o(1)$ and $||u^*_{g_n}||_\psi = O(1)$.

Assumption 3.3.(i) is a high level condition, which is established in Chen, Liao and Sun (2014) under a set of sufficient conditions. Assumption 3.3.(ii) is implied by the triangle array CLTs. Assumption 3.3.(iii) implies that the optimization error $\varepsilon_{2,n}$ in the second-step sieve M estimation is of the same or larger order as κ_n. As $\delta^*_{2,n}$ 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 $\delta^*_{2,n}$ converges to zero at a rate not faster than root-n, which explains the assumption $\kappa_n \delta_{2,n}^{*-1} = o(1)$.

Theorem 3.1 *Suppose that Assumptions 3.2 and 3.3 hold. Then under Assumption 3.1.(i)-(iii), the sieve plug-in estimator $\rho(\widehat{h}_n, \widehat{g}_n)$ satisfies*

$$\frac{\sqrt{n} \left[\rho(\widehat{h}_n, \widehat{g}_n) - \rho(h_{o,n}, g_{o,n}) \right]}{\|v^*_n\|_{sd}} \to_d N(0, 1) \tag{14}$$

where $\|v^*_n\|_{sd}$ is defined in (9). Furthermore, if Assumption 3.1.(iv) is also satisfied, then

$$\frac{\sqrt{n} \left[\rho(\widehat{h}_n, \widehat{g}_n) - \rho(h_o, g_o) \right]}{\|v^*_n\|_{sd}} \to_d N(0, 1). \tag{15}$$

Proof. In Appendix B. ■

4 Consistent Variance Estimation

The Gaussian approximation in Theorem 3.1 suggests a natural method of inference as long as a consistent estimator of $\|v_n^*\|_{sd}$ is available. In this section, we provide such a consistent estimator. An alternative is to consider bootstrap based confidence sets, whose asymptotic validity could be established but is computationally more time-consuming.

For simplicity of notation, we will assume in this and the next two sections that the data are *i.i.d.*,[5] and that the criterion functions $\varphi(Z_1, h)$ and $\psi(Z_2, g, h)$ are respectively twice continuously pathwise differentiable with respect to h and (g, h) in a local neighborhood $\mathcal{N}_n$.

Our estimator is a natural sample analog of the $\|v_n^*\|_{sd}$, but for the sake of completeness, we provide the details below. We define

$$\Delta_\varphi(Z_1, h)[v_{h_1}] = \left.\frac{\partial \varphi(Z_1, h + \tau v_{h_1})}{\partial \tau}\right|_{\tau=0} \text{ and } r_\varphi(Z_1, h)[v_{h_1}, v_{h_2}] = \left.\frac{\partial \Delta_\varphi(Z_1, h + \tau v_{h_1})[v_{h_2}]}{\partial \tau}\right|_{\tau=0}$$

for any $v_{h_1}, v_{h_2} \in \mathcal{V}_{1,n}$. Similarly, we define $\Delta_\psi(Z_2, g, h)[v_{g_1}]$ and $r_\psi(Z_2, g, h)[v_{g_1}, v_{g_2}]$ for any $v_{g_1}, v_{g_2} \in \mathcal{V}_{2,n}$. We define the empirical Riesz representer $\widehat{v}_{h_n}^*$ by

$$\partial_h \rho(\widehat{\alpha}_n)[v_h] = \left\langle v_h, \widehat{v}_{h_n}^* \right\rangle_{n,\varphi} \text{ for any } v_h \in \mathcal{V}_{1,n} \tag{16}$$

where $\langle v_{h_1}, v_{h_2} \rangle_{n,\varphi} = -\frac{1}{n}\sum_{i=1}^n r_\varphi(Z_1, \widehat{h}_n)[v_{h_1}, v_{h_2}]$. Similarly, we define $\widehat{v}_{g_n}^*$ as

$$\partial_g \rho(\widehat{\alpha}_n)[v_g] = \left\langle v_g, \widehat{v}_{g_n}^* \right\rangle_{n,\psi} \text{ for any } v_g \in \mathcal{V}_{2,n} \tag{17}$$

where $\langle v_{g_1}, v_{g_2} \rangle_{n,\psi} = -\frac{1}{n}\sum_{i=1}^n r_\psi(Z_2, \widehat{\alpha}_n)[v_{g_1}, v_{g_2}]$; and $\widehat{v}_{\Gamma_n}^*$ as

$$\Gamma_n(\widehat{\alpha}_n)\left[v_h, \widehat{v}_{g_n}^*\right] = \left\langle v_h, \widehat{v}_{\Gamma_n}^* \right\rangle_{n,\varphi} \text{ for any } v_h \in \mathcal{V}_{1,n}, \tag{18}$$

where

$$\Gamma_n(\widehat{\alpha}_n)\left[v_h, \widehat{v}_{g_n}^*\right] = \frac{1}{n}\sum_{i=1}^n \left.\frac{\partial \Delta_\psi(Z_2, \widehat{g}_n, \widehat{h}_n + \tau v_h)[\widehat{v}_{g_n}^*]}{\partial \tau}\right|_{\tau=0}.$$

Based on these ingredients, we propose to estimate $\|v_n^*\|_{sd}^2$ by

$$\|\widehat{v}_n^*\|_{n,sd}^2 = \frac{1}{n}\sum_{i=1}^n \left|\Delta_\varphi(Z_{1,i}, \widehat{h}_n)[\widehat{v}_{h_n}^* + \widehat{v}_{\Gamma_n}^*] + \Delta_\psi(Z_{2,i}, \widehat{g}_n, \widehat{h}_n)[\widehat{v}_{g_n}^*]\right|^2. \tag{19}$$

Our estimator $\|\widehat{v}_n^*\|_{n,sd}^2$ is nothing but a natural sample analog of $\|v_n^*\|_{sd}$.

Let $\mathcal{W}_{1,n} \equiv \{h \in \mathcal{V}_{1,n} : \|h\|_\varphi \leq 1\}$ and $\mathcal{W}_{2,n} \equiv \{g \in \mathcal{V}_{2,n} : \|g\|_\psi \leq 1\}$ and $\langle v_{g_1}, v_{g_2}\rangle_\psi = E\{-r_\psi(Z_2, \alpha_o)[v_{g_1}, v_{g_2}]\}$ for any $v_{g_1}, v_{g_2} \in \mathcal{V}_2$. We next list the assumptions needed for showing the consistency of the sieve variance estimator.

Assumption 4.1 *The following conditions hold:*

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

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

(iii) $\sup_{\alpha\in\mathcal{N}_n, v_g\in\mathcal{W}_{2,n}} |\partial_g\rho(\alpha)[v_g] - \partial_g\rho(\alpha_o)[v_g]| = o(1)$.

Assumptions 4.1.(i)-(ii) are useful to show that the empirical inner product $\langle v_{g_1}, v_{g_2}\rangle_{n,\psi}$ and the population inner product $\langle v_{g_1}, v_{g_2}\rangle_\psi$ are asymptotically equivalent uniformly over $v_{g_1}, v_{g_2} \in \mathcal{W}_{2,n}$. Assumption 4.1.(iii) implies that the empirical pathwise derivative $\partial_g\rho(\widehat{\alpha}_n)[v_g]$ is close to the theoretical pathwise derivative $\partial_g\rho(\alpha_o)[v_g]$ uniformly over $v_g \in \mathcal{W}_{2,n}$. When the functional $\rho(\alpha)$ is linear in g, Assumption 4.1.(iii) is trivially satisfied.

Let $\mathcal{B}_{2,n}^* \equiv \{v \in \mathcal{V}_{2,n} : \|v - v_{g_n}^*\|_\psi \|v_{g_n}^*\|_\psi^{-1} \leq \delta_{v_g,n}\}$, where $\delta_{v_g,n} = o(1)$ is some positive sequence such that $\widehat{v}_{g_n}^* \in \mathcal{B}_{2,n}^*$ wpa1 and $\langle v_{h_1}, v_{h_2}\rangle_\varphi = E[-r_\varphi(Z_1, h_o)[v_{h_1}, v_{h_2}]]$ for any $v_{h_1}, v_{h_2} \in \mathcal{V}_1$.

Assumption 4.2 *The following conditions hold:*

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

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

(iii) $\sup_{\alpha\in\mathcal{N}_n, v_h\in\mathcal{W}_{1,n}} |\partial_h\rho(\alpha)[v_h] - \partial_h\rho(\alpha_o)[v_h]| = o(1)$;

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

Assumptions 4.2.(i)-(ii) are useful to show the empirical inner product $\langle v_{h_1}, v_{h_2}\rangle_{n,\varphi}$ and the population inner product $\langle v_{h_1}, v_{h_2}\rangle_\varphi$ are asymptotically equivalent uniformly over $v_{h_1}, v_{h_2} \in \mathcal{W}_{1,n}$. Assumption 4.2.(iii) implies that the empirical pathwise derivative $\partial_h\rho(\widehat{\alpha}_n)[v_h]$ is close to the population pathwise derivative $\partial_h\rho(\alpha_o)[v_h]$ uniformly over $v_h \in \mathcal{W}_{1,n}$. Assumption 4.2.(iii) is trivially satisfied, if the functional $\rho(\alpha)$ is linear in h. Assumption 4.2.(iv) is needed to show that the functional $\Gamma(\alpha_o)\left[v_h, v_{g_n}^*\right]$ is consistently estimated by its empirical counterpart $\Gamma_n(\widehat{\alpha}_n)\left[v_h, \widehat{v}_{g_n}^*\right]$ uniformly over $v_h \in \mathcal{W}_{1,n}$.

Assumption 4.3 *Let $\|\cdot\|_2$ denote the $L_2(dF_Z)$-norm, then: (i) the functional $\Delta_\varphi(Z_1, h)[v_h]$ satisfies*

$$\sup_{h\in\mathcal{N}_{h,n}, v_h\in\mathcal{W}_{1,n}} \|\Delta_\varphi(Z_1,h)[v_h] - \Delta_\varphi(Z_1,h_o)[v_h]\|_2 = o(1), \tag{20}$$

$$\textit{and} \sup_{h\in\mathcal{N}_{h,n}, v_h\in\mathcal{W}_{1,n}} \left|\mu_n\left\{\Delta^2_\varphi(Z_1,h)[v_h]\right\}\right| = o_p(1); \tag{21}$$

(ii) the functional $\Delta_\psi(Z_2, \alpha)[v_g]$ satisfies

$$\sup_{\alpha\in\mathcal{N}_n, v_g\in\mathcal{W}_{2,n}} \|\Delta_\psi(Z_2,\alpha)[v_g] - \Delta_\psi(Z_2,\alpha_o)[v_g]\|_2 = o(1), \tag{22}$$

$$\textit{and} \sup_{\alpha\in\mathcal{N}_n, v_g\in\mathcal{W}_{2,n}} \left|\mu_n\left\{\Delta^2_\psi(Z_2,\alpha)[v_g]\right\}\right| = o_p(1); \tag{23}$$

(iii) the following condition holds

$$\sup_{\alpha\in\mathcal{N}_n, v_g\in\mathcal{W}_{2,n}, v_h\in\mathcal{W}_{1,n}} |\mu_n\{\Delta_\varphi(Z_1,h)[v_h]\Delta_\psi(Z_2,\alpha)[v_g]\}| = o_p(1); \tag{24}$$

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

$$\frac{\left\|v^*_{h_n}\right\|_\varphi + \left\|v^*_{\Gamma_n}\right\|_\varphi + \left\|v^*_{g_n}\right\|_\psi}{\left\|\Delta_\varphi(Z_1,h_o)[v^*_{h_n} + v^*_{\Gamma_n}] + \Delta_\psi(Z_2,\alpha_o)[v^*_{g_n}]\right\|_2} = O(1). \tag{25}$$

Conditions (20) and (22) require some local smoothness of $\Delta_\varphi(Z_1,h)[\cdot]$ in $h\in\mathcal{N}_{h,n}$ and $\Delta_\psi(Z_2,\alpha)[\cdot]$ in $\alpha\in\mathcal{N}_n$ respectively. Conditions (21), (23) and (24) can be verified by applying uniform law of large numbers from the empirical process theory. To illustrate Assumption 4.3.(iv), suppose that $v_h(\cdot) = R(\cdot)'\gamma$ for some $\gamma\in\mathbb{R}^{L(n)}$ and $v_g(\cdot) = P(\cdot)'\beta$ for some $\beta\in\mathbb{R}^{K(n)}$ where $R(\cdot) = [r_1(\cdot),\ldots,r_{K(n)}(\cdot)]'$ and $P(\cdot) = [p_1(\cdot),\ldots,p_{L(n)}(\cdot)]'$ are two vectors of basis functions. Let

$$\Delta_\varphi(Z_1,h_o)[R] = \left[\Delta_\varphi(Z_1,h_o)[r_1],\ldots,\Delta_\varphi(Z_1,h_o)[r_{L(n)}]\right]' \text{ and}$$
$$\Delta_\psi(Z_2,\alpha_o)[P] = \left[\Delta_\psi(Z_2,\alpha_o)[p_1],\ldots,\Delta_\psi(Z_2,\alpha_o)[p_{K(n)}]\right]'.$$

In most cases, Assumption 4.3.(iv) is implied by the regularity conditions on the eigenvalues of the matrices $E[\Delta_\varphi(Z_1,h_o)[R]\Delta_\varphi(Z_1,h_o)[R]']$ and $E[\Delta_\psi(Z_2,\alpha_o)[P]\Delta_\psi(Z_2,\alpha_o)[P]']$.

Theorem 4.1 *Suppose that the data are i.i.d. and the conditions in Theorem 3.1 are satisfied. Then*

under Assumptions 4.1, 4.2 and 4.3, we have

$$\left| \|\widehat{v}_n^*\|_{n,sd} \Big/ \|v_n^*\|_{sd} - 1 \right| = o_p(1).$$

Proof. In Appendix C. ■

Based on Theorem 3.1 and Theorem 4.1, we can obtain the conclusion that

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

which can be used to construct confidence bands for $\rho(h_o, g_o)$.

5 An Application

In this section, we illustrate the high level conditions and the main results established in the previous sections in a two-step nonparametric regression example. Suppose we have i.i.d. data $\{y_i, x_i, s_i\}_{i=1}^n$ from the following model:

$$\begin{aligned} y_i &= g_o(\varepsilon_i) + u_i, \\ s_i &= h_o(x_i) + \varepsilon_i, \end{aligned} \tag{26}$$

where $E\left[\varepsilon_i | x_i\right] = 0$, $E\left[u_i | x_i, \varepsilon_i\right] = 0$, h_o and g_o are unknown parameters. The parameter of interest is $\rho\left(g_o\right)$, where the functional $\rho(\cdot)$ is known.

For ease of notation, we assume that y_i, x_i and s_i are univariate random variables. The first-step M estimator is

$$\widehat{h}_n = \arg\max_{h \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^n \left[-\frac{1}{2}(s_i - h(x_i))^2 \right] \tag{27}$$

where $\mathcal{H}_n = \left\{h : h\left(\cdot\right) = R(\cdot)'\gamma,\, \gamma \in \mathbb{R}^{L(n)}\right\}$. Let $R(x) = \left[r_1(x), \ldots, r_{L(n)}(x)\right]'$ and $R_n = [R(x_1), \ldots, R(x_n)]$. The first step M estimator $\widehat{h}_n$ has a closed form expression

$$\widehat{h}_n(\cdot) = R(x)' \left(R_n R_n'\right)^{-1} R_n S_n = R(\cdot)'\widehat{\gamma}_n \tag{28}$$

where $S_n = [s_1, \ldots, s_n]'$. From the first step estimator, we calculate $\widehat{\varepsilon}_i = s_i - \widehat{h}_n(x_i)$ for $i = 1, \ldots, n$. Let

$P(\varepsilon) = \left[p_1(\varepsilon), \ldots, p_{K(n)}(\varepsilon)\right]'$ and $\widehat{P}_n = [P(\widehat{\varepsilon}_1), \ldots, P(\widehat{\varepsilon}_n)]'$. The second-step M estimator is

$$\widehat{g}_n = \arg\max_{g \in \mathcal{G}_n} \frac{1}{n} \sum_{i=1}^{n} \left[-\frac{1}{2} (y_i - g(\widehat{\varepsilon}_i))^2 \right] \tag{29}$$

where $\mathcal{G}_n = \left\{ g : g(\cdot) = P(\cdot)'\beta,\, \beta \in \mathbb{R}^{K(n)} \right\}$. The second step M estimator $\widehat{g}_n$ also has a closed form expression

$$\widehat{g}_n(\varepsilon) = P(\varepsilon)'(\widehat{P}_n'\widehat{P}_n)^{-1}\widehat{P}_n' Y_n = P(\varepsilon)'\widehat{\beta}_n, \tag{30}$$

where $Y_n = [y_1, \ldots, y_n]'$.

Using the specific forms of the criterion functions in the M estimations, we have $\langle v_{h_1}, v_{h_2} \rangle_\varphi = E\left[v_{h_1}(x) v_{h_2}(x)\right]$ for any $v_{h_1}, v_{h_2} \in \mathcal{V}_1$, and $\langle v_{g_1}, v_{g_2} \rangle_\psi = E\left[v_{g_1}(\varepsilon) v_{g_2}(\varepsilon)\right]$ for any $v_{g_1}, v_{g_2} \in \mathcal{V}_2$. As $\rho(\cdot)$ only depends on g_o, we use $\partial\rho(g_o)[\cdot]$ to denote the linear functional defined in (6). As $\mathcal{V}_{2,n}$ is the linear space spanned by the basis functions $P(\varepsilon)$, the Riesz representer $v_{g_n}^*$ of the functional $\partial\rho(g_o)[\cdot]$ has a closed form expression

$$v_{g_n}^*(\varepsilon) = \partial\rho(g_o)[P]' Q_{K(n)}^{-1} P(\varepsilon), \tag{31}$$

where $\partial\rho(g_o)[P]' = [\partial\rho(g_o)[p_1], \ldots, \partial\rho(g_o)[p_{K(n)}]]$ and $Q_{K(n)} = E\left[P(\varepsilon)P(\varepsilon)'\right]$. By definition,

$$\Gamma(\alpha_o)\left[v_h, v_g\right] = E\left[\partial g_o(\varepsilon) v_h(x) v_g(\varepsilon)\right], \tag{32}$$

where $\partial g_o(\varepsilon) = \partial g_o(\varepsilon)/\partial\varepsilon$. As $\mathcal{V}_{1,n}$ is the linear space spanned by the basis functions $R(x)$, the Riesz representer $v_{\Gamma_n}^*$ of the functional $\Gamma(\alpha_o)\left[v_h, v_{g_n}^*\right]$ has the closed form expression

$$v_{\Gamma_n}^*(x) = E\left[\partial g_o(\varepsilon) v_{g_n}^*(\varepsilon) R(x)'\right] Q_{L(n)}^{-1} R(x), \tag{33}$$

where $Q_{L(n)} = E\left[R(x)R(x)'\right]$. Given the sieve Riesz representers $v_{g_n}^*$ and $v_{\Gamma_n}^*$, we have

$$\Delta_\varphi(Z_{1,i}, h_o)[v_{\Gamma_n}^*] + \Delta_\psi(Z_{2,i}, \alpha_o)[v_{g_n}^*] = v_{\Gamma_n}^*(x_i)\varepsilon_i + v_{g_n}^*(\varepsilon_i) u_i \tag{34}$$

which implies that the variance of the plug-in estimator $\rho\left(\widehat{g}_n\right)$ takes the following form:

$$\|v_n^*\|_{sd}^2 = \left\| v_{\Gamma_n}^*(x)\varepsilon \right\|_2^2 + \left\| v_{g_n}^*(\varepsilon) u \right\|_2^2. \tag{35}$$

By definition, the empirical Riesz representers $\widehat{v}^*_{g_n}(\varepsilon)$ and $\widehat{v}^*_{\Gamma_n}(x)$ are

$$\widehat{v}^*_{g_n}(\varepsilon) = \partial\rho(\widehat{g}_n)[P]'\widehat{Q}^{-1}_{n,K(n)}P(\varepsilon) \text{ and } \widehat{v}^*_{\Gamma_n}(x) = n^{-1}\sum_{i=1}^{n}\partial\widehat{g}_n(\widehat{\varepsilon}_i)\widehat{v}^*_{g_n}(\widehat{\varepsilon}_i)R(x_i)'Q^{-1}_{n,L(n)}R(x), \tag{36}$$

respectively, where $\widehat{Q}_{n,K(n)} = n^{-1}\widehat{P}'_n\widehat{P}_n$ and $Q_{n,L(n)} = n^{-1}R_nR'_n$. Hence the variance estimator of the plug-in estimator $\rho(\widehat{g}_n)$ is

$$\|\widehat{v}^*_n\|^2_{n,sd} = n^{-1}\sum_{i=1}^{n}\left[(\widehat{v}^*_{\Gamma_n}(x_i)\widehat{\varepsilon}_i)^2 + (\widehat{v}^*_{g_n}(\widehat{\varepsilon}_i)\widehat{u}_i)^2\right], \tag{37}$$

where $\widehat{u}_i = y_i - \widehat{g}_n(\widehat{\varepsilon}_i)$.

Theorem 5.1 *Under Assumptions D.1, D.2 and D.3 in Appendix D, we have*

$$n^{1/2}\|v^*_n\|^{-1}_{sd}\left[\rho(\widehat{g}_n) - \rho(g_o)\right] \to_d N(0,1).$$

Moreover, if Assumption D.4 in Appendix D also holds, then

$$n^{1/2}\|\widehat{v}^*_n\|^{-1}_{n,sd}\left[\rho(\widehat{g}_n) - \rho(g_o)\right] \to_d N(0,1).$$

6 Practical Implication

This section contains the main contribution of the paper, based on a convenient characterization of $\|\widehat{v}^*_n\|_{n,sd}$. For this purpose, we will assume that a researcher incorrectly believes that the infinite dimensional parameter (h_o, g_o) is in fact finite dimensional, and calculates the estimator of the asymptotic variance using a standard method available in, e.g., Wooldridge (2002, Chapter 12). We will show that our estimator $\|\widehat{v}^*_n\|_{n,sd}$ is in fact identical to such a standard estimator.

For illustration, we will assume that $g_o = (\theta_o, m_o(\cdot))$ and one is interested in the estimation and inference of $\rho(\alpha_o) = \lambda'\theta_o$, and (ii) $\rho(\alpha_o) = \rho(m_o)$, where $\lambda \in R^{d_\theta}$, $\lambda \neq 0$ and $\rho(\cdot)$ only depends on m_o. While it may appear restrictive, the algebra carries over to different parameters as well. Because the notation becomes complicated for some other parameters, we restrict our attention to these parameters.

Let $\mathcal{H}_n = \{h(\cdot) = R_{L(n)}(\cdot)'\gamma : \gamma \in \mathbb{R}^{L(n)}\}$ be the sieve space for a real-valued unknown function h_o in the first step, where $R_{L(n)}(\cdot) = \left[r_1(\cdot), ..., r_{L(n)}(\cdot)\right]'$ denote a $L(n) \times 1$ vector of basis functions. Let $\mathcal{G}_n = \Theta \times \mathcal{M}_n$ be the sieve space for $g_o = (\theta_o, m_o(\cdot))$ in the second step, where $\mathcal{M}_n = \{m(\cdot) =$

$P_{K(n)}(\cdot)'\beta : \beta \in \mathbb{R}^{K(n)}\}$, $P_{K(n)}(\cdot) = \left[p_1(\cdot), ..., p_{K(n)}(\cdot)\right]'$ denote a $K(n) \times 1$ vector of basis functions for $m_o(\cdot)$, and $\overline{K}(n) = \dim(\mathcal{G}_n) = d_\theta + K(n)$. For notational simplicity, we will often omit the $L(n)$ and $K(n)$ subscripts, and write $R_{L(n)}(\cdot) = R(\cdot)$ and $P_{K(n)}(\cdot) = P(\cdot)$. As for misspecification, we assume that there is a researcher who believes that the unknown parameters h_o and g_o should be specified as

$$h_o(\cdot) = R(\cdot)'\gamma_{o,L} \text{ and } g_o(\cdot) = (\theta_o, m_o(\cdot)) = (\theta_o, P(\cdot)'\beta_{o,K}),$$

where L and K happen to be fixed at $L(n)$ and $K(n)$.

First, we provide an explicit characterization of our $\|\widehat{v}_n^*\|_{n,sd}$ for the sieve plug-in estimates $\rho(\widehat{\alpha}_n) = \lambda'\widehat{\theta}_n$ and $\rho(\widehat{\alpha}_n) = \rho(\widehat{m}_n)$. For this purpose, we introduce some notation. Let $\overline{P}(\cdot) = (\mathbf{1}'_{d_\theta}, P(\cdot)')'$ be a $\overline{K}(n) \times 1$ vector and $\overline{p}_a(\cdot)$ denote its a-th component for $a = 1, \ldots, \overline{K}(n)$. Let $r_\psi(Z_2, \alpha)[\overline{P}, \overline{P}]$ and $r_\varphi(Z_1, h)[R, R]$ denote $\overline{K}(n) \times \overline{K}(n)$ and $L(n) \times L(n)$ matrices with the (a, b)-th element being $r_\psi(Z_2, \alpha)[\overline{p}_a, \overline{p}_b]$ and $r_\varphi(Z_1, h)[r_a, r_b]$ respectively. Similarly, we use $\Gamma(\alpha_o)\left[R, \overline{P}\right]$ to define the $L(n) \times \overline{K}(n)$ matrix with the (a, b)-th element being $\Gamma(\alpha_o)[r_a, \overline{p}_b]$. Finally, we let

$$\begin{pmatrix} \widehat{I}_{\theta\theta,n} & \widehat{I}_{\theta m,n} \\ \widehat{I}_{m\theta,n} & \widehat{I}_{mm,n} \end{pmatrix} = \frac{1}{n}\sum_{i=1}^n -r_\psi(Z_{2,i}, \widehat{\alpha}_n)[\overline{P}, \overline{P}],$$
$$\widehat{I}^{(1)}_{\theta\theta,n} = \left(\widehat{I}_{\theta\theta,n} - \widehat{I}_{\theta m,n}\widehat{I}^{-1}_{mm,n}\widehat{I}_{m\theta,n}\right)^{-1},$$
$$\widehat{I}^{(1)}_{mm,n} = \left(\widehat{I}_{mm,n} - \widehat{I}_{m\theta,n}\widehat{I}^{-1}_{\theta\theta,n}\widehat{I}_{\theta m,n}\right)^{-1},$$

and $\widehat{V}_{g,i} = (\Gamma_n(\widehat{\alpha}_n)\left[R, \overline{P}\right])'\left[n^{-1}\sum_{i=1}^n -r_\varphi(Z_{1,i}, \widehat{h}_n)[R, R]\right]^{-1}\Delta_\varphi(Z_{1,i}, \widehat{h}_n)[R] + \Delta_\psi(Z_{2,i}, \widehat{\alpha}_n)[\overline{P}]$. After some straightforward yet tedious algebra, we can establish the following:

Proposition 6.1 *Our estimate of the asymptotic variance of $\rho(\widehat{\alpha}_n) = \lambda'\widehat{\theta}_n$ is equal to*

$$\left\|\widehat{v}^*_{\theta,n}\right\|^2_{sd} = \lambda'\left(\widehat{I}^{(1)}_{\theta\theta,n},\ -\widehat{I}^{(1)}_{\theta\theta,n}\widehat{I}_{\theta m,n}\widehat{I}^{-1}_{mm,n}\right)\frac{\sum_{i=1}^n \widehat{V}_{g,i}\widehat{V}'_{g,i}}{n}\left(\widehat{I}^{(1)}_{\theta\theta,n},\ -\widehat{I}^{(1)}_{\theta\theta,n}\widehat{I}_{\theta m,n}\widehat{I}^{-1}_{mm,n}\right)'\lambda. \tag{38}$$

Our estimate of the asymptotic variance of $\rho(\widehat{\alpha}_n) = \rho(\widehat{m}_n)$ is equal to

$$\left\|\widehat{v}^*_{m,n}\right\|^2_{sd} = \partial\rho(\widehat{m}_n)[P]'\left(-\widehat{I}^{(1)}_{mm,n}\widehat{I}_{m\theta,n}\widehat{I}^{-1}_{\theta\theta,n},\ \widehat{I}^{(1)}_{mm,n}\right)\frac{\sum_{i=1}^n \widehat{V}_{g,i}\widehat{V}'_{g,i}}{n}\left(-\widehat{I}^{(1)}_{mm,n}\widehat{I}_{m\theta,n}\widehat{I}^{-1}_{\theta\theta,n},\ \widehat{I}^{(1)}_{mm,n}\right)'\partial\rho(\widehat{m}_n)[P], \tag{39}$$

where $\partial\rho(\widehat{m}_n)[P] = \left[\partial\rho(\widehat{m}_n)[p_1], \ldots, \partial\rho(\widehat{m}_n)[p_{K(n)}]\right]'$.

Proof. In Appendix C. ■

We now compute the standard estimator of the asymptotic variance when a researcher believes that the unknown parameters h_o and g_o should be specified as

$$h_o(\cdot) = R(\cdot)'\gamma_{o,L} \text{ and } g_o(\cdot) = (\theta_o, m_o(\cdot)) = (\theta_o, P(\cdot)'\beta_{o,K}) \equiv \overline{P}(\cdot)'\beta_{g_o},$$

where L and $\overline{K} = d_\theta + K$ happen to be fixed at $L(n)$ and $\overline{K}(n) = d_\theta + K(n)$, and we denote $(\theta_o', \beta_{o,K}')' \equiv \beta_{g_o}$. The unknown parameter $\gamma_{o,L}$ is estimated by the following parametric M estimation

$$\widehat{\gamma}_h = \underset{\gamma_L \in \mathcal{B}_h}{\arg\max} \frac{1}{n} \sum_{i=1}^{n} \varphi(Z_{1,i}, R(\cdot)'\gamma_L), \tag{40}$$

where $\mathcal{B}_h$ is some compact set in $\mathbb{R}^L$. After the first-step estimate $\widehat{\gamma}_h$ is available, β_{g_o} is estimated by the following second-step parametric M estimation

$$(\widehat{\theta}, \widehat{\beta}_K) = \underset{(\theta,\beta_K)\in\Theta\times\mathcal{B}_m}{\arg\max} \frac{1}{n} \sum_{i=1}^{n} \psi(Z_{2,i}, \theta, P(\cdot)'\beta_K, R(\cdot)'\widehat{\gamma}_h), \tag{41}$$

where Θ and $\mathcal{B}_m$ are some compact sets in $\mathbb{R}^{d_\theta}$ and $\mathbb{R}^K$ respectively.

Define $\beta_g' \equiv (\theta', \beta_K')$ and accordingly $\widehat{\beta}_g' \equiv (\widehat{\theta}', \widehat{\beta}_K')$. Under the standard regularity conditions, one can show that

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta} - \theta_o) \\ \sqrt{n}(\widehat{\beta}_K - \beta_{o,K}) \end{pmatrix} \to_d N(0, \psi_{22}^{-1} V_{22} \psi_{22}^{-1}) \overset{d}{=} N(0, V_{\beta_g}) \tag{42}$$

where

$$V_{22} = E\left[\left[\psi_2(Z_2) + \psi_{21}\varphi_{11}^{-1}\varphi_1(Z_1)\right]\left[\psi_2(Z_2) + \psi_{21}\varphi_{11}^{-1}\varphi_1(Z_1)\right]'\right],$$

$$\varphi_{11} = -E\left[\frac{\partial^2 \varphi(Z_1, h_o(\cdot))}{\partial\gamma_L \partial\gamma_L'}\right],\ \psi_2(Z_2) = \frac{\partial\psi(Z_2, g_o(\cdot), h_o(\cdot))}{\partial\beta_g},\ \varphi_1(Z_1) = \frac{\partial\varphi(Z_1, h_o(\cdot))}{\partial\gamma_L}$$

$$\psi_{21} = E\left[\frac{\partial^2\psi(Z_2, g_o(\cdot), h_o(\cdot))}{\partial\gamma_L\partial\beta_g'}\right] \text{ and } \psi_{22} = -E\left[\frac{\partial^2\psi(Z_2, g_o(\cdot), h_o(\cdot))}{\partial\beta_g\partial\beta_g'}\right].$$

Let $I_{\psi,\theta\theta}$ and $I_{\psi,\beta\beta}$ denote the leading $d_\theta \times d_\theta$ and last $K \times K$ submatrices of ψ_{22}. Accordingly, $I_{\psi,\theta\beta}$ and $I_{\psi,\beta\theta}$ denote the upper-right $d_\theta \times K$ and lower-left $K \times d_\theta$ submatrices of ψ_{22}. Define

$$I_{\psi,\theta\theta}^{(1)} = (I_{\psi,\theta\theta} - I_{\psi,\theta\beta}I_{\psi,\beta\beta}^{-1}I_{\psi,\beta\theta})^{-1} \text{ and } I_{\psi,\beta\beta}^{(1)} = (I_{\psi,\beta\beta} - I_{\psi,\beta\theta}I_{\psi,\theta\theta}^{-1}I_{\psi,\theta\beta})^{-1}.$$

From (42), we can obtain after some algebra that

$$\sqrt{n}(\widehat{\theta}-\theta_o)\to_d N(0,V_\theta) \text{ and } \sqrt{n}(\widehat{\beta}_K-\beta_{o,K})\to_d N(0,V_{\beta_K})$$

where

$$V_\theta=\left(I^{(1)}_{\psi,\theta\theta},\ -I^{(1)}_{\psi,\theta\theta}I_{\psi,\theta\beta}I^{-1}_{\psi,\beta\beta}\right)V_{22}\left(I^{(1)}_{\psi,\theta\theta},\ -I^{(1)}_{\psi,\theta\theta}I_{\psi,\theta\beta}I^{-1}_{\psi,\beta\beta}\right)', \tag{43}$$

$$V_{\beta_K}=\left(-I^{(1)}_{\psi,\beta\beta}I_{\psi,\beta\theta}I^{-1}_{\psi,\theta\theta},\ I^{(1)}_{\psi,\beta\beta}\right)V_{22}\left(-I^{(1)}_{\psi,\beta\beta}I_{\psi,\beta\theta}I^{-1}_{\psi,\theta\theta},\ I^{(1)}_{\psi,\beta\beta}\right)'. \tag{44}$$

Equations (43) and (44) suggest sample analog estimators $\widehat{V}_\theta$ and $\widehat{V}_{\beta_m}$ based on empirical counterparts of $I_{\psi,\theta\theta}$, $I_{\psi,\beta\beta}$, etc.[6] Because

$$\sqrt{n}(\lambda'\widehat{\theta}-\lambda'\theta_o)\to_d N\left(0,\lambda'V_\theta\lambda\right)$$

and

$$\sqrt{n}(\rho(\widehat{m})-\rho(m_o))=\sqrt{n}\partial\rho(m_o)[P]'(\widehat{\beta}_K-\beta_{K,o})+o_p(1)\to_d N\left(0,\partial\rho(m_o)[P]'V_{\beta_K}\partial\rho(m_o)[P]\right) \tag{45}$$

under this parametric (mis)specification, the "parametric" version of the variance estimate is given by $\lambda'\widehat{V}_\theta\lambda$ and $\partial\rho(m_o)[P]'V_{\beta_K}\partial\rho(m_o)[P]$.

Comparing (38) with (43), and also (39) with (44), we obtain the following implication:

Theorem 6.2 *We have* $||\widehat{v}^*_{\theta,n}||^2_{sd}=\lambda'\widehat{V}_\theta\lambda$ *for any* λ, *and* $\left\|\widehat{v}^*_{m,n}\right\|^2_{sd}=\partial\rho(\widehat{m}_n)[P]'\widehat{V}_{\beta_K}\partial\rho(\widehat{m}_n)[P]$.

Proof. In Appendix C. ■

Remark 6.1. Theorem 6.2 implies that one can use the variance-covariance formula of parametric two-step estimation to construct the sieve variance estimates, as long as the number of basis functions used to approximate the unknown functions is the same as in the parametric approximation to the model. As a result, the numerical equivalence established above simplifies the semi/nonparametric inference based on the Gaussian approximation in empirical applications. It should be noted that the above equivalence only holds in finite samples, because the asymptotic theory of the sieve method requires $L(n)$ and $\overline{K}(n)=d_\theta+K(n)$ to diverge to infinity with the sample size n, while the parametric specification has L and $\overline{K}=d_\theta+K$ as finite constants irrespective of the sample size n.

Remark 6.2. Theorem 6.2, by justifying a parametric procedure even when the parametric specification may not be valid, provides a convenient procedure to practitioners. It is even more important because

the asymptotic variances of many two-step sieve estimators do not have simple analytic expressions, and as a consequence, it is often difficult to estimate the asymptotic variance from a correct nonparametric perspective.

Remark 6.3. Theorem 6.2 can be understood as a generalization of a similar result in Ackerberg, Chen and Hahn (2012), who established numerical equivalence for two-step estimators that do not involve generated regressors. We also note that our result is applicable to cases where the second-step estimator is not necessarily $\sqrt{n}$ consistent.

Remark 6.4. Theorem 6.2 can also be understood to be a general result that nests similar results available in the literature. For Example 2.1, Newey, Powell and Vella (1999) establishes that standard parametric standard error formula continues to be valid. Das, Newey and Vella (2003) establishes the same result for Example 2.2. See also Newey (2009). All these papers note that estimators of the asymptotic variance under parametric misspecification are in fact consistent for the correct asymptotic variances, but the equivalence results were all established for their specific models.

Remark 6.5. Theorem 6.2 can also be used to compute a simple consistent estimator of the asymptotic variance for the Olley and Pakes (1996) estimator described in Example 2.3, as long as the parameters are estimated by the method of linear sieve/series. Olley and Pakes (1996) did indeed calculate a series two-step estimator, but cautioned on p. 1279 that they were not aware of a theorem that insures $\sqrt{n}$ consistency and asymptotic normality when the series estimator is used. It is therefore not clear to us how the standard errors for the series based estimator in column (8) of their Table VI were calculated. Assuming that they used a standard parametric procedure, which is likely, our result in Theorem 6.2 provides a theoretical justification of their standard error calculation.

6.1 A simple illustration of Theorem 6.2

Although the asymptotic variance of the two-step sieve estimators does not have a simple analytic expression in general, our result can also be used to derive the analytic expression of the asymptotic variance when an estimator is sufficiently simple. As an example, we consider the simple model studied by Li and Wooldridge (2002), and note that our sieve variance formula coincides with that in the asymptotic

distribution of their Conjecture 2.1. Specifically, we consider the following model

$$y_i = w_i\theta_o + m_o(\varepsilon_i) + u_i, \tag{46}$$

$$s_i = h_o(x_i) + \varepsilon_i, \tag{47}$$

where $E[\varepsilon_i|x_i] = 0$ and $E[u_i|w_i, x_i, \varepsilon_i] = 0$. Li and Wooldridge (2002) study a special case of the above example, because the unknown function $h_o(x_i)$ is parametrically specified as $h_o(x_i) = x_i'\gamma_o$. For simplicity of notation, we assume that θ_o is a scalar.

In the first-step, h_o is estimated by a series nonparametric regression

$$\widehat{h}_n = \underset{h\in\mathcal{H}_n}{\arg\max} -\frac{1}{2n}\sum_{i=1}^{n}[s_i - h(x_i)]^2, \tag{48}$$

where $\mathcal{H}_n = \{h(\cdot) = R(\cdot)'\gamma : \gamma \in \mathbb{R}^{L(n)}\}$ and $R(\cdot)$ is defined in Section 6. In the second step, θ_o and $m_o(\cdot)$ are estimated by

$$(\widehat{\theta}_n, \widehat{m}_n) = \underset{(\theta,m)\in\Theta\times\mathcal{M}_n}{\arg\max} \frac{-1}{2n}\sum_{i=1}^{n}\left[y_i - w_i\theta - m(s_i - \widehat{h}_n(x_i))\right]^2, \tag{49}$$

where $\mathcal{M}_n = \{m(\cdot) = P(\cdot)'\beta : \beta \in \mathbb{R}^{K(n)}\}$ and $P(\cdot)$ is defined in Section 6.

In this example, we have

$$[E\{-r_\varphi(Z_1, h_o)[R, R]\}] = E\left[R(x)R(x)'\right] \text{ and } \Delta_\varphi(Z_1, h_o)[R] = \varepsilon R(x)$$

in the first-step M estimation; and

$$E\left\{-r_\psi(Z_2, \alpha_o)\left[\overline{P}, \overline{P}\right]\right\} = \begin{pmatrix} E\left[w^2\right] & E\left[wP(\varepsilon)'\right] \\ E\left[P(\varepsilon)w\right] & E\left[P(\varepsilon)P(\varepsilon)'\right] \end{pmatrix}, \tag{50}$$

$$\Gamma(\alpha_o)\left[R, \overline{P}\right] = E\left[\partial m_o(\varepsilon)R(x)\begin{bmatrix} w & P(\varepsilon)' \end{bmatrix}\right], \tag{51}$$

$$\Delta_\psi(Z_2, \alpha_o)[\overline{P}] = u\left[w, P(\varepsilon)'\right]', \tag{52}$$

where $\partial m_o(\varepsilon) = \partial m_o(\varepsilon)/\partial\varepsilon$, in the second step M estimation. (We omit the i subscript whenever obvious.)

Equations (50), (51), and (52) imply that[7]

$$\frac{\left\|v^*_{\theta,n}\right\|^2_{sd}}{(I^{(1)}_{\theta\theta,n})^2} = E\left[\partial m_o(\varepsilon)\widetilde{w}_n R(x)'\right] Q^{-1}_{L(n)} Q_{\varepsilon,L(n)} Q^{-1}_{L(n)} E\left[\partial m_o(\varepsilon)\widetilde{w}_n R(x)\right]$$
$$+ 2E\left[\partial m_o(\varepsilon)\widetilde{w}_n R(x)'\right] Q^{-1}_{L(n)} E\left[R(x)\widetilde{w}_n \varepsilon u\right] + E[\widetilde{w}^2_n u^2]$$

where $Q_{L(n)} = E\left[R(x)R(x)'\right]$, $Q_{\varepsilon,L(n)} = E\left[\varepsilon^2 R(x)R(x)'\right]$,

$$\widetilde{w}_n \equiv w - E\left[wP(\varepsilon)'\right]\left\{E\left[P(\varepsilon)P(\varepsilon)'\right]\right\}^{-1} P(\varepsilon) \text{ and}$$
$$I^{(1)}_{\theta\theta,n} = \left(E\left[w^2\right] - E\left[wP(\varepsilon)'\right]\left(E\left[P(\varepsilon)P(\varepsilon)'\right]\right)^{-1} E\left[P(\varepsilon)w\right]\right)^{-1}.$$

Because $E\left[u|\, w, x, \varepsilon\right] = 0$, the second term is zero, and we obtain

$$\frac{\left\|v^*_{\theta,n}\right\|^2_{sd}}{(I^{(1)}_{\theta\theta,n})^2} = E\left[\partial m_o(\varepsilon)\widetilde{w}_n R(x)'\right] Q^{-1}_{L(n)} Q_{\varepsilon,L(n)} Q^{-1}_{L(n)} E\left[\partial m_o(\varepsilon)\widetilde{w}_n R(x)\right] + E[\widetilde{w}^2_n u^2].$$

Noting that

$$\left[E\left[w^2\right] - E\left[wP(\varepsilon)'\right]\left\{E\left[P(\varepsilon)P(\varepsilon)'\right]\right\}^{-1} E\left[P(\varepsilon)w\right]\right]^{-1} = \left(E\left[\left|w - E\left[w|\,\varepsilon\right]\right|^2\right]\right)^{-1} + o(1),$$

we can see that

$$\left\|v^*_{\theta,n}\right\|^2_{sd} = \Phi^{-2} E\left[\partial m_o(\varepsilon)\widetilde{w} R(x)'\right] Q^{-1}_{L(n)} Q_{\varepsilon,L(n)} Q^{-1}_{L(n)} E\left[\partial m_o(\varepsilon)\widetilde{w} R(x)\right] + \Phi^{-2} E[\widetilde{w}^2 u^2] + o\left(1\right),$$

where $\Phi \equiv E\left[\left(w - E\left[w|\,\varepsilon\right]\right)^2\right]$ and $\widetilde{w} \equiv w - E\left[w|\,\varepsilon\right]$.

When h_o is parametrically specified as $h_o(Z_i) = x_i'\gamma_o$ in Li and Wooldridge's (2002), we have

$$\left\|v^*_{\theta,n}\right\|^2_{sd} = \Phi^{-1}\left\{E\left[\partial m_o(\varepsilon)\widetilde{w}x'\right]\Sigma_{22} E\left[\partial m_o(\varepsilon)\widetilde{w}x\right] + E[\widetilde{w}^2 u^2]\right\}\Phi^{-1} + o(1), \tag{53}$$

where $\Sigma_{22} = \left(E\left[xx'\right]\right)^{-1} E\left[\varepsilon^2 xx'\right]\left(E\left[xx'\right]\right)^{-1}$. Letting $E\left[\partial m_o(\varepsilon)\widetilde{w}x'\right] = E\left[\partial m_o(\varepsilon)\widetilde{w}\left(x - E\left[x|\,\varepsilon\right]\right)'\right] \equiv \Psi'$ and $\Omega_1 \equiv E[\widetilde{w}^2 u^2]$, we can rewrite (53) as

$$\left\|v^*_{\theta,n}\right\|^2_{sd} \to \Phi^{-1}\left(\Psi'\Sigma_{22}\Psi + \Omega_1\right)\Phi^{-1}.$$

In other words, $\sqrt{n}(\widehat{\theta}_n-\theta_o) \to_d N\left[0, \Phi^{-1}\left(\Psi'\Sigma_{22}\Psi+\Omega_1\right)\Phi^{-1}\right]$, which coincides with Li and Wooldridge's (2002) conjecture.[8]

When h_o is nonparametrically specified, we note that $E\left[\partial m_o(\varepsilon)\widetilde{w}R(x)'\right]Q_L^{-1}R(x)$ is a nonparametric regression of $\partial m_o(\varepsilon)\widetilde{w}$ on x and hence, we can deduce that

$$\left\|v^*_{\theta,n}\right\|^2_{sd} \to \Phi^{-1}E\left[E\left[\partial m_o(\varepsilon)\widetilde{w}\right|x\right]^2\varepsilon^2+\widetilde{w}^2u^2\Big]\Phi^{-1}. \tag{54}$$

7 Simulation Study

In this section, we study the finite sample performance of the two-step nonparametric M estimator and the proposed inference method. The simulated data is from the following model

$$y_i = w_{1,i}\theta_o + m_o(h_o(x_i)) + u_i, \tag{55}$$

$$s_i = h_o(x_i) + \varepsilon_i, \tag{56}$$

where $\theta_o = 1$; $h_o(x) = 2\cos(\pi x)$, $m_o(w_2) = \sin(\pi w_2)$ and $w_2 = h_o(x)$. For $i = 1, \ldots, n$, we independently draw $(w_{1,i}, x_{*,i}, u_i, \varepsilon_i)'$ from $N(0, I_4)$ and then calculate

$$x_i = \begin{cases} (w_{1,i}+x_{*,i}-1/2)^2\left[1+(w_{1,i}+x_{*,i})^2\right]^{-1}, & \text{in DGP1} \\ 2^{-1/2}(w_{1,i}+x_{*,i}), & \text{in DGP2} \end{cases}. \tag{57}$$

It is clear that x_i has bounded support in DGP1, but unbounded support in DGP2. The data $\{y_i, s_i, w_{1,i}, x_i\}_{i=1}^n$ are generated using the equations in (55) and (56).

The first-step M estimator is $\widehat{h}_n(\cdot) = R(\cdot)'(R_nR_n')^{-1}R_nS_n$ where $R(\cdot)$, R_n and S_n are defined in Section 5. Let $\widehat{w}_{2,i} = \widehat{h}_n(x_i)$ and $\widehat{P}_n = [\overline{P}(\widehat{w}_1), \ldots, \overline{P}(\widehat{w}_n)]'$, where $\overline{P}(\widehat{w}_i)' = [w_{1,i}, P(\widehat{w}_{2,i})']$ and $P(\cdot) = [p_1(\cdot), \ldots, p_K(\cdot)]'$. Define

$$(\widehat{\theta}_n, \widehat{\beta}_n')' = (\widehat{P}_n'\widehat{P}_n)^{-1}\widehat{P}_n'Y_n \tag{58}$$

where $\widehat{\beta}_n$ is a $K \times 1$ vector which is used to construct the estimator of $m_o(\cdot)$: $\widehat{m}_n(\cdot) = P(\cdot)'\widehat{\beta}_n$. The power series are used in both the first-step and second-step M estimations.

We are interested in the inference of the functional value $\rho(g_o) = \theta_o$, where $g_o = (\theta_o, m_o(\cdot))$. Denote

$$I_{n,11} = E\left[w_1^2\right],\ I_{n,22} = E\left[P(w_2)P(w_2)'\right] \text{ and } I_{n,12} = E\left[w_1P(w_2)'\right] = I_{n,21}'. \tag{59}$$

The Riesz representer of the functional $\rho(\cdot)$ has the following form:

$$v_{g_n}^*(w) = \left[v_{\theta,n}^*, -P\left(w_2\right)' I_{n,22}^{-1} I_{n,21} v_{\theta,n}^*\right] \tag{60}$$

where $v_{\theta,n}^* = I_n^{11}$ and $I_n^{11} = \left[I_{n,11} - I_{n,12} I_{n,22}^{-1} I_{n,21}\right]^{-1}$.[9] The Riesz representer $v_{\Gamma_n}^*$ of the functional $\Gamma(\alpha_o)[v_h, v_{g_n}^*]$ has the closed form expression

$$v_{\Gamma_n}^*(x) = -v_{\theta,n}^* E\left[\partial m_o(w_2)\widetilde{w}_{1,n} R(x)'\right] Q_L^{-1} R(x), \tag{61}$$

where $\widetilde{w}_1 = w_1 - P\left(w_2\right)' I_{n,22}^{-1} I_{n,21}$ and $Q_L = E\left[R(x)R(x)'\right]$.[10] Using (55) and (56), we have

$$\|v_n^*\|_{sd}^2 = v_{\theta,n}^{*2}\left[E[u^2\widetilde{w}_1^2] + E\left[\partial m_o(w_2)\widetilde{w}_1 R(x)'\right] Q_L^{-1} Q_{\varepsilon,L} Q_L^{-1} E\left[R(x)\widetilde{w}_1 \partial m_o(w_2)\right]\right] \tag{62}$$

where $Q_{\varepsilon,L} = E\left[\varepsilon^2 R(x)R(x)'\right]$.

We next describe the estimator of $\|v_n^*\|_{sd}^2$. Let

$$\widehat{I}_{n,11} = n^{-1}\sum_{i=1}^n w_{1,i}^2,\ \widehat{I}_{n,22} = n^{-1}\sum_{i=1}^n P(\widehat{w}_{2,i})P(\widehat{w}_{2,i})' \text{ and } \widehat{I}_{n,21} = n^{-1}\sum_{i=1}^n w_{1,i}P(\widehat{w}_{2,i}) = \widehat{I}_{n,12}'.$$

Then $\widehat{v}_{\theta,n}^* = \widehat{I}_n^{11} = [\widehat{I}_{n,11} - \widehat{I}_{n,12}\widehat{I}_{n,22}^{-1}\widehat{I}_{n,21}]^{-1}$. Let

$$\begin{aligned}
\widetilde{w}_{1,i} &= w_{1,i} - P(\widehat{w}_{2,i})'\widehat{I}_{n,22}^{-1}\widehat{I}_{n,21}, & \widehat{u}_i &= y_i - w_{1,i}\widehat{\theta}_n - \widehat{m}_n(\widehat{w}_{2,i}),\\
\widehat{\varepsilon}_i &= s_i - \widehat{h}_n\left(x_i\right), & Q_{\varepsilon,L,n} &= n^{-1}\textstyle\sum_{i=1}^n \widehat{\varepsilon}_i^2 R(x_i)R(x_i)',\\
\partial\widehat{m}_n\left(\cdot\right) &= \partial P\left(\cdot\right)'\widehat{\beta}_n, & Q_{L,n} &= n^{-1}\textstyle\sum_{i=1}^n R(x_i)R(x_i)'.
\end{aligned}$$

The estimator of $\|v_n^*\|_{sd}^2$ is then defined as

$$\|\widehat{v}_n^*\|_{n,sd}^2 = \frac{\widehat{v}_{\theta,n}^{*2}}{n}\left[\sum_{i=1}^n \widehat{u}_i^2\widetilde{w}_{1,i}^2 + \sum_{i=1}^n \partial m_n(\widehat{w}_{2,i})\widetilde{w}_{1,i}R(x_i)'\frac{Q_{L,n}^{-1}Q_{\varepsilon,L,n}Q_{L,n}^{-1}}{n}\sum_{i=1}^n R(x_i)\widetilde{w}_{1,i}\partial m_n(\widehat{w}_{2,i})\right]. \tag{63}$$

The $1-q$ ($q \in (0,1)$) confidence interval of θ_o is

$$CI_{\theta,1-q} = \left[\widehat{\theta}_n - n^{-1/2}\left\|\widehat{v}_n^*\right\|_{n,sd} z_{1-q/2}, \widehat{\theta}_n + n^{-1/2}\left\|\widehat{v}_n^*\right\|_{n,sd} z_{1-q/2}\right] \tag{64}$$

where $z_{1-q/2}$ denotes the $(1-q/2)$-th quantile of the standard normal random variable.

We consider sample sizes $n = 100$, 250 and 500 in this simulation study. For each sample size, we

Figure 7.1. The Mean Square Errors of the Two-step Sieve M Estimators of m_o and θ_o (DGP1)

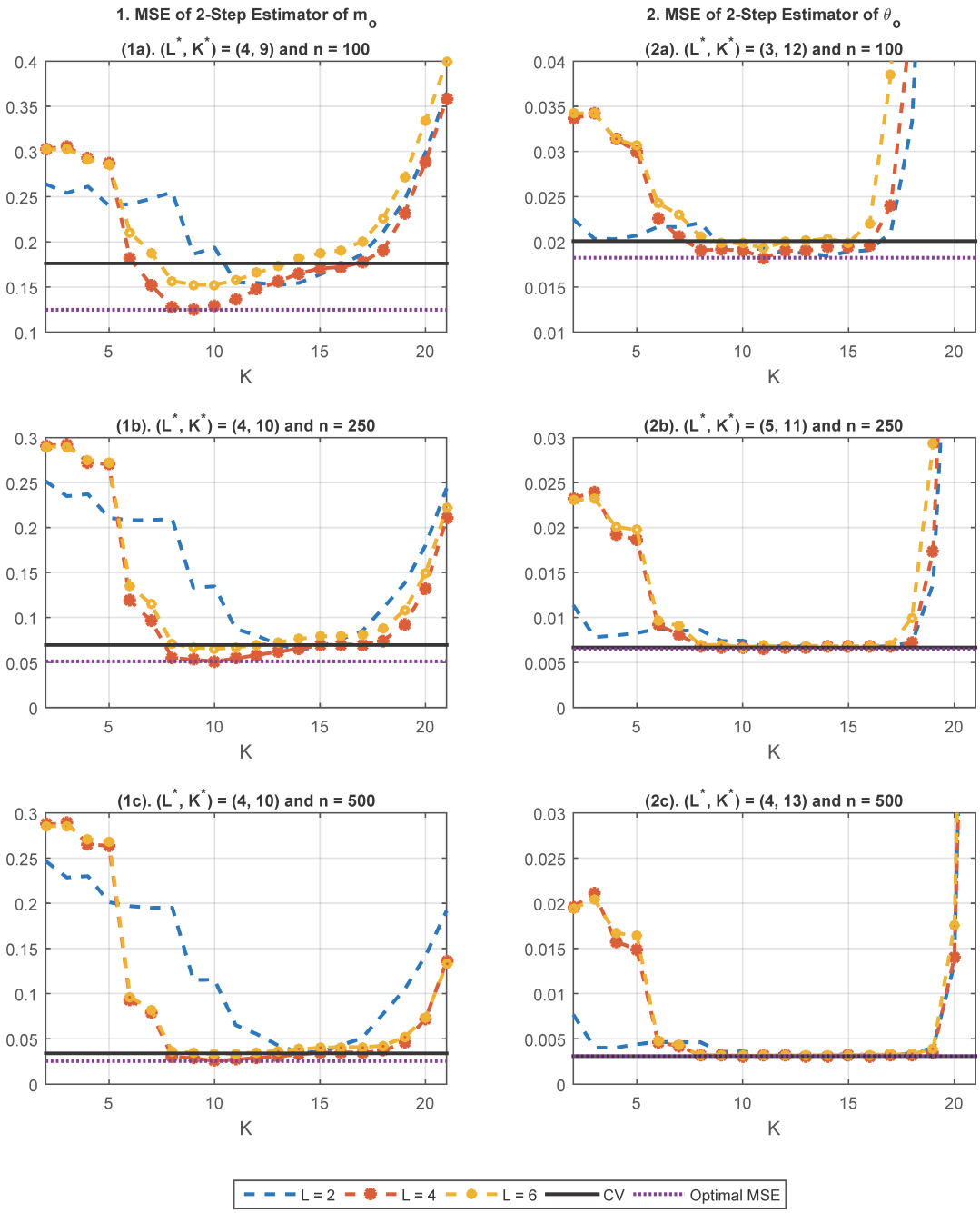

Figure 1: 1. The left panel represents the MSEs of the two-step sieve estimator of m_o for sample sizes n=100, 250 and 500 respectively; 2. the right panel represents the MSEs of the two-step sieve estimator of θ_o for sample sizes n=100, 250 and 500 respectively; 3. L^* and K^* denote the numbers of the series terms which produce sieve estimator of m_o with the smallest finite sample MSE (in the left panel) or sieve estimator of θ_o with the smallest finite sample MSE (in the left panel); 4. the dotted line represents the MSE of the two-step sieve M estimator with $L = L^*$ and $K = K^*$; 5. the solid line represents the MSE of the two-step sieve M estimator with L and K selected by 5-fold cross-validation.

Figure 7.2. The Convergence Probability and the Average Length of the Confidence Interval of θ_o (DGP1)

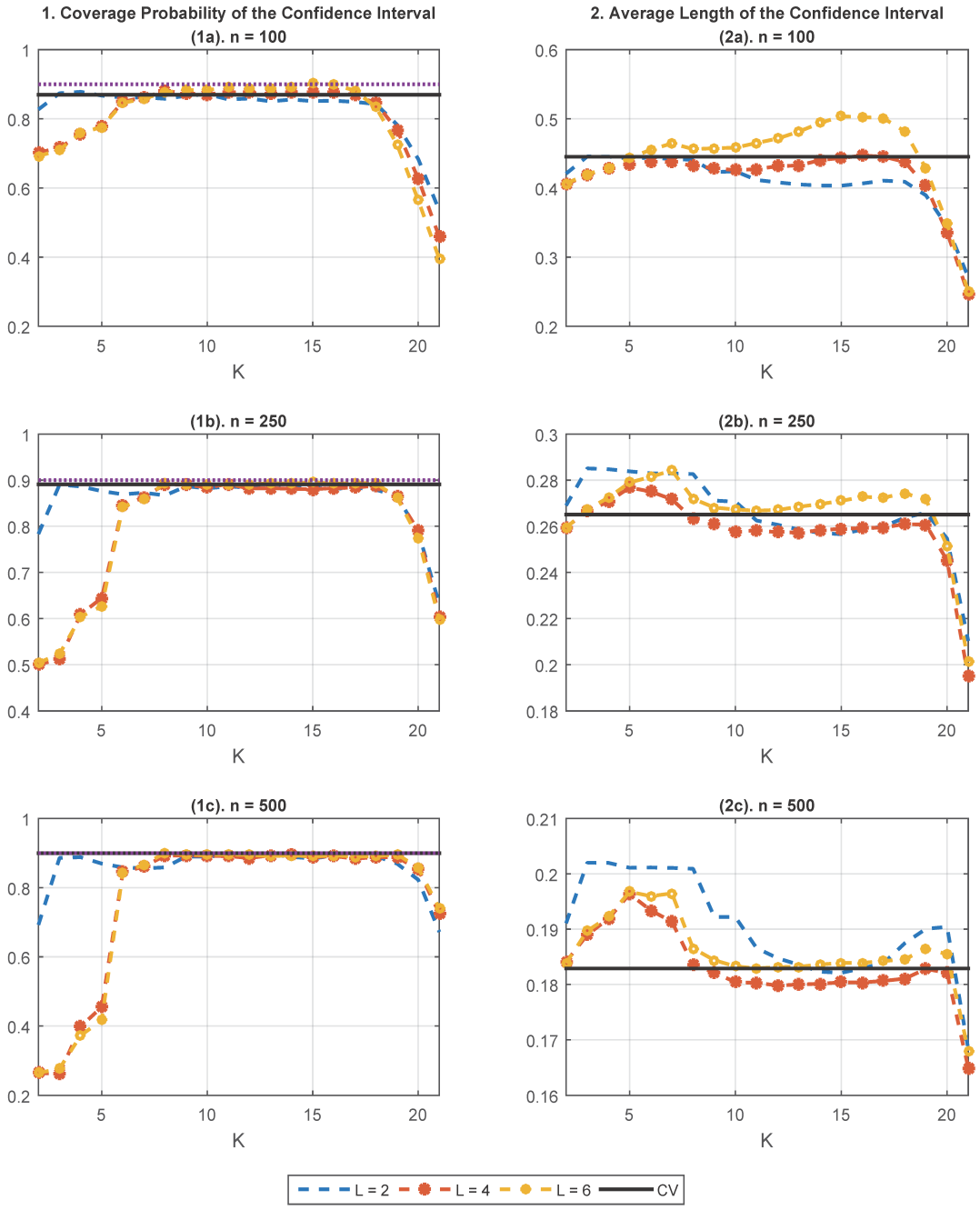

Figure 2: 1. The left panel presents the coverage probability of the confidence interval of θ_o for sample sizes n=100, 250 and 500 respectively; 2. the right panel presents the average length of the confidence interval of θ_o for sample sizes n=100, 250 and 500 respectively; 3. the dotted line in the left panel is the 0.90 line which represents the nominal coverage of the confidence interval; 4. the solid line represents the coverage probability of the confidence interval based on the two-step sieve estimator with K and L selected by 5-fold cross-validation.

Figure 7.3. The Mean Squared Errors of the Two-step Sieve M Estimators of m_o and θ_o (DGP2)

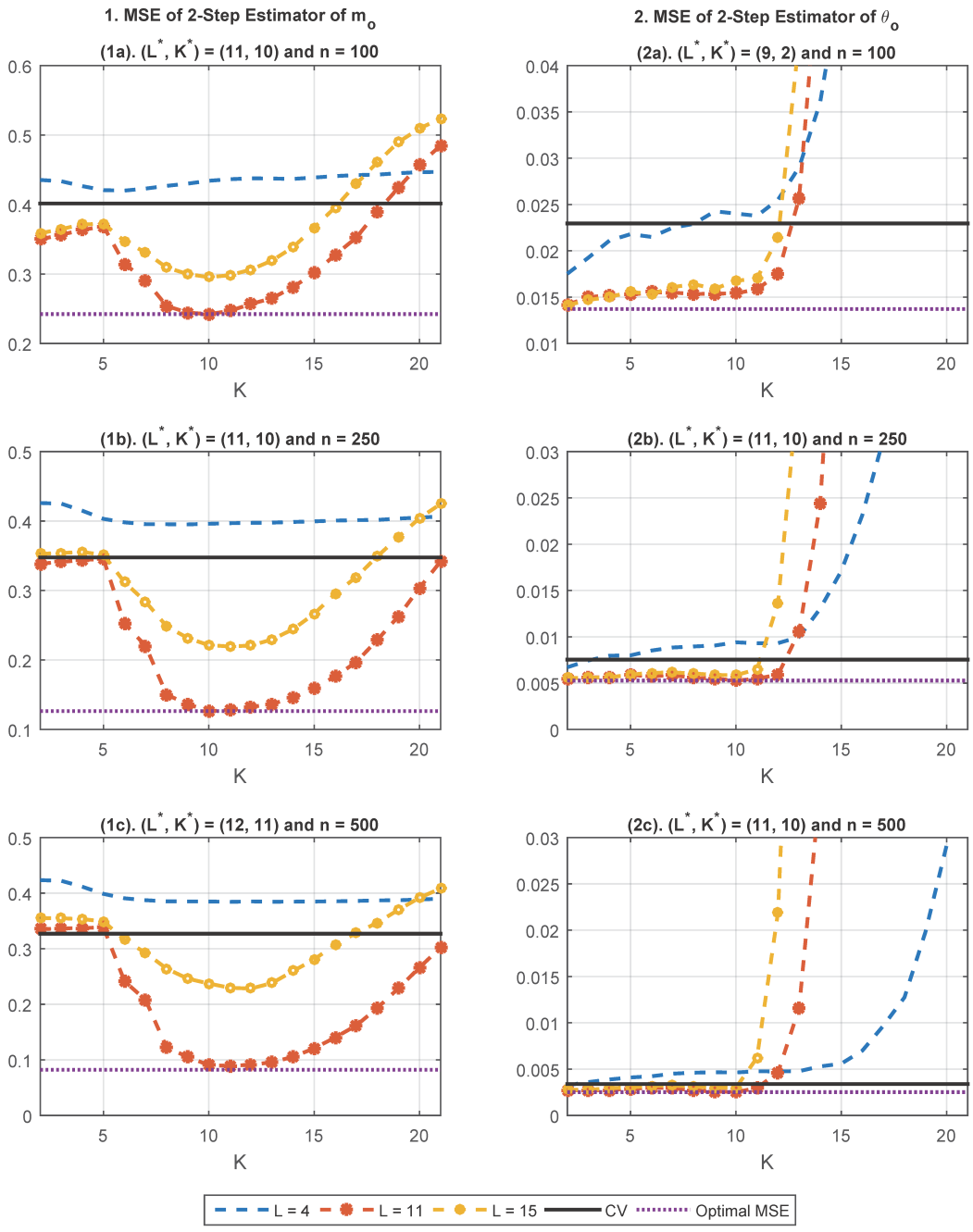

Figure 3: 1. The left panel represents the MSEs of the two-step sieve estimator of m_o for sample sizes n=100, 250 and 500 respectively; 2. the right panel represents the MSEs of the two-step sieve estimator of θ_o for sample sizes n=100, 250 and 500 respectively; 3. L^* and K^* denote the numbers of the series terms which produce sieve estimator of m_o with the smallest finite sample MSE (in the left panel) or sieve estimator of θ_o with the smallest finite sample MSE (in the left panel); 4. the dotted line represents the MSE of the two-step sieve M estimator with $L = L^*$ and $K = K^*$; 5. the solid line represents the MSE of the two-step sieve M estimator with L and K selected by 5-fold cross-validation.

Figure 7.4. The Convergence Probability and the Average Length of the Confidence Interval of θ_o (DGP2)

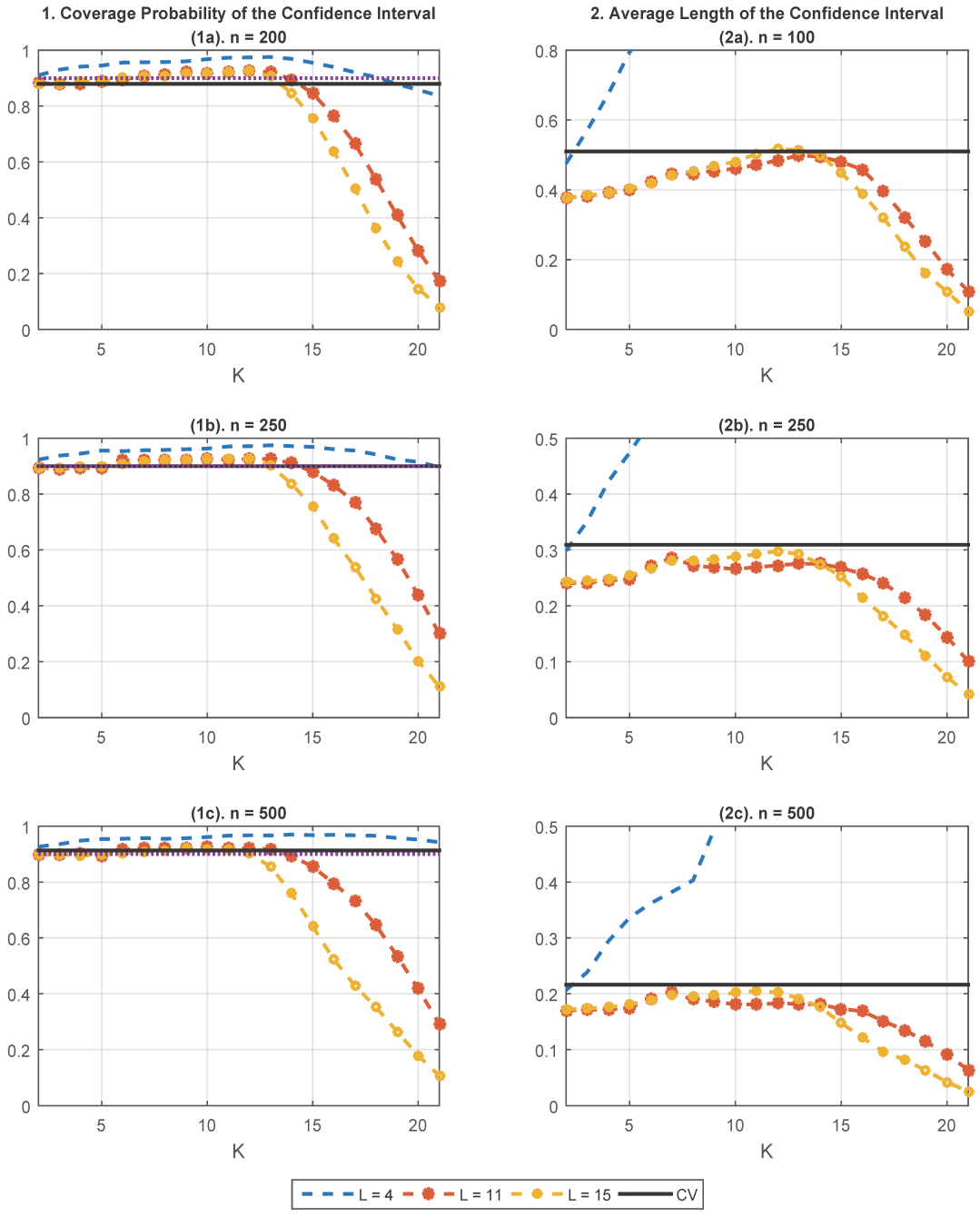

Figure 4: 1. The left panel presents the coverage probability of the confidence interval of θ_o for sample sizes n=100, 250 and 500 respectively; 2. the right panel presents the average length of the confidence interval of θ_o for sample sizes n=100, 250 and 500 respectively; 3. the dotted line in the left panel is the 0.90 line which represents the nominal coverage of the confidence interval; 4. the solid line represents the coverage probability of the confidence interval based on the two-step sieve estimator with K and L selected by 5-fold cross-validation.

generate 10000 simulated samples to evaluate the performances of the two-step sieve estimator and the proposed inference procedure. For each simulated sample, we calculate the sieve estimator of (θ_o, m_o), and the 0.90 confidence interval of θ_o for each combination of (L, K) where $L = 2, \ldots, 16$ and $K = 2, \ldots, 21$. The simulation results are presented in Figures 7.1, 7.2, 7.3 and 7.4.

In Figure 7.1, we see that when the numbers of the series functions (L and K) are too small, the mean square errors (MSEs) of the estimators of m_o and θ_o are big.[11] That is due to the large approximation errors in the first-step and second-step nonparametric estimation when L and K are small. On the other hand, the MSEs of the sieve estimators are also big when L and K are too large. That is because in this case the stochastic estimation error in the two-step nonparametric estimation is big. The optimal L and K which minimize the finite sample MSE of the sieve estimator will balance the trade-off between the approximation error and the stochastic estimation error in the two-step nonparametric estimation. It is interesting to see that the optimal L and K which minimize the MSE of the estimator of m_o are smaller than their counterparts which minimize the MSE of the estimator of θ_o (the only exception is the optimal L when $n = 100$). With the growth of the sample size, the optimal L and K tend to increase. The black solid line in Figure 7.1 represents the MSEs of the sieve estimators based on L and K selected by 5-fold cross-validation.[12] In the right panel of Figure 7.1, we see that as the sample size increases, the MSE of the cross-validated estimator of θ_o approaches the optimal MSE (the dotted line) quickly. Moreover, the solid line and the dotted line are almost identical when the sample size is 250. On the other hand, the nonparametric J-fold cross-validated estimator usually enjoys the optimal convergence rate but it may not necessarily be efficient.[13] In the left panel of Figure 7.1, it is interesting to see that the MSE of the cross-validated sieve estimator of m_o approaches the optimal value with the growth of the sample size.

The properties of the confidence interval of θ_o in DGP1 are displayed in Figure 7.2. When L and K are too small, the confidence interval suffers large size distortion. The coverage probability in this case becomes even worse with the growth of the sample size. This is because the confidence interval is miscentered due to the large approximation error which can not be reduced by increasing the sample size. Meanwhile, the miscentered confidence interval becomes narrower with larger sample sizes which increases the size distortion for the confidence interval with small L and K. On the other hand, the coverage probability of the confidence interval is far below 0.90 when L and K are too large. Two factors may contribute to this phenomenon. First, the bias from the second order stochastic estimation error becomes nontrivial when n is small, and L and K are large. This makes the proposed confidence interval miscentered in finite samples. Second, the variance from the second order stochastic estimation error also becomes nontrivial which means the confidence interval in (64) is too narrow when L and K are too large.

The effects of these two factors become small when the sample size is large, as we can see from Figure 7.2. The coverage probability of the confidence interval based on the cross-validated sieve estimator is slightly below the nominal level when n is small (i.e., $n = 100$), and it approaches the nominal level quickly with a growing sample size. The right panel of Figure 7.2 shows good properties of the confidence interval based on the cross-validated sieve estimator. From the left panel of Figure 7.1, we see that when L is small or large (i.e., $L = 2$ or 6), one can set K to be between 9 and 18 to achieve that that the coverage probabilities of the resulting confidence intervals are close to the nominal level. However, these confidence intervals tend to be much longer when compared with the confidence interval based on the cross-validated sieve estimator.

The simulation results under DGP2 are presented in Figure 7.3 and Figure 7.4. The properties of the two-step sieve M estimator and the proposed confidence interval are similar to what we found in DGP1. We list some important differences. First, when the unknown function estimated in the first-step has unbounded support, the optimal L which produces a two-step M estimator with the smallest MSE is much larger. Second, the ratio between the MSE of the cross-validated estimator of m_o and the optimal MSE does not seem to converge to 1 in all the sample sizes we considered. However, the MSE of the cross-validated estimator of θ_o does approach the optimal value quickly as the sample size increases. Third, when L is small (e.g., $L = 4$), the proposed confidence interval over-covers the unknown parameter θ_o and its length diverges with increasing K. Fourth, the coverage probability of the confidence interval based on the cross-validated sieve estimator is almost identical to the nominal level even when the sample size is small (e.g., $n = 100$).

8 Conclusion

In this paper, we examined statistical properties of two-step sieve M estimation, where both the first and second step models may involve infinite dimensional parameters, and the nonparametric components in the second step model may use the first-step unknown functions as arguments. Our theoretical results expand the applicability of semi/nonparametric methods in estimation and inference of complicated structural models. On a practical level, we established a generalization of the result of Ackerberg, Chen, and Hahn (2012). We show that, even in a more complex semiparametric model with generated regressors, one can implement a two-step linear sieve procedure as if it were a parametric procedure in the spirit of Newey (1984) and Murphy and Topel (1985).

Notes

[1] See Chesher (2003), Altonji and Matzkin (2005), Blundell and Powell (2004, 2007), Florens, Heckman, Meghir and Vytlacil (2008), and Imbens and Newey (2009), among others.

[2] Outside of the sieve estimation framework, there are similar attempts for such a generalization. See, e.g., Ahn (1995), Lewbel and Linton (2002), Mammen, Rothe and Schienle (2012, 2015) and the references therein. Unlike these papers, we provide equivalence and robustness results, which are expected to be appealing to practitioners.

[3] The derivation of these results in the general nonparametric nonlinear models, which are based on standard arguments, can be found in the Supplemental Appendix of this paper.

[4] In some examples (see, e.g., the examples in Subsection 6.1 and Section 5), the two "score functions" $\Delta_\varphi(Z_{1,i}, h_o)[v_h]$ and $\Delta_\psi(Z_{2,i}, \alpha_o)[v_g]$ are orthogonal with each other for any v_h and v_g. Hence when the data is i.i.d., the sieve variance can be more concisely written as $\|v_n^*\|_{sd}^2 = E\left[\left|\Delta_\varphi(Z_{1,i}, h_o)[v_{h_n}^* + v_{\Gamma_n}^*]\right|^2\right] + E\left[\left|\Delta_\psi(Z_{2,i}, \alpha_o)[v_{g_n}^*]\right|^2\right]$.

[5] By combining the results of this paper with those of Chen, Liao and Sun (2014), one can show that analogous results carry over to weakly dependent time series data.

[6] Explicit characterization of $\widehat{V}_\theta$ and $\widehat{V}_{\beta_m}$ is straightforward, and is omitted to keep the notation simple.

[7] This is derived by applying Lemma C.6 in Appendix C.

[8] Conjecture 2.1 in Li and Wooldridge (2002) is on the joint limiting distribution of $\widehat{\beta}_n$ and $\widehat{\theta}_n$, where $\widehat{\beta}_n$ is the first step LS estimate. Although the weak convergence is only on $\widehat{\theta}_n$, we can define the functional $\rho(\alpha) = \lambda'(\beta', \theta')'$ and derive the asymptotic variance of $(\widehat{\beta}_n, \widehat{\theta}_n)$.

[9] The form of $v_{g_n}^*(\cdot)$ in (60) is directly from (110) in Appendix C.

[10] The form of $v_{\Gamma_n}^*(\cdot)$ in (61) is directly from (106) and (108) in Appendix C.

[11] Let $\widehat{m}_{n,L,K}(\cdot)$ denote the second-step M-estimator of $m_o(\cdot)$ based on the simulated sample with sample size n, and the numbers of series functions L and K in the first-step and second step nonparametric estimations respectively. The empirical MSE of $\widehat{m}_{n,L,K}(\cdot)$ is $n^{-1}\sum_{i=1}^n |\widehat{m}_{n,L,K}(w_{2,i}) - m_o(w_{2,i})|^2$. The MSE of $\widehat{m}_{n,L,K}(\cdot)$ refers to the average of the empirical MSEs of the 10000 simulated samples.

[12] The J-fold cross-validation is described in Appendix A.

[13] The cross-validated nonparametric estimator is called efficient if the ratio of its MSE and the optimal MSE converges to 1.

References

[1] Ackerberg, D., X. Chen, and J. Hahn (2012): "A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators," *Review of Economics and Statistics,* 94, 481-498.

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

[3] Blundell, R., and J.L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies* 71, 655-679.

[4] Blundell, R., and J.L. Powell (2007): "Censored Regression Quantiles with Endogenous Regressors," *Journal of Econometrics* 141, 65-83.

[5] Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," In: James J. Heckman and Edward E. Leamer, Editor(s), *Handbook of Econometrics*, 6B, Pages 5549-5632.

[6] Chen, X. and Z. Liao (2014): "Sieve M Inference of Irregular Parameters," *Journal of Econometrics*, 182(1), 70-86

[7] Chen, X., Z. Liao and Y. Sun (2014): "Sieve Inference on Possibly Misspecified Semi-nonparametric Time Series Models," *Journal of Econometrics*, 178(3), 639-658.

[8] Chen, X., O. Linton and I. van Keilegom (2003): "Estimation of Semiparametric Models when the Criterion Function is not Smooth," *Econometrica*, 71, 1591-1608.

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

[10] Chesher, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71(5), 1405–14

[11] Das, M., W. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies,* 70, 33-58.

[12] Florens, J., J. Heckman, C. Meghir and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models with Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica,* 76, 1191-1206.

[13] Hahn, J., and G. Ridder (2013): "The Asymptotic Variance of Semi-parametric Estimators with Generated Regressors," *Econometrica*, 81, 315-340.

[14] Ichimura, H., and S. Lee (2010): "Characterization of the Asymptotic Distribution of Semiparametric M estimators," *Journal of Econometrics,* 159, 252-266.

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

[16] Lee, S. (2007): "Endogeneity in quantile regression models: A Control function approach," *Journal of Econometrics*, 141, 1131-1158.

[17] Lewbel, A., and Linton, O. (2002): "Nonparametric Censored and Truncated Regression," *Econometrica*, 70, 765-779

[18] Li, Q. and M. Wooldridge (2002): "Semiparametric Estimation of Partially Linear Models for Dependent Data with Generated Regressors," *Econometric Theory,* 18, 625-645.

[19] Mammen, E., C. Rothe and M. Schienle (2012): "Nonparametric Regression with Nonparametrically Generated Covariates," *Annals of Statistics*, 40, 1132-1170.

[20] Mammen, E., C. Rothe and M. Schienle (2015): "Semiparametric Estimation with Generated Covariates," *Econometric Theory,* forthcoming.

[21] Murphy, K. and R. Topel (1985): "Estimation and Inference in Two-Step Econometric Models," *Journal of Business and Economic Statistics,* 3, 370-379.

[22] Newey, W. (1984): "A Method of Moments Interpretation of Sequential Estimators," *Economics Letters,* 14, 201-206.

[23] Newey, W. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica,* 62, 1349-1382.

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

[25] Newey, W. (2009): "Two-Step Series Estimation of Sample Selection Models," *Econometrics Journal,* 12, S217–S229.

[26] Olley, G. and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica,* 64, 1263-1297.

[27] Shen, X. (1997): "On Methods of Sieves and Penalization", *Annals of Statistics*, 25, 2555-2591.

[28] Wooldridge, J.M. (2002): *Econometric Analysis of Cross Section and Panel Data*, Cambridge: MIT Press.

Appendix

A A Cross-Validation Method for Selecting the Tuning Parameters

In this appendix, we describe the cross-validation procedure used for selecting the numbers of series terms in the two-step sieve M estimation. We first introduce some notation. For $j = 1, \ldots, J$ (where J is a finite positive integer), let I_j be a subset of $I = \{1, 2, \ldots, n\}$ such that $I_1 \cup \cdots \cup I_J = I$ and $I_{j_1} \cap I_{j_2} = \varnothing$ for any $j_1 \neq j_2$. Let n_j be the number of elements in I_j for $j = 1, \ldots, J$. Let $\mathcal{L}_n$ and $\mathcal{K}_n$ be the user specified sets of candidate values of L and K for the two-step M estimation. For example, in the simulation study of this paper, we set $\mathcal{L}_n = \{2, \ldots, 15\}$ and $\mathcal{K}_n = \{2, \ldots, 21\}$.

For each $j = 1, \ldots, J$ and each $L \in \mathcal{L}_n$, we calculate the first-step M estimator $\widehat{h}_{j,L}$ as

$$n_j^{-1} \sum_{i \in I_j} \varphi(Z_{1,i}, \widehat{h}_{j,L}) \geq \sup_{h \in \mathcal{H}_L} n_j^{-1} \sum_{i \in I_j} \varphi\left(Z_{1,i}, h\right) - O_p(\varepsilon_{1,n_j}^2), \tag{65}$$

where $\mathcal{H}_L$ is the sieve space generated by L basis functions. Given $\widehat{h}_{j,L}$, we calculate the second-step M estimator $\widehat{g}_{j,K}$ as

$$n_j^{-1} \sum_{i \in I_j} \psi(Z_{2,i}, \widehat{g}_{j,K}, \widehat{h}_{j,L}) \geq \sup_{g \in \mathcal{G}_K} \frac{1}{n} \sum_{i=1}^{n} \psi(Z_{2,i}, g, \widehat{h}_{j,L}) - O_p(\varepsilon_{2,n_j}^2), \tag{66}$$

for each $j = 1, \ldots, J$, each $L \in \mathcal{L}_n$ and each $K \in \mathcal{K}_n$, where $\mathcal{G}_K$ is the sieve space generated by K basis functions. The numbers of series terms L^* and K^* are then selected by the following minimization problem:

$$(L^*, K^*) = \underset{L \in \mathcal{L}_n,\, K \in \mathcal{K}_n}{\arg\min} \; J^{-1} \sum_{j=1}^{J} \left[(n - n_j)^{-1} \sum_{i \in I_j^c} \psi(Z_{2,i}, \widehat{g}_{j,K}, \widehat{h}_{j,L}) \right] \tag{67}$$

where I_j^c denotes the complement of I_j for $j = 1, \ldots, J$. The cross-validated two-step M estimator is then calculated by plugging (L^*, K^*) into the estimation problems (1) and (2) in Section 2.

B Proof of Results in Section 3

Proof of Theorem 3.1. Recall that $g^* = g \pm \kappa_n u_{g_n}^*$, where $\kappa_n = o(n^{-\frac{1}{2}})$ and $u_{g_n}^* \equiv v_{g_n}^* \left\| v_n^* \right\|_{sd}^{-1}$. By Assumption 3.3.(iii), $\kappa_n \delta_{2,n}^{*-1} = o\left(1\right)$ and the triangle inequality, we have

$$||\widehat{g}_n^* - g_o||_\psi \leq ||\widehat{g}_n - g_o||_\psi + \kappa_n ||u_{g_n}^*||_\psi = O_p(\delta_{2,n}^*) \tag{68}$$

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

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

$$\begin{aligned}
-O_p\left(\varepsilon_{2,n}^2\right) &\leq \frac{1}{n}\sum_{i=1}^{n}\psi\left(Z_{2,i},\widehat{g}_n,\widehat{h}_n\right)-\frac{1}{n}\sum_{i=1}^{n}\psi\left(Z_{2,i},\widehat{g}_n^*,\widehat{h}_n\right)\\
&=\mu_n\left\{\psi(Z_2,\widehat{g}_n,\widehat{h}_n)-\psi(Z_2,\widehat{g}_n^*,\widehat{h}_n)+\Delta_\psi(Z_2,\widehat{g}_n,\widehat{h}_n)[\pm\kappa_n u_{g_n}^*]\right\}\\
&+\mu_n\left\{\Delta_\psi(Z_2,g_o,h_o)[\pm\kappa_n u_{g_n}^*]-\Delta_\psi(Z_2,\widehat{g}_n,\widehat{h}_n)[\pm\kappa_n u_{g_n}^*]\right\}\\
&-\mu_n\left\{\Delta_\psi(Z_2,g_o,h_o)[\pm\kappa_n u_{g_n}^*]\right\}+\left[K_\psi(\widehat{g}_n,\widehat{h}_n)-K_\psi(\widehat{g}_n^*,\widehat{h}_n)\right]. \qquad (69)
\end{aligned}$$

By Assumption 3.2.(i), we have

$$\mu_n\left\{\psi(Z_2,\widehat{g}_n,\widehat{h}_n)-\psi(Z_2,\widehat{g}_n^*,\widehat{h}_n)+\Delta_\psi(Z_2,\widehat{g}_n,\widehat{h}_n)[\pm\kappa_n u_{g_n}^*]\right\}=O_p(\kappa_n^2) \qquad (70)$$

$$\text{and }\mu_n\left\{\Delta_\psi(Z_2,g_o,h_o)[u_{g_n}^*]-\Delta_\psi(Z_2,\widehat{g}_n,\widehat{h}_n)[u_{g_n}^*]\right\}=O_p(\kappa_n). \qquad (71)$$

Note that $\Gamma(\alpha_o)[\cdot,\cdot]$ is a bilinear functional. Using (68), Assumption 3.2.(ii) and Assumption 3.3.(iii), we deduce that

$$\begin{aligned}
&K_\psi(\widehat{g}_n,\widehat{h}_n)-K_\psi(\widehat{g}_n^*,\widehat{h}_n)\\
&=\mp\kappa_n\Gamma(\alpha_o)[\widehat{h}_n-h_o,u_{g_n}^*]+\frac{||\widehat{g}_n^*-g_o||_\psi^2-||\widehat{g}_n-g_o||_\psi^2}{2}+o_p\left(\kappa_n^2\right)\\
&=\langle\mp\kappa_n u_{\Gamma_n}^*,\widehat{h}_n-h_o\rangle_\varphi+\frac{\kappa_n^2||u_{g_n}^*||_\psi^2}{2}+\langle\pm\kappa_n u_{g_n}^*,\widehat{g}_n-g_o\rangle_\psi+o_p\left(\kappa_n^2\right)\\
&=\langle\mp\kappa_n u_{\Gamma_n}^*,\widehat{h}_n-h_o\rangle_\varphi+\langle\pm\kappa_n u_{g_n}^*,\widehat{g}_n-g_o\rangle_\psi+O_p\left(\kappa_n^2\right). \qquad (72)
\end{aligned}$$

From $\varepsilon_{2,n}=O\left(\kappa_n\right)$, (69), (70), (71) and (72), we get

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

Dividing by κ_n, we obtain

$$\left|\left\langle\widehat{g}_n-g_o,u_{g_n}^*\right\rangle_\psi-\langle\widehat{h}_n-h_o,u_{\Gamma_n}^*\rangle_\varphi-\mu_n\left\{\Delta_\psi(Z_2,g_o,h_o)[u_{g_n}^*]\right\}\right|=O_p(\kappa_n). \qquad (73)$$

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 $\langle g_{o,n}-g_o,u_{g_n}^*\rangle_\psi=0$ and

$$\langle\widehat{g}_n-g_o,u_{g_n}^*\rangle_\psi=\langle\widehat{g}_n-g_{o,n},u_{g_n}^*\rangle_\psi. \qquad (74)$$

From (73), (74) and $\kappa_n=o(n^{-\frac{1}{2}})$, we get

$$\left|\langle\widehat{g}_n-g_{o,n},u_{g_n}^*\rangle_\psi-\langle\widehat{h}_n-h_o,u_{\Gamma_n}^*\rangle_\varphi-\mu_n\left\{\Delta_\psi(Z_2,g_o,h_o)[u_{g_n}^*]\right\}\right|=o_p(n^{-\frac{1}{2}}). \qquad (75)$$

By Assumption 3.1.(i)-(iii) and the Riesz representation theorem,

$$\begin{aligned}
&\sqrt{n}\frac{\rho(\widehat{h}_n,\widehat{g}_n)-\rho(h_{o,n},g_{o,n})}{\|v_n^*\|_{sd}}=\sqrt{n}\frac{\frac{\partial\rho(\alpha_o)}{\partial h}[\widehat{h}_n-h_{o,n}]+\frac{\partial\rho(\alpha_o)}{\partial h}[\widehat{g}_n-g_{o,n}]}{\|v_n^*\|_{sd}}\\
&+\sqrt{n}\frac{\rho(\widehat{h}_n,\widehat{g}_n)-\rho(h_o,g_o)-\frac{\partial\rho(\alpha_o)}{\partial h}[\widehat{h}_n-h_o]-\frac{\partial\rho(\alpha_o)}{\partial g}[\widehat{g}_n-g_o]}{\|v_n^*\|_{sd}}\\
&-\sqrt{n}\frac{\rho(h_{o,n},g_{o,n})-\rho(h_o,g_o)-\frac{\partial\rho(\alpha_o)}{\partial h}[h_{o,n}-h_o]-\frac{\partial\rho(\alpha_o)}{\partial g}[g_{o,n}-g_o]}{\|v_n^*\|_{sd}}\\
&=\sqrt{n}\left[\langle\widehat{h}_n-h_{o,n},u_{h_n}^*\rangle_\varphi+\langle\widehat{g}_n-g_{o,n},u_{g_n}^*\rangle_\psi\right]+o_p(1). \qquad (76)
\end{aligned}$$

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 $\langle h_o-h_{o,n},u_{h_n}^*\rangle_\varphi=0$ and

$$\langle\widehat{h}_n-h_{o,n},u_{h_n}^*\rangle_\varphi=\langle\widehat{h}_n-h_o,u_{h_n}^*\rangle_\varphi. \qquad (77)$$

From the results in (76) and (77), we get

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

which, together with (75) and Assumption 3.3.(i), implies that

$$\begin{aligned}
&\frac{\sqrt{n}\left[\rho(\widehat{h}_n,\widehat{g}_n)-\rho(h_{o,n},g_{o,n})\right]}{\|v_n^*\|_{sd}}\\
&=\sqrt{n}\left[\langle\widehat{h}_n-h_o,u_{h_n}^*+u_{\Gamma_n}^*\rangle_\varphi+\mu_n\left\{\Delta_\psi(Z_2,g_o,h_o)[u_{g_n}^*]\right\}\right]+o_p(1)\\
&=n^{-\frac{1}{2}}\sum_{i=1}^n\left\{\Delta_\varphi(Z_{1,i},h_o)[u_{h_n}^*+u_{\Gamma_n}^*]+\Delta_\psi(Z_{2,i},g_o,h_o)[u_{g_n}^*]\right\}+o_p(1). \qquad (78)
\end{aligned}$$

Now the result in (14) follows immediately by (78) and Assumption 3.3.(ii).

From Assumption 3.1, we get

$$\begin{aligned}
\left|\frac{\rho(h_{o,n},g_{o,n})-\rho(h_o,g_o)}{\|v_n^*\|_{sd}}\right|&\le\left|\frac{\rho(h_{o,n},g_{o,n})-\rho(h_o,g_o)-\frac{\partial\rho(\alpha_o)}{\partial h}[h_{o,n}-h_o]-\frac{\partial\rho(\alpha_o)}{\partial g}[g_{o,n}-g_o]}{\|v_n^*\|_{sd}}\right|\\
&+\frac{1}{\|v_n^*\|_{sd}}\left[\left|\frac{\partial\rho(\alpha_o)}{\partial h}[h_{o,n}-h_o]\right|+\left|\frac{\partial\rho(\alpha_o)}{\partial h}[g_{o,n}-g_o]\right|\right]\\
&=o(n^{-1/2}), \qquad (79)
\end{aligned}$$

which combined with (14) immediately implies (15). ■

C Proof of Results in Section 4 and Section 6

Lemma C.1 *Under Assumption 4.1, the empirical Riesz representer $\widehat{v}^*_{g_n}$ satisfies:*

$$\left\|v^*_{g_n}\right\|^{-1}_{\psi}\left\|\widehat{v}^*_{g_n}-v^*_{g_n}\right\|_{\psi}=o_p(1). \tag{80}$$

Proof. The proof follows the same arguments in the proof of Lemma 5.1 in Chen, Liao and Sun (2014) and hence is omitted. ■

Lemma C.2 *Under Assumption 4.2, the empirical Riesz representers $\widehat{v}^*_{h_n}$ and $\widehat{v}^*_{\Gamma_n}$ satisfy:*

$$\left\|v^*_{h_n}\right\|^{-1}_{\varphi}\left\|\widehat{v}^*_{h_n}-v^*_{h_n}\right\|_{\varphi}=o_p(1) \text{ and } \left\|v^*_{\Gamma_n}\right\|^{-1}_{\varphi}\left\|\widehat{v}^*_{\Gamma_n}-v^*_{\Gamma_n}\right\|_{\varphi}=o_p(1). \tag{81}$$

Proof. The proof follows the same arguments in the proof of Lemma 5.1 in Chen, Liao and Sun (2014) and hence is omitted. ■

Lemma C.3 *Under Assumption 4.3, we have:*
(i) $\sup_{h\in\mathcal{N}_{h,n},v_h\in\mathcal{W}_{1,n}}\left|n^{-1}\sum_{i=1}^n\Delta^2_\varphi(Z_{1,i},h)[v_h]-E\left[\Delta^2_\varphi(Z_1,h_o)[v_h]\right]\right|=o_p(1)$;
(ii) $\sup_{\alpha\in\mathcal{N}_n,v_g\in\mathcal{W}_{2,n}}\left|n^{-1}\sum_{i=1}^n\Delta^2_\psi(Z_{2,i},\alpha)[v_g]-E\left[\Delta^2_\psi(Z_2,\alpha_o)[v_g]\right]\right|=o_p(1)$;
(iii) $\sup_{\alpha\in\mathcal{N}_n,v_h\in\mathcal{W}_{1,n},v_g\in\mathcal{W}_{2,n}}\left|n^{-1}\sum_{i=1}^n\Delta_\varphi(Z_{1,i},h)[v_h]\Delta_\psi(Z_{2,i},\alpha)[v_g]-E\left[\Delta_\varphi(Z_1,h_o)[v_h]\Delta_\psi(Z_2,\alpha_o)[v_g]\right]\right|=o_p(1)$.

Proof of Lemma C.3. (i) By the triangle inequality, Hölder inequality, Assumption 4.3.(i) and (iv), we have

$$\begin{aligned}
&\sup_{h\in\mathcal{N}_{h,n},v_h\in\mathcal{W}_{1,n}}\left|\frac{1}{n}\sum_{i=1}^n\Delta^2_\varphi(Z_{1,i},h)[v_h]-E\left[\Delta^2_\varphi(Z_1,h_o)[v_h]\right]\right|\\
&\leq\sup_{h\in\mathcal{N}_{h,n},v_h\in\mathcal{W}_{1,n}}\left|\mu_n\left\{\Delta^2_\varphi(Z_1,h)[v_h]\right\}\right|+\sup_{h\in\mathcal{N}_{h,n},v_h\in\mathcal{W}_{1,n}}\left|E\left[\Delta^2_\varphi(Z_1,h)[v_h]-\Delta^2_\varphi(Z_1,h_o)[v_h]\right]\right|\\
&\leq o_p(1)+\sup_{h\in\mathcal{N}_{h,n},v_h\in\mathcal{W}_{1,n}}\left\|\Delta_\varphi(Z_1,h)[v_h]-\Delta_\varphi(Z_1,h_o)[v_h]\right\|_2^2\\
&+2\sup_{h\in\mathcal{N}_{h,n},v_h\in\mathcal{W}_{1,n}}\left\|\Delta_\varphi(Z_1,h)[v_h]-\Delta_\varphi(Z_1,h_o)[v_h]\right\|_2\left\|\Delta_\varphi(Z_1,h_o)[v_h]\right\|_2\\
&=o_p(1)
\end{aligned}$$

which shows the first result.

(ii) The second result can be proved using the same arguments in the proof of the first result, but replacing Assumption 4.3.(i) with Assumption 4.3.(ii).

(iii) By the triangle inequality, we have

$$\begin{aligned}
&|E\left[\Delta_\varphi(Z_1,h)[v_h]\Delta_\psi(Z_2,\alpha)[v_g]-\Delta_\varphi(Z_1,h_o)[v_h]\Delta_\psi(Z_2,\alpha_o)[v_g]\right]| \\
&\leq |E\left[\left(\Delta_\varphi(Z_1,h)[v_h]-\Delta_\varphi(Z_1,h_o)[v_h]\right)\left(\Delta_\psi(Z_2,\alpha)[v_g]-\Delta_\psi(Z_2,\alpha_o)[v_g]\right)\right]| \\
&+|E\left[\left(\Delta_\varphi(Z_1,h)[v_h]-\Delta_\varphi(Z_1,h_o)[v_h]\right)\Delta_\psi(Z_2,\alpha_o)[v_g]\right]| \\
&+|E\left[\Delta_\varphi(Z_1,h_o)[v_h]\left(\Delta_\psi(Z_2,\alpha)[v_g]-\Delta_\psi(Z_2,\alpha_o)[v_g]\right)\right]| \qquad (82)
\end{aligned}$$

for any $\alpha\in\mathcal{N}_n$, $v_h\in\mathcal{W}_{1,n}$ and $v_g\in\mathcal{W}_{2,n}$. Using Hölder inequality, Assumption 4.3.(i) and (ii), we have

$$\begin{aligned}
&\sup_{\alpha\in\mathcal{N}_n,v_h\in\mathcal{W}_{1,n},v_g\in\mathcal{W}_{2,n}}|E\left[\left(\Delta_\varphi(Z_1,h)[v_h]-\Delta_\varphi(Z_1,h_o)[v_h]\right)\left(\Delta_\psi(Z_2,\alpha)[v_g]-\Delta_\psi(Z_2,\alpha_o)[v_g]\right)\right]| \\
&\leq \sup_{\alpha\in\mathcal{N}_n,v_h\in\mathcal{W}_{1,n},v_g\in\mathcal{W}_{2,n}}\left\|\Delta_\varphi(Z_1,h)[v_h]-\Delta_\varphi(Z_1,h_o)[v_h]\right\|_2\left\|\Delta_\psi(Z_2,\alpha)[v_g]-\Delta_\psi(Z_2,\alpha_o)[v_g]\right\|_2=o(1). \qquad (83)
\end{aligned}$$

Similarly by Assumption 4.3.(i), (ii) and (iv), we have

$$\sup_{h\in\mathcal{N}_{h,n},v_h\in\mathcal{W}_{1,n},v_g\in\mathcal{W}_{2,n}}\left\|\Delta_\varphi(Z_1,h)[v_h]-\Delta_\varphi(Z_1,h_o)[v_h]\right\|_2\left\|\Delta_\psi(Z_2,\alpha_o)[v_g]\right\|_2=o(1) \qquad (84)$$

and

$$\sup_{\alpha\in\mathcal{N}_n,v_h\in\mathcal{W}_{1,n},v_g\in\mathcal{W}_{2,n}}\left\|\Delta_\psi(Z_2,\alpha)[v_g]-\Delta_\psi(Z_2,\alpha_o)[v_g]\right\|_2\left\|\Delta_\varphi(Z_1,h_o)[v_h]\right\|_2=o(1). \qquad (85)$$

Now the claimed result follows immediately from Assumption 4.3.(iii), (82), (83), (84) and (85). ■

Proof of Theorem 4.1. Let c denote some generic finite positive constant. As the data are *i.i.d.*, by definition, we have

$$\begin{aligned}
\|v_n^*\|_{sd}^2&=\mathrm{Var}\left[n^{-\frac{1}{2}}\sum\nolimits_{i=1}^n\left(\Delta_\varphi(Z_{1,i},h_o)[v_{h_n}^*+v_{\Gamma_n}^*]+\Delta_\psi(Z_{2,i},\alpha_o)[v_{g_n}^*]\right)\right] \\
&=E\left[\left|\Delta_\varphi(Z_{1,i},h_o)[v_{h_n}^*+v_{\Gamma_n}^*]+\Delta_\psi(Z_{2,i},\alpha_o)[v_{g_n}^*]\right|^2\right] \\
&=\left\|\Delta_\varphi(Z_{1,i},h_o)[v_{h_n}^*+v_{\Gamma_n}^*]+\Delta_\psi(Z_{2,i},\alpha_o)[v_{g_n}^*]\right\|_2^2. \qquad (86)
\end{aligned}$$

By the triangle inequality,

$$\frac{\|\widehat{v}_n^*-v_n^*\|_{sd}}{\|v_n^*\|_{sd}}\leq\frac{\left\|\Delta_\varphi(Z_1,h_o)[\widehat{v}_{h_n}^*-v_{h_n}^*]\right\|_2+\left\|\Delta_\varphi(Z_1,h_o)[\widehat{v}_{\Gamma_n}^*-v_{\Gamma_n}^*]\right\|_2+\left\|\Delta_\psi(Z_2,\alpha_o)[\widehat{v}_{g_n}^*-v_{g_n}^*]\right\|_2}{\|v_n^*\|_{sd}}. \qquad (87)$$

Using (86), Assumption 4.3.(iv) and the result in (80), we have

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

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

$$\frac{\left\|\Delta_\psi(Z_2,\alpha_o)[\widehat{v}^*_{g_n}-v^*_{g_n}]\right\|_2}{\|v^*_n\|_{sd}}\leq\frac{\left\|v^*_{g_n}\right\|_\psi}{\|v^*_n\|_{sd}}\frac{\left\|\widehat{v}^*_{g_n}-v^*_{g_n}\right\|_\psi}{\left\|v^*_{g_n}\right\|_\psi}\sup_{v_g\in\mathcal{W}_{g_n}}\|\Delta_\psi(Z_2,\alpha_o)[v_g]\|_2=o_p(1). \tag{88}$$

Similarly, we have

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

which together with (87) and (88) implies that

$$\|v^*_n\|^{-1}_{sd}\left\|\widehat{v}^*_n-v^*_n\right\|_{sd}=o_p(1). \tag{89}$$

Using (89) and the triangle inequality, we get

$$\begin{aligned}o_p(1)=\frac{\|\widehat{v}^*_n-v^*_n\|_{sd}}{\|v^*_n\|_{sd}}&\geq\left|\frac{\|\widehat{v}^*_n\|_{sd}}{\|v^*_n\|_{sd}}-1\right|=\left|\frac{\|\widehat{v}^*_n\|_{sd}}{\|\widehat{v}^*_n\|_{n,sd}}\frac{\|\widehat{v}^*_n\|_{n,sd}}{\|v^*_n\|_{sd}}-1\right|\\&=\left|\|\widehat{v}^*_n\|^{-1}_{n,sd}\|\widehat{v}^*_n\|_{sd}\left(\|v^*_n\|^{-1}_{sd}\|\widehat{v}^*_n\|_{n,sd}-1\right)+\left(\|\widehat{v}^*_n\|^{-1}_{n,sd}\|\widehat{v}^*_n\|_{sd}-1\right)\right|.\end{aligned} \tag{90}$$

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

$$\begin{aligned}\frac{\left\|\widehat{v}^*_{h_n}+\widehat{v}^*_{\Gamma_n}\right\|_\varphi}{\|\widehat{v}^*_n\|_{sd}}&\leq\frac{\left\|v^*_{h_n}+v^*_{\Gamma_n}\right\|_\varphi+\left\|\widehat{v}^*_{h_n}-v^*_{h_n}\right\|_\varphi+\left\|\widehat{v}^*_{\Gamma_n}-v^*_{\Gamma_n}\right\|_\varphi}{\|\widehat{v}^*_n\|_{sd}}\\&\leq\frac{\left\|v^*_{h_n}\right\|_\varphi+\left\|v^*_{\Gamma_n}\right\|_\varphi+\left\|v^*_{h_n}\right\|_\varphi\frac{\left\|\widehat{v}^*_{h_n}-v^*_{h_n}\right\|_\varphi}{\left\|v^*_{h_n}\right\|_\varphi}+\left\|v^*_{\Gamma_n}\right\|_\varphi\frac{\left\|\widehat{v}^*_{\Gamma_n}-v^*_{\Gamma_n}\right\|_\varphi}{\left\|v^*_{\Gamma_n}\right\|_\varphi}}{\|\widehat{v}^*_n\|_{sd}}\\&=\frac{\left\|v^*_{h_n}\right\|_\varphi}{\|\widehat{v}^*_n\|_{sd}}\left(1+\frac{\left\|\widehat{v}^*_{h_n}-v^*_{h_n}\right\|_\varphi}{\left\|v^*_{h_n}\right\|_\varphi}\right)+\frac{\left\|v^*_{\Gamma_n}\right\|_\varphi}{\|\widehat{v}^*_n\|_{sd}}\left(1+\frac{\left\|\widehat{v}^*_{\Gamma_n}-v^*_{\Gamma_n}\right\|_\varphi}{\left\|v^*_{\Gamma_n}\right\|_\varphi}\right)\\&=\frac{\|v^*_n\|_{sd}}{\|\widehat{v}^*_n\|_{sd}}\left[\frac{\left\|v^*_{h_n}\right\|_\varphi}{\|v^*_n\|_{sd}}(1+o_p(1))+\frac{\left\|v^*_{\Gamma_n}\right\|_\varphi}{\|v^*_n\|_{sd}}(1+o_p(1))\right]\\&=\frac{\|v^*_n\|_{sd}}{\|\widehat{v}^*_n\|_{sd}}O_p(1)=\left(\frac{1}{\|\widehat{v}^*_n\|_{sd}/\|v^*_n\|_{sd}}-1\right)O_p(1)+O_p(1)=O_p(1)\end{aligned} \tag{91}$$

where the first two inequalities are by the triangle inequality, the second equality is by (81), the third equality is by (86) and (25), and the last equality is by the first inequality in (90). Similarly, we can show that

$$\left\|\widehat{v}^*_{g_n}\right\|_\psi \|\widehat{v}^*_n\|^{-1}_{sd} = O_p(1). \tag{92}$$

By the triangle inequality, we get

$$\begin{aligned}\left|\frac{\|\widehat{v}^*_n\|^2_{n,sd} - \|\widehat{v}^*_n\|^2_{sd}}{\|\widehat{v}^*_n\|^2_{sd}}\right| = \Bigg|&\frac{\frac{1}{n}\sum_{i=1}^n \left[\Delta_\varphi(Z_{1,i},\widehat{h}_n)[\widehat{v}^*_{h_n} + \widehat{v}^*_{\Gamma_n}] + \Delta_\psi(Z_{2,i},\widehat{\alpha}_n)[\widehat{v}^*_{g_n}]\right]^2}{\|\widehat{v}^*_n\|^2_{sd}} \\ &- \frac{E_Z\left[\left|\Delta_\varphi(Z_1,h_o)[\widehat{v}^*_{h_n} + \widehat{v}^*_{\Gamma_n}] + \Delta_\psi(Z_2,\alpha_o)[\widehat{v}^*_{g_n}]\right|^2\right]}{\|\widehat{v}^*_n\|^2_{sd}}\Bigg| \\ &\le |I_{1,n}| + |I_{2,n}| + 2\,|I_{3,n}|\end{aligned} \tag{93}$$

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),

$$I_{1,n} = \frac{\frac{1}{n}\sum_{i=1}^n \Delta^2_\varphi(Z_{1,i},\widehat{h}_n)[\widehat{v}^*_{h_n} + \widehat{v}^*_{\Gamma_n}] - E_{Z_1}\left[\Delta^2_\varphi(Z_1,h_o)[\widehat{v}^*_{h_n} + \widehat{v}^*_{\Gamma_n}]\right]}{\|\widehat{v}^*_n\|^2_{sd}},$$

$$I_{2,n} = \frac{\frac{1}{n}\sum_{i=1}^n \Delta^2_\psi(Z_{2,i},\widehat{\alpha}_n)[\widehat{v}^*_{g_n}] - E_{Z_2}\left[\Delta^2_\psi(Z_2,\alpha_o)[\widehat{v}^*_{g_n}]\right]}{\|\widehat{v}^*_n\|^2_{sd}}$$

and

$$\begin{aligned}I_{3,n} = &\frac{\frac{1}{n}\sum_{i=1}^n \Delta_\varphi(Z_{1,i},\widehat{h}_n)[\widehat{v}^*_{h_n} + \widehat{v}^*_{\Gamma_n}]\Delta_\psi(Z_{2,i},\widehat{\alpha}_n)[\widehat{v}^*_{g_n}]}{\|\widehat{v}^*_n\|^2_{sd}} \\ &- \frac{E_Z\left[\Delta_\varphi(Z_1,h_o)[\widehat{v}^*_{h_n} + \widehat{v}^*_{\Gamma_n}]\Delta_\psi(Z_2,\alpha_o)[\widehat{v}^*_{g_n}]\right]}{\|\widehat{v}^*_n\|^2_{sd}}.\end{aligned}$$

By Lemma C.3.(i) and (91), we have

$$\begin{aligned}|I_{1,n}| &= \frac{\left\|\widehat{v}^*_{h_n} + \widehat{v}^*_{\Gamma_n}\right\|^2_\varphi}{\|\widehat{v}^*_n\|^2_{sd}} \frac{\left|\frac{1}{n}\sum_{i=1}^n \Delta^2_\varphi(Z_{1,i},\widehat{h}_n)[\widehat{v}^*_{h_n} + \widehat{v}^*_{\Gamma_n}] - E_{Z_1}\left[\Delta^2_\varphi(Z_1,h_o)[\widehat{v}^*_{h_n} + \widehat{v}^*_{\Gamma_n}]\right]\right|}{\left\|\widehat{v}^*_{h_n} + \widehat{v}^*_{\Gamma_n}\right\|^2_\varphi} \\ &\le O_p(1) \sup_{h\in\mathcal{N}_{h,n}, v_h\in\mathcal{W}_{1,n}} \left|\frac{1}{n}\sum_{i=1}^n \Delta^2_\varphi(Z_{1,i},h)[v_h] - E_{Z_1}\left[\Delta^2_\varphi(Z_1,h_o)[v_h]\right]\right| = o_p(1).\end{aligned} \tag{94}$$

Similarly, by Lemma C.3.(ii) and (92), we have

$$|I_{2,n}| = \frac{\left\|\widehat{v}^*_{g_n}\right\|^2_\psi}{\|\widehat{v}^*_n\|^2_{sd}} \frac{\left|\frac{1}{n}\sum_{i=1}^n \Delta^2_\psi(Z_{2,i},\widehat{\alpha}_n)[\widehat{v}^*_{g_n}] - E_{Z_2}\left[\Delta^2_\psi(Z_2,\alpha_o)[\widehat{v}^*_{g_n}]\right]\right|}{\left\|\widehat{v}^*_{g_n}\right\|^2_\psi} = o_p(1). \tag{95}$$

For the last term $I_{3,n}$, note that

$$\begin{aligned} |I_{3,n}| &\leq \frac{\left\|\widehat{v}^*_{h_n} + \widehat{v}^*_{\Gamma_n}\right\|_\varphi \left\|\widehat{v}^*_{g_n}\right\|_\psi}{\|\widehat{v}^*_n\|^2_{sd}} \\ &\quad \times \sup_{\alpha\in\mathcal{N}_n, v_g\in\mathcal{W}_{2,n}, v_h\in\mathcal{W}_{1,n}} \left| \begin{array}{c} n^{-1}\sum_{i=1}^n \Delta_\varphi(Z_{1,i},h)[v_h]\Delta_\psi(Z_{2,i},\alpha)[v_g] \\ -E_Z\left[\Delta_\varphi(Z_1,h)[v_h]\Delta_\psi(Z_2,\alpha)[v_g]\right] \end{array} \right| \\ &= \frac{\left\|\widehat{v}^*_{h_n} + \widehat{v}^*_{\Gamma_n}\right\|_\varphi \left\|\widehat{v}^*_{g_n}\right\|_\psi}{\|\widehat{v}^*_n\|^2_{sd}} o_p(1) = o_p(1), \end{aligned} \tag{96}$$

where the first equality is by Lemma C.3.(iii), the last equality is by (91) and (92).

From the results in (93), (94), (95) and (96), we deduce that

$$\left| \frac{\|\widehat{v}^*_n\|^2_{n,sd} - \|\widehat{v}^*_n\|^2_{sd}}{\|\widehat{v}^*_n\|^2_{sd}} \right| = o_p(1). \tag{97}$$

It is clear that (90) and (97) imply that $\left|\|\widehat{v}^*_n\|_{n,sd} / \|v^*_n\|_{sd} - 1\right| = o_p(1)$, which finishes the proof. ■

We divide the proof of Proposition 6.1 into establishing several lemmas. We start with an explicit characterization of $\|v^*_n\|^2_{sd}$ when the data are *i.i.d.* and when both h and g are real-valued functions for the sake of simplified notation.

Let $\mathcal{H}_n = \left\{h(\cdot) = R(\cdot)'\gamma_L : \gamma_L \in \mathbb{R}^L\right\}$ be a sieve space for $h_o(\cdot)$ in the first step, where $R(\cdot)$ is defined in Section 6. Let $\mathcal{G}_n = \{g(\cdot) = \overline{P}(\cdot)'\beta_g : \beta_g \in \mathbb{R}^{\overline{K}}\}$ be the sieve space for $g_o(\cdot)$ in the second step, where $\overline{P}(\cdot)$ is defined in Section 6. Define

$$R_\varphi(h_o)[v_{h_1}, v_{h_2}] \equiv \langle v_{h_1}, v_{h_2}\rangle_\varphi \text{ and } R_\psi(\alpha_o)[v_{g_1}, v_{g_2}] \equiv \langle v_{g_1}, v_{g_2}\rangle_\psi \tag{98}$$

for any $v_{h_1}, v_{h_2} \in \mathcal{V}_1$ and $v_{g_1}, v_{g_2} \in \mathcal{V}_2$. Finally, we let $\Delta_\varphi(R) \equiv \Delta_\varphi(Z_1, h_o)[R]$ be $L \times 1$ and $\Delta_\psi(\overline{P}) \equiv \Delta_\psi(Z_2, \alpha_o)[\overline{P}]$ be $\overline{K} \times 1$, and

$$\Delta(P,P) \equiv \left(\begin{array}{cc} \Delta_\varphi(R)\Delta'_\varphi(R) & \Delta_\varphi(R)\Delta'_\psi(\overline{P}) \\ \Delta_\psi(\overline{P})\Delta'_\varphi(R) & \Delta_\psi(\overline{P})\Delta'_\psi(\overline{P}) \end{array} \right) \quad \text{be } (L+\overline{K}) \times (L+\overline{K}).$$

Recall that for *i.i.d.* data, we have

$$\|v_n^*\|_{sd}^2 = E\left[\left|\Delta_\varphi(Z_1, h_o)[v_{h_n}^* + v_{\Gamma_n}^*] + \Delta_\psi(Z_2, \alpha_o)[v_{g_n}^*]\right|^2\right].$$

Lemma C.4 *Let $\partial_h\rho(\alpha_o)[R]$ and $\partial_g\rho(\alpha_o)[\overline{P}]$ denote $L\times 1$ and $\overline{K}\times 1$ vectors with the j-th elements being $\partial_h\rho(\alpha_o)[r_j]$ and $\partial_g\rho(\alpha_o)[\overline{p}_j]$ respectively. Let $R_\varphi(h_o)[R,R]$ and $R_\psi(\alpha_o)[\overline{P},\overline{P}]$ be $L\times L$ and $\overline{K}\times\overline{K}$ matrices with the (a,b)-th elements being $R_\varphi(h_o)[r_a,r_b]$ and $R_\psi(\alpha_o)[\overline{p}_a,\overline{p}_b]$ respectively. Let $\Gamma(\alpha_o)\left[R, v_{g_n}^*\right]$ be $L\times 1$ vector with the j-th element being $\Gamma(\alpha_o)\left[r_j, v_{g_n}^*\right]$ and $\Gamma\left(\alpha_o\right)\left[R,\overline{P}\right]$ be $L\times\overline{K}$ matrix with the (a,b)-th element being $\Gamma\left(\alpha_o\right)[r_a,\overline{p}_b]$. We then have*

$$v_{h_n}^*(\cdot) = R(\cdot)'\left[R_\varphi(h_o)[R,R]\right]^{-1}\partial_h\rho(\alpha_o)[R] \equiv R(\cdot)'\beta_{h_n}^*; \tag{99}$$

$$v_{g_n}^*(\cdot) = \overline{P}(\cdot)'\left[R_\psi(\alpha_o)[\overline{P},\overline{P}]\right]^{-1}\partial_g\rho(\alpha_o)[\overline{P}_2] \equiv \overline{P}(\cdot)'\beta_{g_n}^*; \tag{100}$$

$$v_{\Gamma_n}^*(\cdot) = R(\cdot)'\left[R_\varphi(h_o)[R,R]\right]^{-1}\Gamma(\alpha_o)\left[R, v_{g_n}^*\right] \equiv R(\cdot)'\beta_{\Gamma_n}^* \tag{101}$$

and

$$\beta_{\Gamma_n}^* = \left[R_\varphi(h_o)[R,R]\right]^{-1}\Gamma\left(\alpha_o\right)\left[R,\overline{P}\right]\beta_{g_n}^*. \tag{102}$$

Proof. Let $\beta_{h_n}^*$ be such that $v_{h_n}^* = R(\cdot)'\beta_{h_n}^*$. Note that we have

$$\partial_h\rho(\alpha_o)\left[r_j\right] = \left\langle r_j, v_{h_n}^*\right\rangle_\varphi = R_\varphi(h_o)\left[r_j, R'(\cdot)\beta_{h_n}^*\right] = R_\varphi(h_o)\left[r_j, R\right]'\beta_{h_n}^*$$

for all j. It follows that $\partial_h\rho(\alpha_o)\left[R\right] = R_\varphi(h_o)\left[R,R\right]\beta_{h_n}^*$ and hence $\beta_{h_n}^* = \left(R_\varphi(h_o)\left[R,R\right]\right)^{-1}\partial_h\rho(\alpha_o)\left[R\right]$, which proves (99). We can prove (100) and (101) similarly, and the details are omitted. As for (102), by definition we have

$$\Gamma(\alpha_o)\left[R, v_{g_n}^*\right] = \Gamma\left(\alpha_o\right)\left[R,\overline{P}\right]\beta_{g_n}^*.$$

Using (101), we obtain

$$\beta_{\Gamma_n}^* = \left[R_\varphi(h_o)[R,R]\right]^{-1}\Gamma(\alpha_o)\left[R, v_{g_n}^*\right] = \left[R_\varphi(h_o)[R,R]\right]^{-1}\Gamma\left(\alpha_o\right)\left[R,\overline{P}\right]\beta_{g_n}^*.$$

■

Lemma C.5

$$\|v_n^*\|_{sd}^2 = \left(\begin{array}{cc} \beta_{h_n}^{*\prime}, & \beta_{g_n}^{*\prime}\end{array}\right)\left(\begin{array}{cc} I_L & 0 \\ A_n' & I_{\overline{K}}\end{array}\right)E\left[\Delta(P,P)\right]\left(\begin{array}{cc} I_L & A_n \\ 0 & I_{\overline{K}}\end{array}\right)\left(\begin{array}{c}\beta_{h_n}^* \\ \beta_{g_n}^*\end{array}\right). \tag{103}$$

where

$$A_n \equiv \left[R_\varphi(h_o)[R,R]\right]^{-1}\Gamma\left(\alpha_o\right)\left[R,\overline{P}\right] \quad \textit{is } L\times\overline{K}. \tag{104}$$

Proof. Because $\|v_n^*\|_{sd}^2 = \text{Var}\left[\Delta_\varphi(Z_{1,i}, h_o)[v_{h_n}^* + v_{\Gamma_n}^*] + \Delta_\psi(Z_{2,i}, \alpha_o)[v_{g_n}^*]\right]$, it follows that

$$\|v_n^*\|_{sd}^2 = \left(\begin{array}{cc} \beta_{h_n}^{*\prime} + \beta_{\Gamma_n}^{*\prime}, & \beta_{g_n}^{*\prime} \end{array} \right) E\left[\Delta(P,P)\right] \left(\begin{array}{cc} \beta_{h_n}^{*\prime} + \beta_{\Gamma_n}^{*\prime}, & \beta_{g_n}^{*\prime} \end{array} \right)'. \tag{105}$$

Combining with (102) in Lemma C.4, we then obtain

$$v_{\Gamma_n}^*(\cdot) = R(\cdot)' A_n \beta_{g_n}^* \tag{106}$$

Finally, combining (105) and (106), we obtain the desired conclusion. ■

In the rest of this appendix, we provide explicit expressions of sieve variances for the two functional considered in Section 4 when $g_o = (\theta_o, m_o(\cdot))$. The first functional is $\rho(\alpha) = \lambda'\theta$ for some $\lambda \neq 0$ in $\mathbb{R}^{d_\theta}$, where $\alpha = (h(\cdot), g(\cdot)) = (h(\cdot), \theta, m(\cdot))$.

Lemma C.6 *For $\rho(\alpha) = \lambda'\theta$, we have*

$$\left\|v_{\theta,n}^*\right\|_{sd}^2 = \lambda' \left(I_{\theta\theta,n}^{(1)}, -I_{\theta\theta,n}^{(1)} I_{\theta m,n} I_{mm,n}^{-1}\right) E\left[V_{g,i} V_{g,i}'\right] \left(I_{\theta\theta,n}^{(1)}, -I_{\theta\theta,n}^{(1)} I_{\theta m,n} I_{mm,n}^{-1}\right)' \lambda, \tag{107}$$

where $V_{g,i} = \left(\Gamma(\alpha_o)\left[R, \overline{P}\right]\right)' \left[R_\varphi(h_o)[R,R]\right]^{-1} \Delta_\varphi(Z_{1,i}, h_o)[R] + \Delta_\psi(Z_{2,i}, \alpha_o)[\overline{P}]$.

Proof. In this case, it is clear that we have $v_g(\cdot) = (v_\theta, v_m(\cdot))$, where $v_m(\cdot) = P(\cdot)'\beta_K$. Hence for any $v_g \in \mathcal{V}_{g_n}$ with $v_g = (v_\theta, v_m(\cdot))$ and $v_m(\cdot) = P_2'(\cdot)\beta_K$, we can write

$$\langle v_g, v_g \rangle_\psi = (v_\theta', \beta_K') R_\psi(\alpha_o)[\overline{P}, \overline{P}](v_\theta', \beta_K')'.$$

By definition, we have

$$R_\psi(\alpha_o)[\overline{P}, \overline{P}] = \left(\begin{array}{cc} I_{\theta\theta,n} & I_{\theta m,n} \\ I_{m\theta,n} & I_{mm,n} \end{array} \right)$$

where $I_{\theta\theta,n}$ and $I_{mm,n}$ are $d_\theta \times d_\theta$ and $L \times L$ matrices respectively.

Because $\partial_g \rho(\alpha_o)\left[\overline{P}\right] = (\lambda', \mathbf{0}_L')'$, we can use (100) in Lemma C.4 along with the partitioned inverse formula, and conclude

$$\beta_{g_n}^* \equiv \left(\left(\beta_{\theta_n}^*\right)', (\beta_K^*)'\right)' = \left(\lambda' I_{\theta\theta,n}^{(1)}, -\lambda' I_{\theta\theta,n}^{(1)} I_{\theta m,n} I_{mm,n}^{-1}\right)', \tag{108}$$

where $I_{\theta\theta,n}^{(1)} = \left(I_{\theta\theta,n} - I_{\theta m,n} I_{mm,n}^{-1} I_{m\theta,n}\right)^{-1}$. Likewise, we can conclude that $\beta_{h_n}^* = 0$. We therefore obtain

$$\left(\begin{array}{cc} \beta_{h_n}^{*\prime} & \beta_{g_n}^{*\prime} \end{array} \right) \left(\begin{array}{cc} I_L & 0 \\ A_n' & I_{\overline{K}} \end{array} \right) \left(\begin{array}{c} \Delta_\varphi(R) \\ \Delta_\psi(\overline{P}) \end{array} \right) = \beta_{g_n}^{*\prime} A_n' \Delta_\varphi(R) + \beta_{g_n}^{*\prime} \Delta_\psi(\overline{P}) = \beta_{g_n}^{*\prime} V_g. \tag{109}$$

The conclusion follows by combining (108) and (109) with Lemma C.5. ■

From (108), we see that the Riesz representer $v_g^*(\cdot)$ of the functional $\rho(\alpha_o) = \lambda'\theta_o$ takes the following form:

$$v_g^*(\cdot) = \left(\lambda' I_{\theta\theta,n}^{(1)}, -\lambda' I_{\theta\theta,n}^{(1)} I_{\theta m,n} I_{mm,n}^{-1}\right)' \overline{P}(\cdot) = \left(\lambda' I_{\theta\theta,n}^{(1)} \mathbf{1}_{d_\theta}, -\lambda' I_{\theta\theta,n}^{(1)} I_{\theta m,n} I_{mm,n}^{-1} P(\cdot)\right). \tag{110}$$

Lemma C.7 *For $\rho(\alpha) = \rho(m_o)$, we have*

$$\left\|v_{m,n}^*\right\|_{sd}^2 = \partial\rho(m_o)[P]' \left(-I_{mm,n}^{(1)} I_{m\theta,n} I_{\theta\theta,n}^{-1},\ I_{mm,n}^{(1)}\right) E\left[V_{g,i} V_{g,i}'\right] \left(-I_{mm,n}^{(1)} I_{m\theta,n} I_{\theta\theta,n}^{-1},\ I_{mm,n}^{(1)}\right)' \partial\rho(m_o)[P]. \tag{111}$$

Proof. Because $\partial_g \rho(\alpha_o)\left[\overline{P}_2\right] = \left(\mathbf{0}_{d_\theta}', \partial\rho(m_o)[P]'\right)'$, we obtain

$$\beta_{g_n}^* = \left(R_\psi(\alpha_o)[\overline{P}_2, \overline{P}_2]\right)^{-1} \left(\mathbf{0}_{d_\theta}', \partial\rho(m_o)[P]'\right)' = \left(-\partial\rho(m_o)[P]' I_{mm,n}^{(1)} I_{m\theta,n} I_{\theta\theta,n}^{-1},\ \partial\rho(m_o)[P]' I_{mm,n}^{(1)}\right)'$$

so we have

$$\left(\beta_{g_n}^*\right)' = \partial\rho(m_o)[P]' \left(-I_{mm,n}^{(1)} I_{m\theta,n} I_{\theta\theta,n}^{-1},\ I_{mm,n}^{(1)}\right)$$

Likewise, we can conclude that $\beta_{h_n}^* = 0$. We therefore obtain

$$\left(\begin{array}{cc} \beta_{h_n}^{*\prime} & \beta_{g_n}^{*\prime} \end{array}\right) \left(\begin{array}{cc} I_{k_1(n)} & 0 \\ A_n' & I_{k_2(n)} \end{array}\right) \left(\begin{array}{c} \Delta_\varphi(P_1) \\ \Delta_\psi(\overline{P}_2) \end{array}\right) = \beta_{g_n}^{*\prime} A_n' \Delta_\varphi(P_1) + \beta_{g_n}^{*\prime} \Delta_\psi(\overline{P}_2) = \beta_{g_n}^{*\prime} V_g.$$

As in the previous lemma, the conclusion follows by using Lemma C.5. ■

Proof of Proposition 6.1. Recalling that our estimator of the asymptotic variance is nothing but the sample analog of $\left\|v_{\theta,n}^*\right\|_{sd}^2$, we can obtain from Lemma C.6 that

$$\left\|\widehat{v}_{\theta,n}^*\right\|_{sd}^2 = \left(\widehat{\beta}_{\theta_n}^{*\prime}, -\widehat{\beta}_{\theta_n}^{*\prime} \widehat{I}_{\theta m,n} \widehat{I}_{mm,n}^{-1}\right) \frac{\sum_{i=1}^n \widehat{V}_{g,i} \widehat{V}_{g,i}'}{n} \left(\widehat{\beta}_{\theta_n}^{*\prime}, -\widehat{\beta}_{\theta_n}^{*\prime} \widehat{I}_{\theta m,n} \widehat{I}_{mm,n}^{-1}\right)',$$
$$\left\|\widehat{v}_{m,n}^*\right\|_{sd}^2 = \left(-\widehat{\beta}_{m_n}^{*\prime} \widehat{I}_{m\theta,n} \widehat{I}_{\theta\theta,n}^{-1},\ \widehat{\beta}_{m_n}^{*\prime}\right) \frac{\sum_{i=1}^n \widehat{V}_{g,i} \widehat{V}_{g,i}'}{n} \left(-\widehat{\beta}_{m_n}^{*\prime} \widehat{I}_{m\theta,n} \widehat{I}_{\theta\theta,n}^{-1},\ \widehat{\beta}_{m_n}^{*\prime}\right)',$$

which proves Proposition 6.1. ■

Proof of Theorem 6.2. By definition, we have

$$\Gamma_n(\widehat{\alpha}_n)\left[R, \overline{P}\right] = n^{-1} \sum_{i=1}^n \frac{\partial^2 \psi(Z_{2,i}, \widehat{g}(\cdot), \widehat{h}(\cdot))}{\partial\gamma_L \partial\beta_g'},$$
$$n^{-1} \sum_{i=1}^n r_\psi(Z_{2,i}, \widehat{\alpha}_n)[\overline{P}, \overline{P}] = n^{-1} \sum_{i=1}^n \frac{\partial^2 \psi\left(Z_{2,i}, \widehat{g}(\cdot), \widehat{h}(\cdot)\right)}{\partial\beta_g \partial\beta_g'},$$
$$\Delta_\varphi(Z_1, \widehat{h}_n)[R] = \frac{\partial\varphi(Z_1, \widehat{h}_n(\cdot))}{\partial\gamma_L} \text{ and } \Delta_\psi(Z_2, \widehat{\alpha}_n)[\overline{P}] = \frac{\partial\psi(Z_2, \widehat{g}(\cdot), \widehat{h}(\cdot))}{\partial\beta_g}.$$

It follows that $\widehat{I}_{\psi,\theta\theta} = \widehat{I}_{\theta\theta,n}$, etc., $\widehat{V}_{22} = n^{-1}\sum_{i=1}^{n} \widehat{V}_{g,i}\widehat{V}_{g,i}'$, from which we obtain the conclusion. ■

D Proof of Results in Section 5

In this appendix, we provide the sufficient conditions of Theorem 5.1. For any column vector a, let $\|a\|$ denote its ℓ_2-norm. For any square matrix A, the operator norm is denoted by $||A||$. Let $\|h\|_{\mathcal{H}} = \|h\|_2$ and $\|g\|_{\mathcal{G}} = \|g\|_2$. Let C denote some generic finite constant larger than 1. By definition, $g_{o,n}$ and $h_{o,n}$ are the projections of g_o and h_o under the L_2-norm respectively. Let $\mathcal{X}$ denote the support of x. For ease of notations, we omit the dependence of $L(n)$ and $K(n)$ on n, and write $L(n) = L$ and $K(n) = K$.

Assumption D.1 *(i) The data $\{y_i, x_i, s_i\}_{i=1}^n$ is i.i.d.; (ii) $E\left[\varepsilon_i^4 \middle| x_i\right] < C$ and $E[\varepsilon_i^2 \middle| x_i] > C^{-1}$; (iii) there exist $\rho_h > 0$ and $\gamma_{o,L} \in \mathbb{R}^L$ such that*

$$\|h_{o,L} - h_o\|_\infty = O(L^{-\rho_h})$$

where $h_{o,L}(\cdot) \equiv R(\cdot)'\gamma_{o,L}$; (iv) the eigenvalues of Q_L are between C^{-1} and C for all L; (v) there exists a nondecreasing sequence ζ_L such that $\sup_{x\in\mathcal{X}} \|R(x)\| \le \zeta_L$.

We assume that the support of ε_i is $\mathcal{E} = [a, b]$, where a, b are finite constants. Define $\mathcal{E}_\eta = [a-\eta, b+\eta]$ for some small $\eta > 0$. For d a nonnegative integer, let $|g|_d = \max_{|\tau|\le d} \sup_{\varepsilon\in\mathcal{E}} |\partial^\tau g(\varepsilon)|$ for any $g \in \mathcal{G}$.

Assumption D.2 *(i) $E[u_i^4 \middle| \varepsilon_i] < C$ and $E[u_i^2 \middle| \varepsilon_i] > C^{-1}$; (ii) $g_o(\varepsilon)$ is twice continuously differentiable; (iii) there exist $\rho_g > 0$ and $\beta_{o,K} \in \mathbb{R}^K$ such that*

$$|g_{o,K} - g_o|_d = O(K^{-\rho_g})$$

where $g_{o,K}(\cdot) = P(\cdot)'\beta_{o,K}$ and $d = 1$; (iv) the eigenvalues of Q_K are between C^{-1} and C for all K; (v) for $j = 0, 1, 2$, there exists a nondecreasing sequence $\xi_{j,K}$ such that $\sup_{\varepsilon\in\mathcal{E}_\eta} \left\|\partial^j P(\varepsilon)\right\| \le \xi_{j,K}$.

Assumptions D.1 and D.2 impose restrictions on the moments of the error terms u and ε, and the smoothness of unknown functions h_o and g_o. Let $\upsilon_{j,K} = \sup_{\varepsilon\in\mathcal{E}_\eta} \left\|\partial^j P(\varepsilon)'\beta_{o,K}\right\|$ for $j = 1, 2$. Under Assumptions D.1 and D.2, and restrictions on the number of sieve basis functions in Assumption D.4 below, we show in Lemma 1.1 of the Supplemental Appendix that

$$||\widehat{h}_n - h_o|| = O_p(\delta_{h,n}^*) \text{ and } \|\widehat{g}_n - g_o\| = O_p(\delta_{2,n}^*)$$

where $\delta_{h,n}^* = L^{1/2}n^{-1/2} + L^{-\rho_h}$ and $\delta_{2,n}^* = K^{1/2}n^{-1/2} + K^{-\rho_g} + \upsilon_{1,K}\delta_{h,n}^*$. Let $\delta_{2,n} = \delta_{2,n}^* \log(\log(n))$, then we have $\widehat{g}_n \in \mathcal{N}_{g,n}$ wpa1 where $\mathcal{N}_{g,n} = \{g \in \mathcal{G}_n : \|g - g_o\|_2 \le \delta_{2,n}\}$.

The sequence $\upsilon_{1,K}$ effects the rate of convergence of $\widehat{g}_n$ derived in Lemma 1.1 of the Supplemental Appendix. When the unknown function $g_o(\cdot)$ is well defined and is continuously differentiable in the extended support $\mathcal{E}_\eta$, and $\beta_{o,K}$ satisfies

$$\max_{|j|\le 2} \sup_{\varepsilon\in\mathcal{E}_\eta} \left|\partial^j P(\varepsilon)'\beta_{o,K} - \partial^j g_o(\varepsilon)\right| = O(K^{-\rho_g}), \tag{112}$$

then we have

$$\upsilon_{j,K} \leq \sup_{\varepsilon \in \mathcal{E}_\eta} \left| \partial^j P(\varepsilon)' \beta_{o,K} - \partial^j g_o(\varepsilon) \right| + \sup_{\varepsilon \in \mathcal{E}_\eta} \left| \partial^j g_o(\varepsilon) \right| = O(1). \tag{113}$$

Let $\omega_{\min}(Q_K)$ denote the smallest eigenvalue of Q_K. Then by the Cauchy inequality and Assumptions D.2.(ii)-(v),

$$\upsilon_{j,K} \leq \sup_{\varepsilon \in \mathcal{E}_\eta} \left\| \partial^j P(\varepsilon) \right\| \left\| \beta_{o,K} \right\| \leq \sup_{\varepsilon \in \mathcal{E}_\eta} \left\| \partial^j P(\varepsilon) \right\| \omega_{\min}^{-1}(Q_K) \left[\left\| g_{o,K} - g_o \right\|_2 + \left\| g_o \right\|_2 \right] \leq C \xi_{j,K}, \tag{114}$$

which provides a upper bound for $\upsilon_{j,K}$ when $g_o(\cdot)$ does not satisfy the extended support condition (113).

Assumption D.3 *(i)* $||v_{g_n}^*||_2 \geq C$ *for all* n*; (ii) the functional* $\rho(\cdot)$ *satisfies*

$$\sup_{g \in \mathcal{N}_{g,n}} \left| \frac{\rho(g) - \rho(g_o) - \partial \rho(g_o)[g - g_o]}{\|v_n^*\|_{sd}} \right| = o(n^{-1/2});$$

(iii) $\left| \|v_n^*\|_{sd}^{-1} \, \partial \rho(g_o)[g_{o,n} - g_o] \right| = o(n^{-1/2})$*; (iv)* $\sup_{g \in \mathcal{N}_{g,n}} \|\partial \rho(g)[P] - \partial \rho(g_o)[P]\| = o(1)$.

Assumption D.3.(i) requires that the L_2-norm of the Riesz representer $v_{g_n}^*$ is uniformly bounded from below by some fixed constant. This condition together with the condition $E[u_i^2 \big| \varepsilon_i] > C^{-1}$ in Assumption D.2.(i) imply that the sieve variance is bounded away from zero. Assumptions D.3.(ii)-(iv) imposes conditions on the functional $\rho(\cdot)$. These conditions can be easily verified when the functional $\rho(\cdot)$ is linear.

Assumption D.4 *The following conditions hold:*
(i) $n^{-1/2}(K+L)^{1/2}(\xi_{0,K} + \zeta_L)(\log(n))^{1/2} = o(1)$*;*
(ii) $n^{-1}(L\xi_{1,K}^2 \log(n) + \zeta_L \xi_{1,K}) = o(1)$*;*
(iii) $n^{-1/2}\zeta_L(L\xi_{2,K} + L^{1/2}\xi_{1,K})(n^{-1/2}K^{1/2} + K^{-\rho_g} + \upsilon_{1,K} n^{-1/2} L^{1/2}) \log(n) = o(1)$*;*
(iv) $n^{-1/2}\zeta_L(L + L^{1/2}\upsilon_{1,K} + L\upsilon_{2,K}) \log(n) = o(1)$*;*
(v) $nL^{1-2\rho_h} + K^{-\rho_g} = o(1)$.

Assumption D.4 imposes restrictions on the numbers of basis functions L and K. These conditions are needed to show the asymptotic normality of the two-step plug-in sieve estimator.

Assumption D.5 *The following conditions hold:*
(i) $||v_{g_n}^*||_2 \leq C$ *for all* n.
(ii) $(n^{-1}K\xi_{1,K}^2 + (\zeta_L^2 + \xi_{0,K}^2 + \xi_{1,K}^2)K^{-2\rho_g}) \log(n) = o(1)$*;*
(iii) $n^{-1}(\zeta_L^2 + \xi_{0,K}^2 + \xi_{1,K}^2)\upsilon_{1,K}^2 L \log(n) = o(1)$.

Assumption D.5 is needed to show the consistency of sieve variance estimator. Assumption D.5.(i) imposes upper bound on the L_2-norm of the Riesz representer $||v_{g_n}^*||_2$, which holds when the linear functional $\partial \rho(g_o)[\cdot]$ is bounded on the space $\mathcal{V}_2$. Assumptions D.5.(ii)-(iii) include extra conditions on the numbers of basis functions. When the power series are used in the second step estimation, we have

$\xi_{j,K} \leq CK^{1+2j}$. When the power series are also used in the first step estimation, $\zeta_L \leq CL$ and the sufficient conditions for Assumptions D.4.(i)-(iv) and D.5.(ii)-(iii) become

$$n^{-1}(K^7 + LK^6 + L^2K^{11/2} + L^{5/2}K^5 + L^4)\log(n) = o(1) \text{ and } (L + K^3)K^{-\rho_g}\log(n) = o(1), \quad (115)$$

when the extended support condition (113) holds. If the spline or trigonometric series are used in the first step, we have $\zeta_L \leq CL^{1/2}$ and the sufficient conditions for Assumptions D.4.(i)-(iv) become

$$n^{-1}(K^7 + LK^6 + L^{3/2}K^{11/2} + L^2K^5 + L^3)\log(n) = o(1) \text{ and } (L^{1/2} + K^3)K^{-\rho_g}\log(n) = o(1), \quad (116)$$

when the extended support condition (113) holds. When the extended support condition (113) does not hold, the sufficient conditions for Assumptions D.4.(i)-(iv) become

$$n^{-1}(L^4K^{10} + K^{12})\log(n) = o(1) \text{ and } (L + K^3)K^{-\rho_g}\log(n) = o(1), \quad (117)$$

when the power series are used in the first step, and

$$n^{-1}(L^3K^{10} + K^{12})\log(n) = o(1) \text{ and } (L^{1/2} + K^3)K^{-\rho_g}\log(n) = o(1), \quad (118)$$

when the splines or trigonometric series are used in the first step.

Proof of Theorem 5.1. (i) By Lemma 1.3, Lemma 1.4 and Lemma 1.5 in the Supplemental Appendix, we see that Assumptions 3.1, 3.2 and 3.3 hold. Hence, Theorem 5.1.(i) follows by Theorem 3.1.

(ii) Lemma 1.6 in the Supplemental Appendix implies that Assumptions 4.1 and 4.2 hold. Assumptions 4.3.(i)-(iii) are only used in proving Lemma C.3 which is verified in Lemma 1.7 of the Supplemental Appendix. Assumptions 4.3.(iv) are proved in Lemma 1.8 of the Supplemental Appendix. Hence by Theorem 4.1, Theorem 5.1.(i) and the continuous mapping Theorem, Theorem 5.1.(ii) holds. ■