SIEVE SEMIPARAMETRIC TWO-STEP GMM UNDER WEAK DEPENDENCE

By

Xiaohong Chen and Zhipeng Liao

July 2015

COWLES FOUNDATION DISCUSSION PAPER NO. 2012

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles. yale.edu/

Sieve Semiparametric Two-Step GMM Under Weak Dependence*

Xiaohong Chen†, and Zhipeng Liao‡

First version: April 2011; Revised version: June 2015

Abstract

This paper considers semiparametric two-step GMM estimation and inference with weakly dependent data, where unknown nuisance functions are estimated via sieve extremum estimation in the first step. We show that although the asymptotic variance of the second-step GMM estimator may not have a closed form expression, it can be well approximated by sieve variances that have simple closed form expressions. We present consistent or robust variance estimation, Wald tests and Hansen's (1982) over-identification tests for the second step GMM that properly reflect the first-step estimated functions and the weak dependence of the data. Our sieve semiparametric two-step GMM inference procedures are shown to be numerically equivalent to the ones computed as if the first step were parametric. A new consistent random-perturbation estimator of the derivative of the expectation of the non-smooth moment function is also provided.

JEL Classification: C12, C22, C32, C140

Keywords: Sieve two-step GMM; Weakly dependent data; Auto-correlation robust inference; Semiparametric over-identification test; Numerical equivalence; Random perturbation derivative estimator

*We gratefully acknowledge insightful comments from Jinyong Hahn, who was a co-author of the initial version. We appreciate useful suggestions from the coeditor, the associate editor, three anonymous referees, D. Ackerberg, L. Benkard, R. Blundell, S. Khan, W. Newey, A. Pakes, F. Penaranda, D. Pouzo, G. Sanchez, E. Sentana and participants in July 2011 Econometrics Society Australasian Meeting in Adelaide, in May 2012 Cemmap Masterclass on Semi-nonparametric Models and Methods in London, in May 2012 Carlos III - CEMFI joint seminar, in August 2012 NBER-NSF CEME Conference on The Econometrics of Dynamic Games at NYU, and in November 2012 Info-metrics & Nonparametric Inference at UC Riverside. Any errors are the responsibility of the authors.

†Corresponding author. Cowles Foundation for Research in Economics, Yale University, Box 208281, New Haven, CT 06520 USA. Tel.: +1 203 432 5852. Fax: +1 203 432 6167. Email: xiaohong.chen@yale.edu

‡Department of Economics, UCLA, Box 951477, Los Angeles, CA 90095 USA. Email: zhipeng.liao@econ.ucla.edu

1 Introduction

Flexible parametric estimation as a substitute for full blown nonparametric estimation has now become a standard tool kit in empirical analysis in nonlinear models with weakly dependent data including time series, panel time series, spatial and network models. See, e.g., Engle et al (1986), Engle and Gonzalez-Rivera (1991), Engle and Ng (1993), Bansal and Viswanathan (1993), Gallant and Tauchen (1989, 1996), Gallant, Hansen and Tauchen (1990), Gallant, Hsieh and Tauchen (1991), Conley and Dupor (2003), Engle and Rangel (2008), Chen and Ludvigson (2009), Engle (2010), Kawai (2011), Chen, Favilukis and Ludvigson (2013), Lee and Robinson (2013), to name only a few. See Chen (2007, 2013) for additional references. In our view, the strategy of using flexible parametric estimation can be interpreted as nonparametric estimation, where the researchers make a nonparametric "promise" to increase the complexity of the parametric models as the sample size grows. In other words, these empirical papers are in fact engaged in nonparametric estimation. Natural questions that arise are (1) under what conditions this interpretation can be rigorously justified; and (2) how one should modify the inference procedures in light of such a nonparametric interpretation.

In this paper, we shall provide formal justifications of such empirical practices in a broad context of *sieve* semiparametric two-step GMM estimation and inference for models with weakly dependent data. We consider simple inference on a finite dimensional parameter θ_o that is (over-) identified by a set of unconditional moment restrictions depending on unknown infinite dimensional nuisance functions $h_o()$. The unknown $h_o()$ is identified as a maximizer to a non-random criterion over some function space, and is consistently estimated by a sieve extremum estimator $\widehat{h}_n()$ in the first step. And the unknown θ_o is estimated by Hansen's (1982) GMM estimator $\widehat{\theta}_n$ in the second step, based on the sample moment restrictions depending on $\widehat{h}_n$.

Our sieve semiparametric two-step GMM is a special case of the more general semiparametric two-step GMM with *any* consistent nonparametric estimator of $h_o()$ in the first step. The existing literature has largely focused on the situation where θ_o is root-n consistently estimable, where n is the sample size. Newey (1994), Chen, Linton and van Keilegom (2003, CLvK), Chen (2007, theorem 4.1) and others already establish the root-n consistency and asymptotic normality (CAN) of the second step GMM estimator $\widehat{\theta}_n$, $\sqrt{n}(\widehat{\theta}_n - \theta_o) \rightarrow_d \mathcal{N}(0, V_\theta)$. These general theories do not impose any specific structures on $h_o()$ or its consistent estimators, rendering the characterization and estimation of the asymptotic variance V_θ difficult in diverse empirical applications. CLvK (2003) and Chen (2007) relax the smooth moment conditions imposed in Newey (1994) and allow for dependent data, but without providing any variance estimator for $\widehat{\theta}_n$. CLvK (2003) and Armstrong, Bertanha and Hong (2014) establish bootstrap consistency in terms of approximating the asymptotic normal distribution of $\widehat{\theta}_n$, but without any variance estimation, either. To the best of our knowledge, there is no published results on the long-run variance (LRV) estimation, auto-correlation robust inference and overidentification test of semiparametric two-step GMM with any nonparametric first step and weakly dependent data.

In this paper, we provide a characterization of the asymptotic variance V_θ of our sieve semiparametric two-step GMM estimator $\widehat{\theta}_n$ with weakly dependent data. We show that although the V_θ may not have a closed form expression, it can be well approximated by sieve variances that have simple closed form expressions. Next, we provide simple valid inference procedures, such as confidence sets construction, Wald tests and over-identification tests, for the semiparametric two-step GMM that properly reflect the first-step sieve estimated nuisance functions and the weak dependence of the data. In particular, we propose different inference procedures using asymptotically pivotal statistics based on two kinds of estimators of V_θ. The first one is a kernel based heteroskedasticity and autocorrelation consistent (HAC) estimator that is inspired by Newey and West (1987), Andrews (1991) and others for parametric time series models. The second one is a robust orthonormal series estimator that is inspired by Phillips (2005) and Sun (2013) for parametric time series models. In addition, we provide a new consistent random-perturbation estimator of the derivative of the expectation of the non-smooth moment function, which is used for the semiparametric variance estimation and inference based on the second-step GMM. This new derivative estimator is extremely easy to compute and is an attractive alternative to numerical derivative estimator of non-smooth moments for multivariate θ.

Our paper is the first to provide these inference results for semiparametric two-step GMM with sieve extreme estimation in the first step, allowing for non-smooth moment and dependent data. Our inference results are useful not only to financial and macro nonlinear time series models, but also to semiparametric structural models in IO, labor, trade, social networks and others with temporal or/and spatial dependent data.

There are two kinds of smoothing parameters needed for inferences based on sieve semiparametric two-step GMMs for dependent data. The first is to choose the sieve (approximating) dimension in the first-step estimation of $h_o()$; the second is to choose the bandwidth parameter in the LRV estimation for the second-step GMM procedure. It is known that sieve extremum estimators have the so-called "small bias property" (SBP). That is, when the Euclidean parameter is root-n consistently estimable, the sieve dimension could be chosen to achieve the optimal nonparametric convergence rates. In particular, the regularity conditions (in Appendix A) for $\widehat{\theta}_n$ to be root-n consistent allow the sieve dimension in the first step to be chosen in a data-driven way, such as Lepski method, AIC and others, that either balances the bias and the standard deviation or makes the bias a smaller order of the standard deviation (of $\widehat{h}_n$). In our simulation studies, we used the simple AIC for selecting the sieve dimension. Our inference results in Sections 3 and 4 allow for the second-step LRV estimation bandwidth parameters to be chosen as if the GMM moment restrictions depends on a "parametric" first step.

We also derive results that are expected to have a practical appeal; we show that in terms of implementation in finite samples, empirical researchers can ignore the semiparametric nature of the model and obtain simple estimators of the V_θ and conduct inference using existing softwares "as if" $h_o()$ were parametrically specified. That is, from the computational point of view, we could assume that the linear sieve approximation in the first step provides a "correct" parametric specification, and based on which we derive another parametric asymptotic variance, $V_{\theta,P}$, of $\widehat{\theta}_n$. While the *semiparametric* asymptotic variance V_θ may not have a closed-form expression in general, the *parametric* asymptotic variance $V_{\theta,P}$ has a closed-form expression. Hence it is easy to compute estimate of $V_{\theta,P}$ using existing softwares for parametric two-step GMM with weakly dependent data. We show that our estimate of the semiparametric asymptotic variance V_θ is numerically identical to the estimate of the parametric asymptotic variance $V_{\theta,P}$. This result generalizes those in Newey (1994) and Ackerberg, Chen and Hahn (2012) to more general overidentified semiparametric GMM with any linear sieve extremum estimation in the first step. It greatly simplifies the computation of standard errors and inference based on semiparametric two-step GMM with weakly dependent data, and provides a formal first-order asymptotic justification for flexible parametric estimation and inference in empirical work under weak dependence.

The rest of the paper is organized as follows. Section 2 characterizes the semiparametric asymptotic variance V_θ of $\widehat{\theta}_n$ for weakly dependent data. Section 3 presents inference results based on kernel HAC estimate of V_θ. Section 4 presents inference results based on robust orthogonal series estimate of V_θ. Section 5 provides numerical equivalent ways to compute estimates of V_θ. Section 6 proposes new consistent estimators of average derivatives of non-smooth moment functions. Section 7 conducts simulation experiments to investigate the finite sample performances of our inference methods. Section 8 concludes by mentioning extensions to sieve semiparametric two-step GMM when $\widehat{\theta}_n$ converges to θ_o at a slower than root-n rate. Most of the regularity conditions and the proofs are contained in Appendix.

2 Sieve Semiparametric Two-step GMM Estimator

This section introduces a sieve semiparametric two-step GMM estimator $\widehat{\theta}_n$, and characterizes its semiparametric asymptotic variance V_θ for weakly dependent data.

2.1 The Model and the Estimator

The model Suppose that the data $\{Z_t = (Y_t', X_t')'\}_{t=1}^n$ is weakly dependent and is defined on a complete probability space. We denote Θ for a finite dimensional parameter set (a compact subset of $\mathbb{R}^{d_\theta}$) and $\mathcal{H}$ for an infinite dimensional parameter set. Let $\theta_o \in int(\Theta)$ and $h_o \in \mathcal{H}$ denote the pseudo-true unknown finite and infinite dimensional parameters. Let $g(\cdot,\cdot,\cdot) : \mathbb{R}^{d_z} \times \Theta \times \mathcal{H} \to \mathbb{R}^{d_g}$ be a vector measurable functions with $d_g \geq d_\theta$. Let $Q_n() : \mathcal{H} \to \mathbb{R}$ be a non-random criterion function. A semiparametric structural model specifies that

$$E\left[\frac{1}{n}\sum_{i=1}^{n} g(Z_i, \theta, h_o(\cdot,\theta))\right] = 0 \quad \text{at } \theta = \theta_o \in \Theta, \tag{1}$$

and for any fixed $\theta \in \Theta$, $h_o(\cdot,\theta) \in \mathcal{H}$ solves

$$Q_n(h_o) = \sup_{h\in\mathcal{H}} Q_n(h). \tag{2}$$

If $h_o()$ were known, the finite dimensional structural parameter θ_o is (over-)identified by d_g ($\geq d_\theta$) moment conditions (1). But $h_o()$ is in fact unknown, except that it is identified as a maximizer of a non-random criterion function $Q_n()$ over $\mathcal{H}$. As in Newey (1994) and CLvK, we allow that the function $h_o \in \mathcal{H}$ can depend on the parameters θ and the data. We usually suppress the arguments of the function h_o for notational convenience; thus: $(\theta, h) \equiv (\theta, h(\cdot,\theta))$, $(\theta, h_o) \equiv (\theta, h_o(\cdot,\theta))$, and $(\theta_o, h_o) \equiv (\theta_o, h_o(\cdot,\theta_o))$. We also allow the moment functions $g(Z_i, \theta, h(\cdot))$ to depend on the entire functions $h(\cdot)$ and not just their values at observed data points.

Definition of sieve semiparametric two-step GMM estimators In the **first-step** the unknown nuisance functions $h_o()$ is estimated via an *approximate sieve extremum estimation*, i.e.,

$$\widehat{Q}_n(\widehat{h}_n) \geq \sup_{h\in\mathcal{H}_{k(n)}} \widehat{Q}_n(h) - o_p(n^{-1}), \tag{3}$$

where $\widehat{Q}_n()$ is a random criterion function such that $\sup_{h\in\mathcal{H}_{k(n)}} |\widehat{Q}_n(h) - Q_n(h)| = o_p(1)$, and $\mathcal{H}_{k(n)}$ is a sieve space for $\mathcal{H}$ (i.e., a sequence of approximating parameter spaces that become dense in

$\mathcal{H}$ as $k(n) \to \infty$). In the **second-step**, the first-step sieve extremum estimator $\widehat{h}_n$ is plugged into some unconditional moment conditions and the unknown θ_o is estimated by GMM

$$\widehat{\theta}_n = \arg\min_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^{n} g\left(Z_i, \theta, \widehat{h}_n\right) \right]' W_n \left[\frac{1}{n} \sum_{i=1}^{n} g\left(Z_i, \theta, \widehat{h}_n\right) \right], \tag{4}$$

where W_n is a $d_g \times d_g$ positive definite (possibly random) matrix.

Discussion Our definition of sieve semiparametric two-step GMM estimation consists of equations (3) and (4). As demonstrated in Chen (2007), sieve extremum estimation in the first-step is very flexible and can estimate unknown functions in most nonparametric models. More precisely, a semiparametric structural model specifies a non-random criterion $Q_n()$ that is maximized at $h_o() \in \mathcal{H}$, which in turn suggests a special case of the first-step sieve extremum estimation. For example, if an economic model specifies h_o as a solution to $\sup_{\mathcal{H}} E\left[\frac{1}{n}\sum_{i=1}^{n} \varphi(Z_i, h)\right]$ for some measurable function $\varphi(\cdot,\cdot) : \mathbb{R}^{d_z} \times \mathcal{H} \to \mathbb{R}$, then the first step usually takes a form of sieve M-estimation (e.g., least squares, quantile, quasi maximum-likelihood) with $Q_n(h) = E\left[\frac{1}{n}\sum_{i=1}^{n} \varphi(Z_i, h)\right]$ and

$$\widehat{Q}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \varphi(Z_i, h). \tag{5}$$

If an economic model specifies a conditional moment restriction $E[\rho(Z, h_o)|X] = 0$, the first-step could be a sieve MD estimation with $Q_n(h) = E\left[\frac{-1}{2n}\sum_{i=1}^{n} m(X_i, h)' m(X_i, h)\right]$ and

$$\widehat{Q}_n(h) = -\frac{1}{2n} \sum_{i=1}^{n} \widehat{m}(X_i, h)' \widehat{m}(X_i, h) \tag{6}$$

where $\widehat{m}(X, h)$ is a consistent estimate of the conditional mean function $m(X, h) = E[\rho(Z, h)|X]$. See Chen (2007) for additional examples of different criterion functions $Q_n()$, $\widehat{Q}_n()$ and different sieves $\mathcal{H}_{k(n)}$.

2.2 Asymptotic Normality of Sieve Semiparametric Two-step GMM Estimator

In this subsection we characterize the asymptotic variance V_θ of the semiparametric two-step GMM estimator. Practitioners who are not interested in the asymptotic justification and only care about the practical implications may want to skip the rest of this section, and just read Section 5.

Heuristic review of the existing theory To simplify the presentation, in the rest of the paper we assume that $\{Z_i = (Y_i', X_i')'\}_{i=1}^n$ is strictly stationary weakly dependent, and that $Z_i = (Y_i', X_i')'$ has the same distribution as that of $Z = (Y', X')'$. Let $Q(h) = Q_n(h)$ and $G(\theta, h) = E\left[g\left(Z, \theta, h\right)\right]$. For any $(\theta, h) \in \Theta \times \mathcal{H}$, we denote the ordinary derivative of $G(\theta, h)$ with respect to θ as $\Gamma_1(\theta, h)$. For any $\theta \in \Theta$, we say that $G(\theta, h)$ is pathwise differentiable at $h \in \mathcal{H}$ in the direction v, if $\{h + \tau v : \tau \in [0, 1]\} \subset \mathcal{H}$ and

$$\Gamma_2(\theta, h)[v] = \left(\Gamma_{2,1}(\theta, h)[v], ..., \Gamma_{2,d_g}(\theta, h)[v]\right)' \equiv \left.\frac{\partial G\left[\theta, h(\cdot) + \tau v(\cdot)\right]}{\partial \tau}\right|_{\tau=0}$$

exits.[1] Let $\alpha_o = (\theta_o, h_o)$, $\Gamma_1 = \Gamma_1(\alpha_o)$ and W be the probability limit of W_n. Throughout the paper we assume that $\Gamma_1' W \Gamma_1$ is non-singular.

Let $G_n(\theta, h) = \frac{1}{n}\sum_{i=1}^n g\left(Z_i, \theta, h\right)$ and $\|\theta\|_E = \sqrt{\theta'\theta}$. For general semiparametric two-step GMM estimation with *any* consistent nonparametric $\widehat{h}_n$ in the first step, under mild condition we have

$$(\widehat{\theta}_n - \theta_o) = -\left(\left(\Gamma_1' W \Gamma_1\right)^{-1} \Gamma_1' W + o_p\left(1\right)\right) G_n(\theta_o, \widehat{h}_n) \quad \text{and} \quad \left\|\widehat{\theta}_n - \theta_o\right\|_E \asymp \left\|G_n(\theta_o, \widehat{h}_n)\right\|_E.$$

Suppose the stochastic equicontinuty condition holds:

$$\frac{\left\|G_n(\theta_o, \widehat{h}_n) - G_n(\theta_o, h_o) - \{G(\theta_o, \widehat{h}_n) - G(\theta_o, h_o)\}\right\|_E}{\|G_n(\theta_o, h_o)\|_E + \left\|G(\theta_o, \widehat{h}_n) - G(\theta_o, h_o)\right\|_E} = o_p(1). \tag{7}$$

Then

$$\left\|\widehat{\theta}_n - \theta_o\right\|_E \lesssim \|G_n(\theta_o, h_o)\|_E + \left\|G(\theta_o, \widehat{h}_n) - G(\theta_o, h_o)\right\|_E.$$

Note that $\sqrt{n} G_n(\theta_o, h_o) = O_p(1)$ under mild conditions, we have: $\sqrt{n}\left\|\widehat{\theta}_n - \theta_o\right\|_E \to \infty$ whenever $\sqrt{n}\left\|G(\theta_o, \widehat{h}_n)\right\|_E \to \infty$; and $\sqrt{n}(\widehat{\theta}_n - \theta_o) = O_p(1)$ if $\sqrt{n} G(\theta_o, \widehat{h}_n) = O_p(1)$. Suppose the nonlinear remainder condition holds:

$$\frac{\left\|G(\theta_o, \widehat{h}_n) - G(\theta_o, h_o) - \Gamma_2(\alpha_o)[\widehat{h}_n - h_o]\right\|_E}{\left\|\Gamma_2(\alpha_o)[\widehat{h}_n - h_o]\right\|_E} = o_p(1). \tag{8}$$

[1] Note that $g\left(Z, \theta, h\right)$ is a d_g dimensional vector of moment functions. Hence $\Gamma_2(\theta, h)[\cdot]$ is a d_g dimensional vector of functionals.

Then $\sqrt{n}(\widehat{\theta}_n - \theta_o) = O_p(1)$ if

$$\sqrt{n}\Gamma_2(\alpha_o)[\widehat{h}_n - h_o] = O_p(1) \tag{9}$$

and hence $\widehat{\theta}_n$ satisfies

$$\sqrt{n}(\widehat{\theta}_n - \theta_o) = -\left(\Gamma_1' W \Gamma_1\right)^{-1} \Gamma_1' W \sqrt{n}\left[G_n(\alpha_o) + \Gamma_2(\alpha_o)[\widehat{h}_n - h_o]\right] + o_p(1). \tag{10}$$

Without specifying how h_o is estimated in the first step, CLvK directly assumes that

$$\sqrt{n}\left[G_n(\alpha_o) + \Gamma_2(\alpha_o)[\widehat{h}_n - h_o]\right] \to_d \mathcal{N}(0, V_1) \text{ for a finite positive definite } V_1; \tag{11}$$

while Newey (1994) assumes that there is a zero-mean and finite second moment "adjustment" term $\mathbf{\Delta}_i^*$ such that

$$\sqrt{n}\Gamma_2(\alpha_o)[\widehat{h}_n - h_o] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{\Delta}_i^* + o_p(1) = O_p(1). \tag{12}$$

Condition (11) and equation (10) imply that $\widehat{\theta}_n$ is $\sqrt{n}$-consistent and asymptotically normally distributed (CAN) with the asymptotic variance $V_\theta = (\Gamma_1' W \Gamma_1)^{-1} (\Gamma_1' W V_1 W \Gamma_1) (\Gamma_1' W \Gamma_1)^{-1}$, where the long-run variance (LRV) V_1 captures the first-order asymptotic effect of the first-step nonparametric estimation of $h_o()$.

Unfortunately, without specifying any primitive nature of the unknown function $h_o()$, it is often difficult to verify any of these conditions ((9), (11) or (12)), and the LRV V_1 (and V_θ) typically has no analytic expression for complicated semiparametric models.

Riesz representation for (9) We assume that $\mathcal{H}$ is a vector space of functions endowed with a pseudo-metric $||\cdot||_{\mathcal{H}}$, which is a problem specific strong-norm metric with respect to the θ-argument and a pseudo-metric with respect to all the other arguments. For example when $\mathcal{H}$ is a class of continuous functions mapping from $\mathbf{Z} \times \Theta$ to $\mathbb{R}$ and having finite sup-norms, we may take $||h||_{\mathcal{H}} = \sup_\theta ||h(\cdot,\theta)||_\infty = \sup_\theta \sup_z |h(z,\theta)|$ or $||h||_{\mathcal{H}} = \sup_\theta ||h(\cdot,\theta)||_{L_r(P)} = \sup_\theta \{\int |h(Z,\theta)|^r dP\}^{1/r}$ for $1 \leq r < \infty$. Under mild conditions, any first-step nonparametric estimator $\widehat{h}$, being it a kernel, local linear regression or sieve estimator, is consistent: $||\widehat{h} - h_o||_{\mathcal{H}} = o_p(1)$.

Since h_o is the unique maximizer of $Q(h)$ over $\mathcal{H}$, within any shrinking $||\cdot||_{\mathcal{H}}-$neighborhood, $\mathcal{B}_o$, of h_o, we can define a local pseudo-metric

$$\|h-h_o\| = \left\{ - \left[\frac{\partial^2}{\partial \tau^2} Q(h_o + \tau(h-h_o)) \right] \Big|_{\tau=0} \right\}^{1/2} \text{ for any } h \in \mathcal{B}_o, \tag{13}$$

where $||\cdot||_{\mathcal{H}}$ is chosen such that $\|h-h_o\| \leq const. \times ||h-h_o||_{\mathcal{H}}$ for any $h \in \mathcal{B}_o$. See, for example, Chen and Shen (1998) for M estimation $Q(h) = E\left[\varphi\left(Z, h\right)\right]$ with

$$\|h-h_o\|^2 = - \left[\frac{\partial^2}{\partial \tau^2} E\left\{ \varphi\left(Z, h_o + \tau(h-h_o)\right)\right\} \right] \Big|_{\tau=0}, \tag{14}$$

and Ai and Chen (2003) for MD estimation $Q(h) = -E\left[m(X,h)'m(X,h)\right]/2$ with

$$\|h-h_o\|^2 = E\left(\frac{\partial m(X, h_o+\tau(h-h_o))}{\partial \tau}\Big|_{\tau=0}' \frac{\partial m(X, h_o+\tau(h-h_o))}{\partial \tau}\Big|_{\tau=0} \right). \tag{15}$$

Let $\mathcal{V}$ be the closed linear span of $\mathcal{H} - \{h_o\}$ under $\|\cdot\|$. Let $\langle \cdot, \cdot \rangle$ be the inner-product induced by $\|\cdot\|$.

We assume that the linear functional $\Gamma_{2,j}(\alpha_o)[\cdot] : (\mathcal{V}, \|\cdot\|) \to \mathbb{R}$ is *bounded*, i.e.

$$\sup_{v \in \mathcal{V}, v \neq 0} \left\{ \frac{|\Gamma_{2,j}(\alpha_o)[v]|}{\|v\|} \right\} < \infty \text{ for all } j = 1, ..., d_g.$$

By Riesz representation theorem, for each $j = 1, ..., d_g$, the functional $\Gamma_{2,j}(\alpha_o)[\cdot]$ is bounded if and only if (iff) there is a Riesz representer $v_j^* \in \mathcal{V}$ such that

$$\Gamma_{2,j}(\alpha_o)[v] = \left\langle v, v_j^* \right\rangle \text{ for all } v \in \mathcal{V} \text{ and } \left\| v_j^* \right\| = \sup_{v \in \mathcal{V}, v \neq 0} \frac{|\Gamma_{2,j}(\alpha_o)[v]|}{\|v\|} < \infty. \tag{16}$$

As we will see later in this subsection, the Riesz representers v_j^* for $j = 1, ..., d_g$ play an important role in the asymptotic variance V_θ of $\widehat{\theta}_n$. However, v_j^* for $j = 1, ..., d_g$ may not have closed form solutions in general, which limits their usefulness in the empirical applications.[2] We next provide a sieve approximation of the Riesz representers v_j^* for $j = 1, ..., d_g$ which always have explicit expressions.

[2] For example, an additive nonparametric regression in the first step and an average derivative in the second step.

Since $(\mathcal{V}, \|\cdot\|)$ is a Hilbert space, there is an increasing sequence of finite-dimensional Hilbert spaces $(\mathcal{V}_{k(n)}, \|\cdot\|)$ that is dense in $(\mathcal{V}, \|\cdot\|)$ as $k(n) \to \infty$. Denote $k(n) = \dim(\mathcal{V}_{k(n)})$. For each $k(n) < \infty$, the restricted linear functional $\Gamma_{2,j}(\alpha_o)[\cdot] : \mathcal{V}_{k(n)} \to \mathbb{R}$ is always bounded and hence there always exists a sieve Riesz representer $v^*_{j,k(n)} \in \mathcal{V}_{k(n)}$ such that

$$\Gamma_{2,j}(\alpha_o)[v] = \left\langle v, v^*_{j,k(n)} \right\rangle \text{ for all } v \in \mathcal{V}_{k(n)} \text{ and } \left\| v^*_{j,k(n)} \right\| = \sup_{v \in \mathcal{V}_{k(n)}, v \neq 0} \frac{|\Gamma_{2,j}(\alpha_o)[v]|}{\|v\|} < \infty. \tag{17}$$

Moreover, the Riesz representer $v^*_j \in \mathcal{V}$ defined in (16) exists iff $\lim_{k(n)\to\infty} ||v^*_{j,k(n)}|| < \infty$. For such a case we have $||v^*_j|| = \lim_{k(n)\to\infty} ||v^*_{j,k(n)}||$ and $\lim_{k(n)\to\infty} ||v^*_{j,k(n)} - v^*_j|| = 0$.

Although the Riesz representer $v^*_j \in \mathcal{V}$ may not have a closed form solution, the sieve Riesz representer $v^*_{j,k(n)} \in \mathcal{V}_{k(n)}$ always has a closed form expression. To see this, let h_o be a real valued function, and let $\{p_j\}_{j=1}^{\infty}$ be a complete basis for the infinite dimensional Hilbert space $(\mathcal{V}, \|\cdot\|)$. Let $P_{k(n)}(\cdot) = \left(p_1(\cdot), ..., p_{k(n)}(\cdot)\right)'$. Then $\mathcal{V}_{k(n)} = \{v(\cdot) = P_{k(n)}(\cdot)'\gamma : \gamma \in \mathbb{R}^{k(n)}\}$ become dense in $(\mathcal{V}, \|\cdot\|)$ as $k(n) \to \infty$. By definition, the sieve Riesz representer $v^*_{j,k(n)}(\cdot) = P_{k(n)}(\cdot)'\gamma^*_{j,k(n)} \in \mathcal{V}_{k(n)}$ of $\Gamma_{2,j}(\alpha_o)[\cdot] : \mathcal{V}_{k(n)} \to \mathbb{R}$ solves the following optimization problem:

$$\left\| v^*_{j,k(n)} \right\|^2 = \sup_{\gamma \in \mathbb{R}^{k(n)}, \gamma \neq 0} \frac{\gamma' F_{j,k(n)} F'_{j,k(n)} \gamma}{\gamma' R_{k(n)} \gamma}, \tag{18}$$

where $F_{j,k(n)} = \Gamma_{2,j}(\alpha_o)[P_{k(n)}(\cdot)] \equiv \left(\Gamma_{2,j}(\alpha_o)[p_1(\cdot)], ..., \Gamma_{2,j}(\alpha_o)[p_{k(n)}(\cdot)]\right)'$ is a $k(n) \times 1$ vector, and $R_{k(n)}$ is a $k(n) \times k(n)$ positive definite matrix such that

$$\gamma' R_{k(n)} \gamma \equiv - \left[\frac{\partial^2}{\partial \tau^2} Q(h_o(\cdot) + \tau P_{k(n)}(\cdot)'\gamma) \right] \Bigg|_{\tau=0} \quad \text{for all } \gamma \in \mathbb{R}^{k(n)}. \tag{19}$$

It is obvious that the optimal solution of γ in (18) has a closed-form expression:

$$\gamma^*_{j,k(n)} = (R_k)^- F_{j,k(n)}.$$

The sieve Riesz representer is then given by

$$v^*_{j,k(n)}(\cdot) = P_{k(n)}(\cdot)'\gamma^*_{j,k(n)} = P_{k(n)}(\cdot)' \left(R_{k(n)}\right)^- F_{j,k(n)} \in \mathcal{V}_{k(n)} \tag{20}$$

for $j = 1, ..., d_g$, where $\left(R_{k(n)}\right)^-$ is a generalized inverse of $R_{k(n)}$.

Root-n CAN of sieve semiparametric two-step GMM For simplicity we let $\mathcal{H}_{k(n)}$ be an increasing sequence of approximating parameter spaces that become dense in $\mathcal{H}$ under $||\cdot||_{\mathcal{H}}$ as $k(n) = \dim(\mathcal{H}_{k(n)}) \to \infty$ (i.e., for any $h \in \mathcal{H}$ there is an element $\pi_n h$ in $\mathcal{H}_{k(n)}$ satisfying $||h - \pi_n h||_{\mathcal{H}} \to 0$ as $k(n) \to \infty$). Under mild conditions and for weakly dependent data, the first-step sieve extremum estimator is consistent under $||\cdot||_{\mathcal{H}}$ (see, e.g., Chen (2007)). Define $h_{o,n} \in \arg\min_{h\in\mathcal{H}_{k(n)}} \|h - h_o\|$. Let $\mathcal{V}_{k(n)}$ be a closed linear span of $\mathcal{H}_{k(n)} - \{h_{o,n}\}$ under $\|\cdot\|$. Since $\mathcal{V}$ is the closed linear span of $\mathcal{H} - \{h_o\}$ under $\|\cdot\|$ and $\|\cdot\| \leq const. \times ||\cdot||_{\mathcal{H}}$, we have that the closure of $\cup_{k(n)}\mathcal{V}_{k(n)}$ is dense in $\mathcal{V}$ under $\|\cdot\|$. Let $\mathbf{v}_n^* = \mathbf{v}_{k(n)}^* = (v_{1,k(n)}^*, ..., v_{d_g,k(n)}^*)'$ be the sieve Riesz representer as defined in (17) that corresponds to the sieve $\mathcal{V}_{k(n)}$.

For each fixed z, $\Delta(z, h_o)[\cdot] : (\mathcal{V}, \|\cdot\|) \to \mathbb{R}$ is a linear map such that

$$E\left(\Delta(Z, h_o)[v]\right) = \left[\frac{\partial}{\partial\tau}Q(h_o + \tau v)\right]\bigg|_{\tau=0}.$$

See, for example, Chen and Shen (1998) for M estimation

$$\Delta(Z, h_o)[v] = \left.\frac{\partial\varphi(Z, h_o + \tau v)}{\partial\tau}\right|_{\tau=0}, \tag{21}$$

and Ai and Chen (2003) for MD estimation

$$\Delta(Z, h_o)[v] = -\left(\left.\frac{\partial m(X, h_o + \tau v)}{\partial\tau}\right|_{\tau=0}\right)' \rho(Z, h_o). \tag{22}$$

Suppose that $\max_{j=1,\dots,d_g} \lim_{k(n)\to\infty} ||v_{j,k(n)}^*|| < \infty$, then the Riesz representer $\mathbf{v}^* = (v_1^*, ..., v_{d_g}^*)'$ given in (16) exists. Under mild additional conditions (see Appendix A for details), Newey's condition (12) will be satisfied with the adjustment term Δ_i^* given by

$$\boldsymbol{\Delta}_i^* = \Delta(Z_i, h_o)\left[\mathbf{v}^*\right] = \left[\Delta(Z, h_o)[v_1^*], \dots, \Delta(Z, h_o)[v_{d_g}^*]\right]'. \tag{23}$$

Equations (11), (12) and (23) immediately lead to

$$V_1 = Avar\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{g\left(Z_i, \alpha_o\right) + \Delta(Z_i, h_o)\left[\mathbf{v}^*\right]\right\}\right). \tag{24}$$

Expression (24) for the LRV V_1 would be very useful if the Riesz representer $\mathbf{v}^*$ (and hence $\mathbf{\Delta}_i^* = \Delta(Z_i, h_o)[\mathbf{v}^*]$) could be computed in a closed-form, which is, unfortunately, not the case for complicated semiparametric problems. Let

$$S_{i,n}^* = S_i(\alpha_o)[\mathbf{v}_n^*] = g(Z_i, \alpha_o) + \Delta(Z_i, h_o)[\mathbf{v}_n^*] \tag{25}$$

be the sieve score, where $\Delta(Z, h_o)[\mathbf{v}_n^*] = \left[\Delta(Z, h_o)[v_{1,k(n)}^*], \ldots, \Delta(Z, h_o)[v_{d_g,k(n)}^*]\right]'$ is a "sieve influence function" approximating the possibly unknown "adjustment" term $\mathbf{\Delta}_i^* = \Delta(Z_i, h_o)[\mathbf{v}^*]$. Let

$$V_{1,n}^* = E\left(n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} S_{i,n}^* S_{j,n}^{*\prime}\right) \tag{26}$$

be the sieve LRV. Since the sieve Riesz representer $\mathbf{v}_n^*$ can be computed in a closed-form (20), the sieve score $S_{i,n}^*$ and the sieve LRV $V_{1,n}^*$ have closed-form expressions.

The next theorem establishes the $\sqrt{n}-$CAN of a sieve semiparametric two-step GMM estimator.

Theorem 2.1 *Let* $\max_{j=1,\ldots,d_g} \lim_{k(n)\to\infty} ||v_{j,k(n)}^*|| < \infty$ *and Assumptions A.1 and A.2 in Appendix A hold. Then: the sieve semiparametric two-step GMM estimator satisfies* $\sqrt{n}(\widehat{\theta}_n - \theta_o) \to_d \mathcal{N}(0, V_\theta)$*, where*

$$V_\theta = \left(\Gamma_1' W \Gamma_1\right)^{-1}\left(\Gamma_1' W V_1 W \Gamma_1\right)\left(\Gamma_1' W \Gamma_1\right)^{-1}, \tag{27}$$

and

$$V_1 = \lim_{n\to\infty} V_{1,n}^* = Avar\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{g\left(Z_i, \alpha_o\right) + \Delta(Z_i, h_o)[\mathbf{v}^*]\right\}\right). \tag{28}$$

We may want to consider choosing W to minimize the asymptotic variance (27) of $\widehat{\theta}_n$. An obvious choice is $W = V_1^{-1}$, in which case the V_θ simplifies to

$$V_\theta^o = \left(\Gamma_1' V_1^{-1} \Gamma_1\right)^{-1}. \tag{29}$$

We call the estimator $\widehat{\theta}_n$ with $W_n = V_1^{-1} + o_p(1)$ a "semiparametric two-step optimally weighted GMM". It is more efficient than other semiparametric two-step GMM estimators with $W_n \neq$

$V_1^{-1} + o_p(1)$, but we cannot say that it is an estimator that achieves the semiparametric efficiency bound. In this paper we call any weight matrix satisfying $W_n = V_1^{-1} + o_p(1)$ a "limited information optimal weight matrix".

2.3 Estimation of Semiparametric Asymptotic Variance

Theorem 2.1 can be a basis of inference about the unknown parameter θ_o. Equation (27) suggests that V_θ can be estimated by

$$\widehat{V}_\theta = \left(\widehat{\Gamma}_1' W_n \widehat{\Gamma}_1\right)^{-1} \left(\widehat{\Gamma}_1' W_n \widehat{V}_1 W_n \widehat{\Gamma}_1\right) \left(\widehat{\Gamma}_1' W_n \widehat{\Gamma}_1\right)^{-1}, \tag{30}$$

where $\widehat{\Gamma}_1$ and $\widehat{V}_1$ are estimates of Γ_1 and V_1 respectively.

If the moment function $g(Z, \theta, h)$ is differentiable in θ_o, then a standard textbook-level analysis and the consistency of $\widehat{\alpha}_n = (\widehat{\theta}_n, \widehat{h}_n)$ can be used to show that

$$\widehat{\Gamma}_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial g(Z_i, \widehat{\alpha}_n)}{\partial \theta'} \tag{31}$$

would be consistent for Γ_1. See Section 6 for an alternative estimate of Γ_1 when $g(Z, \theta, h)$ is not differentiable in θ_o.

Theorem 2.1 states that $V_1 = \lim_{n\to\infty} V_{1,n}^*$, where the sieve LRV $V_{1,n}^*$ can be estimated based on an estimate of the sieve score $S_{i,n}^* = S_i(\alpha_o)\,[\mathbf{v}_n^*]$:

$$\widehat{S}_{i,n}^* = \widehat{S}_i(\widehat{\alpha}_n)\,[\widehat{\mathbf{v}}_n^*] = g(Z_i, \widehat{\alpha}_n) + \widehat{\Delta}(Z_i, \widehat{h}_n)[\widehat{\mathbf{v}}_n^*], \tag{32}$$

where $\widehat{\Delta}(Z, h)[\cdot]$ is some estimate of $\Delta(Z, h)\,[\cdot]$ for any h in a local neighborhood of h_o. See, for example, Chen, Liao and Sun (2014) for sieve M estimation

$$\widehat{\Delta}(Z, \widehat{h}_n)[v] = \left. \frac{\partial \varphi(Z, \widehat{h}_n + \tau v)}{\partial \tau} \right|_{\tau=0} \tag{33}$$

and Ai and Chen (2003) for sieve MD estimation

$$\widehat{\Delta}(Z, \widehat{h}_n)[v] = -\left(\left. \frac{\partial \widehat{m}(X, \widehat{h}_n + \tau v)}{\partial \tau} \right|_{\tau=0} \right)' \rho(Z, \widehat{h}_n). \tag{34}$$

For $j=1,...,d_g$, when the moment function $g_j(Z,\theta,h)$ is pathwise differentiable in h_o, the functional $\Gamma_{2,j}(\alpha_o)[\cdot]$ can be estimated by

$$\Gamma_{2,j,n}(\widehat{\alpha}_n)[\cdot] = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial g_j(Z_i,\widehat{\alpha}_n)}{\partial h}[\cdot],$$

(see Section 6 for an alternative estimate of $\Gamma_{2,j}(\alpha_o)[\cdot]$ when $g_j(Z,\theta,h)$ is not pathwise differentiable in h_o.) The estimate $\widehat{v}^*_{j,k(n)}$ of $v^*_{j,k(n)}$ is the Riesz representer of the functional $\Gamma_{2,j,n}(\widehat{\alpha}_n)[\cdot]$ on $\mathcal{V}_{k(n)}$, i.e. $\widehat{v}^*_{j,k(n)}$ ($j=1,...,d_g$) satisfies

$$\Gamma_{2,j,n}(\widehat{\alpha}_n)[v] = \left\langle v, \widehat{v}^*_{j,k(n)}\right\rangle_n \text{ for all } v\in\mathcal{V}_{k(n)} \text{ and } \left\|\widehat{v}^*_{j,k(n)}\right\|_n = \sup_{v\in\mathcal{V}_{k(n)},v\neq 0}\frac{|\Gamma_{2,j,n}(\widehat{\alpha}_n)[v]|}{\|v\|_n}<\infty, \quad (35)$$

where $\|\cdot\|_n$ is the empirical semi-norm associated with the theoretical semi-norm $\|\cdot\|$, defined as

$$||v||_n = \left\{-\left[\frac{\partial^2}{\partial\tau^2}\widehat{Q}_n(\widehat{h}_n+\tau v)\right]\Big|_{\tau=0}\right\}^{1/2} \text{ for any } v\in\mathcal{V}, \quad (36)$$

and $\langle\cdot,\cdot\rangle_n$ is the empirical inner product induced by the empirical semi-norm $\|\cdot\|_n$. Again if h_o is a real valued function and $\mathcal{V}_{k(n)} = \{v(\cdot) = P_{k(n)}(\cdot)'\gamma : \gamma\in\mathbb{R}^{k(n)}\}$, then $\widehat{\mathbf{v}}^*_n = (\widehat{v}^*_{1,k(n)},...,\widehat{v}^*_{d_g,k(n)})'$ defined in (35) can be computed in a closed form: for $j=1,...,d_g$,

$$\widehat{v}^*_{j,k(n)} = P_{k(n)}(\cdot)'\widehat{\gamma}^*_{j,k(n)} = P_{k(n)}(\cdot)'\left(\widehat{R}_{k(n)}\right)^{-}\widehat{F}_{j,k(n)}, \quad (37)$$

where $\widehat{F}_{j,k(n)} = \Gamma_{2,j,n}(\widehat{\alpha}_n)[P_{k(n)}(\cdot)]$ and $\widehat{R}_{k(n)}$ is such that

$$\gamma'\widehat{R}_{k(n)}\gamma \equiv -\left[\frac{\partial^2}{\partial\tau^2}\widehat{Q}_n(\widehat{h}_n+\tau P_{k(n)}(\cdot)'\gamma)\right]\Big|_{\tau=0} \quad \text{for all } \gamma\in\mathbb{R}^{k(n)}. \quad (38)$$

These are useful in establishing the numerical equivalence results in Section 5.

The following lemma states the consistency of the empirical Riesz representer $\widehat{\mathbf{v}}^*_n$ in (35) for the theoretical sieve Riesz representer $\mathbf{v}^*_n$.

Lemma 2.1 *Let Assumption A.3 in Appendix A hold with some positive sequence $\delta_{w,n}=o(1)$. Then:*

$$\max_{j=1,...,d_g}||\widehat{v}^*_{j,k(n)} - v^*_{j,k(n)}|| = O_p(\delta_{w,n}) = o_p(1).$$

The above lemma serves as a key ingredient to study properties of two classes of estimates of V_θ, which are considered in the subsequent two sections.

3 Inference Based on Consistent LRV Estimate

In this section we provide inference and over-identifying specification test based on a consistent estimate of the LRV V_1 given in (28).

3.1 Consistent LRV Estimation and Wald Test

We can rewrite the LRV V_1 in (28) as $V_1 = \lim_{n\to\infty} V_{1,n}^*$ with

$$V_{1,n}^* = \sum_{i=-n+1}^{n-1} \Upsilon_i(\alpha_o)\left[\mathbf{v}_n^*, \mathbf{v}_n^*\right], \tag{39}$$

where

$$\Upsilon_i(\alpha_o)\left[\mathbf{v}_n^*, \mathbf{v}_n^*\right] = \begin{cases} \frac{1}{n}\sum_{l=i+1}^{n} E\left[S_{l,n}^* S_{l-i,n}^{*\prime}\right] & \text{for } i \geq 0 \\ \frac{1}{n}\sum_{l=-i+1}^{n} E\left[S_{l,n}^* S_{l+i,n}^{*\prime}\right] & \text{for } i < 0 \end{cases}$$

with $S_{l,n}^* = S_l(\alpha_o)\left[\mathbf{v}_n^*\right]$ given in (25).

The intuition from Newey-West, e.g., suggests the following strategy for estimating (39). Let $\mathcal{K}(\cdot)$ be a kernel function that satisfies Assumption B.1. We define a kernel-based estimator of V_1 as

$$\widehat{V}_{1,n} = \sum_{i=-n+1}^{n-1} \mathcal{K}\left(\frac{i}{M_n}\right) \Upsilon_{n,i}(\widehat{\alpha}_n)\left[\widehat{\mathbf{v}}_n^*, \widehat{\mathbf{v}}_n^*\right], \tag{40}$$

where $M_n \to \infty$ as $n \to \infty$, and

$$\Upsilon_{n,i}\left(\widehat{\alpha}_n\right)\left[\widehat{\mathbf{v}}_n^*, \widehat{\mathbf{v}}_n^*\right] = \begin{cases} \frac{1}{n}\sum_{l=i+1}^{n} \widehat{S}_{l,n}^* \widehat{S}_{l-i,n}^{*\prime} & \text{for } i \geq 0 \\ \frac{1}{n}\sum_{l=-i+1}^{n} \widehat{S}_{l,n}^* \widehat{S}_{l+i,n}^{*\prime} & \text{for } i < 0 \end{cases} \tag{41}$$

with $\widehat{S}_{l,n}^* = \widehat{S}_l(\widehat{\alpha}_n)\left[\widehat{\mathbf{v}}_n^*\right]$ given in (32). We could also estimate V_1 by a centered version of $\widehat{V}_{1,n}$:

$$\widehat{V}_{c,1,n} = \sum_{i=-n+1}^{n-1} \mathcal{K}\left(\frac{i}{M_n}\right) \overline{\Upsilon}_{n,i}(\widehat{\alpha}_n)\left[\widehat{\mathbf{v}}_n^*, \widehat{\mathbf{v}}_n^*\right], \tag{42}$$

where

$$\overline{\Upsilon}_{n,i}(\widehat{\alpha}_n)\left[\widehat{\mathbf{v}}_n^*, \widehat{\mathbf{v}}_n^*\right] = \begin{cases} \frac{1}{n}\sum_{l=i+1}^{n} \left(\widehat{S}_{l,n}^* - \widehat{S}_n^*\right)\left(\widehat{S}_{l-i,n}^* - \widehat{S}_n^*\right)' & \text{for } i \geq 0 \\ \frac{1}{n}\sum_{l=-i+1}^{n} \left(\widehat{S}_{l,n}^* - \widehat{S}_n^*\right)\left(\widehat{S}_{l+i,n}^* - \widehat{S}_n^*\right)' & \text{for } i < 0 \end{cases},$$

and $\widehat{S}_n^* = \frac{1}{n}\sum_{l=1}^{n}\widehat{S}_{l,n}^*$.

Theorem 3.1 *Let conditions of Theorem 2.1 and Lemma 2.1 hold. Suppose that Assumptions B.1 and B.2 in Appendix B hold. Then:*

$$\widehat{V}_{1,n} = V_1 + o_p(1); \tag{43}$$

$$\widehat{V}_{c,1,n} = V_1 + o_p(1). \tag{44}$$

Using the kernel LRV estimate $\widehat{V}_{1,n}$ (or $\widehat{V}_{c,1,n}$) and the estimate of the average derivative $\widehat{\Gamma}_1$, we can define

$$\widehat{V}_{\theta,n} = \left(\widehat{\Gamma}_1' W_n \widehat{\Gamma}_1\right)^{-1}\left(\widehat{\Gamma}_1' W_n \widehat{V}_{1,n} W_n \widehat{\Gamma}_1\right)\left(\widehat{\Gamma}_1' W_n \widehat{\Gamma}_1\right)^{-1}. \tag{45}$$

If $\widehat{\Gamma}_1 \to_p \Gamma_1$ and $W_n \to_p W$, then invoking Theorem 3.1, we have that

$$\widehat{V}_{\theta,n} \to_p \left(\Gamma_1' W \Gamma_1\right)^{-1}\left(\Gamma_1' W V_1 W \Gamma_1\right)\left(\Gamma_1' W \Gamma_1\right)^{-1} = V_\theta. \tag{46}$$

By the consistency of $\widehat{V}_{\theta,n}$, the asymptotic normality of $\widehat{\theta}_n$ and the Slutsky theorem, we also have that

$$\sqrt{n}\widehat{V}_{\theta,n}^{-1/2}(\widehat{\theta}_n - \theta_o) \to_d \mathcal{N}\left(0, I_{d_\theta}\right). \tag{47}$$

The above weak convergence is directly applicable for conducting inference about θ_o. For example the standard Wald test of $\theta = \theta_o$ follows from

$$C_n = n(\widehat{\theta}_n - \theta_o)'\widehat{V}_{\theta,n}^{-1}(\widehat{\theta}_n - \theta_o) \to_d \chi^2_{d_\theta} \tag{48}$$

where $\chi^2_{d_\theta}$ denotes a chi-square distributed random variable with degree of freedom d_θ.

3.2 Over-identification Test

In this subsection, we present an over-identification test of the moment restrictions $E\left[g\left(Z, \theta_o, h_o\right)\right] = 0$ by taking into account the fact that the nonparametric component h_o has to be estimated in the first step.

Inspired by Hansen's (1982) over-identification J test of the parametric moment restrictions $E\left[g\left(Z,\theta_o\right)\right]=0$, we will construct our over-identification test of $E\left[g\left(Z,\theta_o,h_o\right)\right]=0$ based on a limited information optimal weight matrix $W_n=V_1^{-1}+o_p(1)$ and a semiparametric two-step optimally weighted GMM estimator.

Let $\widetilde{\alpha}_n=(\widetilde{\theta}_n,\widehat{h}_n)$ be a preliminary consistent estimator of $\alpha_o=(\theta_o,h_o)$, where $\widetilde{\theta}_n$ could be a sieve semiparametric two-step GMM estimator with an arbitrary weight matrix W_n (say an identity, but the details are not important). We compute $\widetilde{S}_{i,n}^{*}=\widehat{S}_i\left(\widetilde{\alpha}_n\right)\left[\widetilde{\mathbf{v}}_n^{*}\right]=g(Z_i,\widetilde{\alpha}_n)+\widehat{\Delta}(Z_i,\widehat{h}_n)[\widetilde{\mathbf{v}}_n^{*}]$ and $\Upsilon_{n,i}\left(\widetilde{\alpha}_n\right)\left[\widetilde{\mathbf{v}}_n^{*},\widetilde{\mathbf{v}}_n^{*}\right]$ as in (41), and then compute the weight matrix

$$\widetilde{W}_n=\left(\sum_{i=-n+1}^{n-1}\mathcal{K}\left(\frac{i}{M_n}\right)\Upsilon_{n,i}\left(\widetilde{\alpha}_n\right)\left[\widetilde{\mathbf{v}}_n^{*},\widetilde{\mathbf{v}}_n^{*}\right]\right)^{-1} \tag{49}$$

as in (40). We can then go on to compute the second-step GMM estimator $\widehat{\theta}_n$ as

$$\widehat{\theta}_n=\underset{\theta\in\Theta}{\operatorname{argmin}}\left[\frac{1}{n}\sum_{i=1}^{n}g\left(Z_i,\theta,\widehat{h}_n\right)\right]'\widetilde{W}_n\left[\frac{1}{n}\sum_{i=1}^{n}g\left(Z_i,\theta,\widehat{h}_n\right)\right]. \tag{50}$$

Theorem 3.1 implies that $\widetilde{W}_n=V_1^{-1}+o_p(1)$. Thus $\widehat{\theta}_n$ is a semiparametric two-step optimally weighted GMM with its asymptotic variance given by $V_\theta^o=\left(\Gamma_1'V_1^{-1}\Gamma_1\right)^{-1}$.

Our J test statistics is based on $\widehat{\theta}_n$. Although we can in principle work with the weight matrix $\widetilde{W}_n$ above, we recommend using a centered version of the weight matrix defined by

$$\widehat{W}_{c,n}=\left(\widehat{V}_{c,1,n}\right)^{-1}=\left(\sum_{i=-n+1}^{n-1}\mathcal{K}\left(\frac{i}{M_n}\right)\overline{\Upsilon}_{n,i}(\widehat{\alpha}_n)\left[\widehat{\mathbf{v}}_n^{*},\widehat{\mathbf{v}}_n^{*}\right]\right)^{-1}.$$

Based on the argument in Hall (2000) for parametric moment restrictions $E\left[g\left(Z,\theta_o\right)\right]=0$, we conjecture that the J test based on the centered weight matrix $\widehat{W}_{c,n}$ might be more powerful in finite samples.

To summarize, our over-identification test statistic is

$$J_n=\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}g(Z_i,\widehat{\theta}_n,\widehat{h}_n)\right]'\widehat{W}_{c,n}\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}g(Z_i,\widehat{\theta}_n,\widehat{h}_n)\right]. \tag{51}$$

Proposition 3.2 *Let the conditions of Theorem 3.1 hold. Then: under the null of correct specification* $E[g(Z, \theta_o, h_o)] = 0$ *with* $d_g > d_\theta$,

$$J_n \rightarrow_d \chi^2_{d_g - d_\theta}.$$

The inference and specification test proposed in this section are based on the consistency of kernel LRV estimate, which is derived under the assumption that the bandwidth M_n diverges to infinity. However, inference procedures based on such large bandwidth asymptotics may suffer nontrivial size distortion in finite samples, because the bandwidth M_n is always finite in empirical applications. We shall investigate alternative fixed-bandwidth asymptotics in the next section.

4 Inference Based on Orthonormal Series LRV Estimate

This section provides semiparametric inference and over-identifying specification test based on an orthonormal series LRV estimate.

4.1 Series LRV Estimate and Robust F Test

The series LRV estimation method projects the process of interest onto some orthonormal basis functions and uses the average of the out-product of each projection coefficients as the LRV estimator. The series LRV estimate is convenient for the empirical implementation because it is easy to compute and is automatically positive definite in finite samples. For parametric models, Phillips (2005) established the consistency of the series LRV estimate when the number of the basis functions M_n diverges to infinity; Sun (2013) suggested that the Wald test based on the fixed-M asymptotic theory has more accurate size in finite sample than the test based on the increasing-M asymptotics. We adopt this approach to semiparametric two-step GMM framework.

Let $\{\phi_m\}_{m=1}^{\infty}$ be a sequence of orthonormal basis functions in $L_2[0,1]$ that satisfy $\int_0^1 \phi_m(r)dr = 0$.

We define the following orthonormal series projection

$$\hat{\Lambda}_m = \frac{(\widehat{\Gamma}_1' W_n \widehat{\Gamma}_1)^{-1} \widehat{\Gamma}_1' W_n}{\sqrt{n}} \sum_{i=1}^{n} \phi_m \left(\frac{i}{n} \right) \widehat{S}_{i,n}^* \tag{52}$$

and construct the direct series estimator $\hat{\Lambda}_m \hat{\Lambda}_m'$ for each $m = 1, 2, ..., M$. Taking a simple average of these direct estimators yields the series estimator of V_θ

$$\widehat{V}_{R,n} = \frac{1}{M} \sum_{m=1}^{M} \hat{\Lambda}_m \hat{\Lambda}_m', \tag{53}$$

where M is the number of basis functions used and is the smoothing parameter in the present setting.

Let Ω_θ denote the square root matrix of V_θ, i.e. $\Omega_\theta^2 = V_\theta$. The following Lemma contains the key results to derive the asymptotic properties of $\widehat{V}_{R,n}$.

Lemma 4.1 *Let conditions of Theorem 2.1 and Lemma 2.1 hold. Suppose that Assumptions C.1 and C.2 in Appendix C hold. Then: for fixed finite* M*,*

$$\Omega_\theta^{-1} \hat{\Lambda}_m \to_d B_{d_\theta, m}(1),$$

where $\{B_{d_\theta,m}(1)\}_{m=1}^{M}$ *are* $d_\theta \times 1$ *independent standard Gaussian random vectors.*

Robust t- and F-tests By Lemma 4.1 and the Continuous Mapping Theorem (CMT), we can deduce that

$$\begin{aligned} \Omega_\theta^{-1} \widehat{V}_{R,n} \Omega_\theta^{-1} &= \Omega_\theta^{-1} \frac{1}{M} \sum_{m=1}^{M} \hat{\Lambda}_m \hat{\Lambda}_m' \Omega_\theta^{-1} \\ &\to_d \frac{1}{M} \sum_{m=1}^{M} B_{d_g,m}(1) B_{d_g,m}'(1) \equiv \frac{V_{R,\infty}}{M}. \end{aligned}$$

When θ_o is a scaler, the robust t statistic is

$$t_{R,n} \equiv \frac{\sqrt{n}(\widehat{\theta}_n - \theta_o)}{\sqrt{M \widehat{V}_{R,n}}} = \frac{\sqrt{n} \Omega_\theta^{-1} (\widehat{\theta}_n - \theta_o)}{\sqrt{M \Omega_\theta^{-1} \widehat{V}_{R,n} \Omega_\theta^{-1}}} \to_d \frac{B(1)}{V_{R,\infty}^{1/2}} \stackrel{d}{=} t(M), \tag{54}$$

where $t(M)$ denotes the Student-t distribution with M degree freedom.

When θ_o is a vector, then the robust F statistic is

$$
\begin{aligned}
F_{R,n} &\equiv \frac{n}{M}\frac{M-d_\theta+1}{d_\theta}(\widehat{\theta}_n-\theta_o)'\widehat{V}_{R,n}^{-1}(\widehat{\theta}_n-\theta_o) \\
&= \frac{M-d_\theta+1}{d_\theta}\left[\sqrt{n}(\widehat{\theta}_n-\theta_o)'\Omega_\theta^{-1}\right]\left(M\Omega_\theta^{-1}\widehat{V}_{R,n}\Omega_\theta^{-1}\right)^{-1}\left[\Omega_\theta^{-1}\sqrt{n}(\widehat{\theta}_n-\theta_o)\right] \\
&\to_d \ \frac{M-d_\theta+1}{d_\theta}B_{d_\theta}'(1)V_{R,\infty}^{-1}B_{d_\theta}(1) \overset{d}{=} \mathcal{F}(d_\theta, M-d_\theta+1),
\end{aligned} \tag{55}
$$

where $\mathcal{F}(a,b)$ denotes a F-distributed random variable with degree of freedom (a,b).

4.2 Robust Over-identification Test

We next construct the over-identification test statistic using the series LRV estimate. Let $\widetilde{\alpha}_n = (\widetilde{\theta}_n, \widehat{h}_n)$ and $\widetilde{S}_{i,n}^* = \widehat{S}_i(\widetilde{\alpha}_n)[\widetilde{\mathbf{v}}_n^*]$ be the same as defined in subsection 3.2. We define the following weight matrix

$$
\widehat{W}_{R,n} = \left[\frac{1}{nM}\sum_{m=1}^{M}\left(\sum_{i=1}^{n}\phi_m\left(\frac{i}{n}\right)\widetilde{S}_{i,n}^*\right)\left(\sum_{i=1}^{n}\phi_m\left(\frac{i}{n}\right)\widetilde{S}_{i,n}^{*\prime}\right)\right]^{-1}.
$$

The semiparametric two-step GMM estimate $\widehat{\theta}_{R,n}$ of θ_o based on $\widehat{W}_{R,n}$ is defined as

$$
\widehat{\theta}_{R,n} = \operatorname*{argmin}_{\theta\in\Theta}\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}g(Z_i,\theta,\widehat{h}_n)\right]'\widehat{W}_{R,n}\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}g(Z_i,\theta,\widehat{h}_n)\right]. \tag{56}
$$

Then the robust J test statistic is

$$
J_{R,n} = \left[n^{-\frac{1}{2}}\sum_{i=1}^{n}g(Z_i,\widehat{\theta}_{R,n},\widehat{h}_n)\right]'\widehat{W}_{R,n}\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}g(Z_i,\widehat{\theta}_{R,n},\widehat{h}_n)\right]. \tag{57}
$$

The following proposition extends theorem 1 in Sun and Kim (2012) for a parametric GMM model $E[g(Z,\theta_o)]=0$ to a semiparametric two-step GMM model $E[g(Z,\theta_o,h_o)]=0$ with unknown functions $h_o()$.

Proposition 4.1 *Let the conditions of Lemma 4.1 hold. Then under the null of correct specification $E[g(Z,\theta_o,h_o)]=0$ with $d_g > d_\theta$,*

$$
J_{R,n}^* \equiv \frac{M-(d_g-d_\theta)+1}{M(d_g-d_\theta)}J_{R,n} \to_d \mathcal{F}(d_g-d_\theta, M-(d_g-d_\theta)+1).
$$

5 Estimation of Asymptotic Variance - Practical Interpretation

In this section, we consider how the sieve semiparametric two-step GMM estimator $\widehat{\theta}_n$ and the two estimators of V_θ relate to what one obtains if the estimation problem is approached from a purely parametric perspective. This can be viewed as an extension of Ackerberg, Chen and Hahn (2012) to weakly dependent setting. To simplify the presentation we restrict to the case in which h_o is real-valued and is estimated by a sieve M estimation in the first step.[3] The exact finite sample numerical equivalence results hold when the first-step unknown function h_o is approximated via a linear sieve space $\mathcal{H}_{k(n)} = \{h(\cdot) = P_{k(n)}(\cdot)'\beta : \beta \in \mathbb{R}^{k(n)}\}$. That is, the first-step sieve M estimator $\widehat{h} = P_{k(n)}(\cdot)'\widehat{\beta}$ solves

$$\max_{\beta_1,\cdots,\beta_{k(n)}} \frac{1}{n}\sum_{i=1}^{n}\varphi\left(Z_i, p_1(\cdot)\beta_1 + \cdots + p_{k(n)}(\cdot)\beta_{k(n)}\right). \tag{58}$$

If a researcher believes that the unknown function h_o is indeed parametrically specified, i.e., $h_o(\cdot) = p_1(\cdot)\beta_{o,1} + \cdots + p_K(\cdot)\beta_{o,K}$, and if $K = k(n)$, then the parametric two-step GMM estimator of θ_o starting with (58) will be identical to the sieve semiparametric two-step GMM estimator $\widehat{\theta}_n$ in (4). This means that for the purpose of computing $\widehat{\theta}_n$, it is harmless to "pretend" that the h_o is parametrically specified. We now show that the same idea holds for the estimated variance.

5.1 Asymptotics Based on Parametric Belief

We will assume that an applied researcher believes that $h_o(\cdot) = P_K'(\cdot)\beta_{o,P}$, and estimates $\beta_{o,P}$ by the maximizer $\widehat{\beta}_{n,P}$ of the following M estimation

$$\widehat{\beta}_{n,P} = \underset{\beta\in\mathcal{B}_P}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n}\varphi_P(Z_i, \beta) \tag{59}$$

[3] The numerical equivalence results for semiparametric two-step GMM when h_o is vector-valued or/and is estimated via sieve MD can be established very similarly, but at the expense of tedious notation and hence is omitted.

where $\mathcal{B}_P$ is some compact parameter space in $\mathbb{R}^K$ and $\varphi_P(Z,\beta) \equiv \varphi(Z, P_K'(\cdot)\beta)$. In the second step, he goes on to estimate θ_o by the following GMM procedure

$$\widehat{\theta}_{n,P} = \underset{\theta\in\Theta}{\text{argmin}} \left[n^{-\frac{1}{2}}\sum_{i=1}^{n} g_P(Z_i,\theta,\widehat{\beta}_{n,P})\right]' W_n \left[n^{-\frac{1}{2}}\sum_{i=1}^{n} g_P(Z_i,\theta,\widehat{\beta}_{n,P})\right], \tag{60}$$

where $g_P(Z,\theta,\beta) \equiv g(Z,\theta,P_K'(\cdot)\beta)$.

Let $R_{o,P} = -E\left[\frac{\partial^2\varphi_P(Z_i,\beta_{o,P})}{\partial\beta\partial\beta'}\right]$ be a nonsingular $K\times K$ matrix, $\Gamma_{1,P} = \frac{\partial E[g_P(Z,\theta_o,\beta_{o,P})]}{\partial\theta'}$ be a $d_g \times d_\theta$ matrix, and $\Gamma_{2,P} = \frac{E[\partial g_P(Z,\theta_o,\beta_{o,P})]}{\partial\beta'}$ be a $d_g\times K$ matrix. The applied researcher would then derive the standard asymptotic properties of the parametric two-step estimator $\widehat{\theta}_{n,P}$, which is summarized in the following proposition.

Proposition 5.1 *Under standard regularity conditions for parametric two-step GMM estimation such as in Newey and McFadden (1994), we have: the estimator $\widehat{\theta}_{n,P}$ defined in (60) satisfies $\sqrt{n}\left(\widehat{\theta}_{n,P}-\theta_o\right) \to_d \mathcal{N}(0,V_{\theta,P})$, where*

$$V_{\theta,P} = \left(\Gamma_{1,P}'W\Gamma_{1,P}\right)^{-1}\Gamma_{1,P}'WV_{1,P}W\Gamma_{1,P}\left(\Gamma_{1,P}'W\Gamma_{1,P}\right)^{-1} \tag{61}$$

and

$$V_{1,P} = \lim_{n\to\infty} Var\left\{n^{-\frac{1}{2}}\sum_{i=1}^{n}\left[g_P(Z_i,\theta_o,\beta_{o,P}) + \Gamma_{2,P}(R_{o,P})^{-1}\frac{\partial\varphi_P(Z_i,\beta_{o,P})}{\partial\beta}\right]\right\}. \tag{62}$$

It is clear that the semiparametric asymptotic variance V_θ in (27), which was derived under the nonparametric specification, is typically different from the parametric asymptotic variance $V_{\theta,P}$ in (61), which was derived under the parametric specification of h. The difference is easily understood because $\Gamma_{1,P} \neq \Gamma_1$ and $V_{1,P}\neq V_1$.

5.2 Numerical Equivalence

Using the parametric two-step GMM estimate, we can compute $\widehat{\Gamma}_{1,P} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial g_P(Z_i,\widehat{\theta}_{n,P},\widehat{\beta}_{n,P})}{\partial\theta'}$ as a consistent estimator of $\Gamma_{1,P}$. Define

$$\widehat{S}_{i,P,n} = g_P\left(Z_i,\widehat{\theta}_{n,P},\widehat{\beta}_{n,P}\right) + \widehat{\Gamma}_{2,P,n}\left(\widehat{R}_{n,P}\right)^{-}\frac{\partial\varphi_P\left(Z_i,\widehat{\beta}_{n,P}\right)}{\partial\beta}, \tag{63}$$

where $\widehat{\Gamma}_{2,P,n} \equiv \frac{1}{n}\sum_{i=1}^{n}\frac{\partial g_P(Z_i,\widehat{\theta}_{n,P},\widehat{\beta}_{n,P})}{\partial\beta'}$ and $\widehat{R}_{n,P} \equiv -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\varphi_P(Z_i,\widehat{\beta}_{n,P})}{\partial\beta\partial\beta'}$ are standard sample analog estimators of $\Gamma_{2,P}$ and $R_{o,P}$.[4] It turns out that when $K = k(n)$, $\widehat{S}_{i,P,n}$ in (63) is numerically equivalent to $\widehat{S}^*_{i,n}$ in (32) (see Appendix D for details). This implies the numerical equivalence results below.

Kernel LRV estimation case A typical consistent kernel estimator of $V_{1,P}$ will be

$$\widehat{V}_{1,P} = \sum_{i=-n+1}^{n-1}\mathcal{K}\left(\frac{i}{M_n}\right)\widehat{\Upsilon}_{n,P}(i),$$

where

$$\widehat{\Upsilon}_{n,P}(i) = \begin{cases} \frac{1}{n}\sum_{l=i+1}^{n}\widehat{S}_{l,P,n}\widehat{S}'_{l-i,P,n} & \text{for } i \geq 0 \\ \frac{1}{n}\sum_{l=-i+1}^{n}\widehat{S}_{l,P,n}\widehat{S}'_{l+i,P,n} & \text{for } i < 0 \end{cases}.$$

Then

$$\widehat{V}_{\theta,P,n} = \left(\widehat{\Gamma}'_{1,P}W_n\widehat{\Gamma}_{1,P}\right)^{-1}\widehat{\Gamma}'_{1,P}W_n\widehat{V}_{1,P}W_n\widehat{\Gamma}_{1,P}\left(\widehat{\Gamma}'_{1,P}W_n\widehat{\Gamma}_{1,P}\right)^{-1}$$

is a consistent kernel estimator of $V_{\theta,P}$ under the parametric specification.

Theorem 5.2 *Suppose that the parametric specification sets $h(\cdot) = p_1(\cdot)\beta_1 + \cdots + p_K(\cdot)\beta_K$ with $K = k(n)$, the sieve dimension of the linear sieve space $\mathcal{H}_{k(n)}$. Then: $\widehat{V}_{\theta,P,n} = \widehat{V}_{\theta,n}$ for all n, where $\widehat{V}_{\theta,n}$ is defined in (45).*

Orthonormal series LRV estimation case The numerical equivalence also applies to the series LRV estimate. For the parametric specification (59), we construct $\hat{\Lambda}_{m,P}$ as

$$\hat{\Lambda}_{m,P} = \frac{(\widehat{\Gamma}'_{1,P}W_n\widehat{\Gamma}_{1,P})^{-1}\widehat{\Gamma}'_{1,P}W_n}{\sqrt{n}}\sum_{i=1}^{n}\phi_m\left(\frac{i}{n}\right)\widehat{S}_{i,P,n}$$

[4]For the ease of notation, we prove the numerical equivalence of the variance estimates assuming that the moment functions and the criterion function in the first step M estimation are smooth. The main result will not change in the scenario where the moment and/or criterion functions are non-smooth.

and an orthonormal series LRV estimate $\widehat{V}_{R,P,n}$ of $V_{\theta,P}$ as

$$\widehat{V}_{R,P,n} = \frac{1}{M}\sum_{m=1}^{M}\hat{\Lambda}_{m,P}\hat{\Lambda}'_{m,P}.$$

If $K = k(n)$ (say chosen by AIC), then $\widehat{V}_{R,P,n} = \widehat{V}_{R,n}$ for all n, where $\widehat{V}_{R,n}$ is defined in (53).

Although it is incorrect from a theoretical prospective, the numerical equivalence result provides a computationally practical justification of the parametric belief in the first-step.

6 Extension to GMM with Non-smooth Moment Functions

In this section, we provide consistent estimates of average derivatives of possibly non-smooth moment functions.[5] Our estimates are based on random perturbation of the moment functions. To fix the idea, let η_θ be some $d_\theta \times 1$ random vector with mean zero and variance I_{d_θ} that is independent of the data. Let $\mu_n[f] = n^{-1}\sum_{i=1}^{n}\{f(Z_i) - E[f(Z_i)]\}$ denote the empirical process indexed by f. Then we define

$$\begin{aligned} D_{n,\theta}(\eta_\theta, \widehat{\alpha}_n) &\equiv n^{-\frac{1}{2}}\sum_{i=1}^{n} g(Z_i, \widehat{\theta}_n + n^{-\frac{1}{2}}\eta_\theta, \widehat{h}_n) - n^{-\frac{1}{2}}\sum_{i=1}^{n} g(Z_i, \widehat{\theta}_n, \widehat{h}_n) \\ &= \sqrt{n}\mu_n\left[g(Z_i, \widehat{\theta}_n + n^{-\frac{1}{2}}\eta_\theta, \widehat{h}_n) - g(Z_i, \widehat{\theta}_n, \widehat{h}_n)\right] + \Gamma_1(\widetilde{\theta}_n, \widehat{h}_n)\eta_\theta, \end{aligned} \tag{64}$$

where $\Gamma_1(\widetilde{\theta}_n, \widehat{h}_n) = \left[\Gamma'_{1,1}(\widetilde{\theta}_{1,n}, \widehat{h}_n), ..., \Gamma'_{1,d_g}(\widetilde{\theta}_{d_g,n}, \widehat{h}_n)\right]'$, and $\widetilde{\theta}_{j,n}$ $(j = 1, ..., d_\theta)$ are some values between $\widehat{\theta}_n$ and $\widehat{\theta}_n + n^{-\frac{1}{2}}\eta_\theta$. By equation (64) and the stochastic equicontinuity, we obtain:

$$D_{n,\theta}(\eta_\theta, \widehat{\alpha}_n) = \Gamma_1(\widetilde{\theta}_n, \widehat{h}_n)\eta_\theta + o_p(1). \tag{65}$$

By the continuity of the function $\Gamma_1(\cdot)$ in the local neighborhood of α_o, we know that $\Gamma_1(\widetilde{\theta}_n, \widehat{h}_n)$ is a consistent estimate of Γ_1. Motivated by the expression (65), we propose the following resampling procedures to estimate Γ_1:

[5] Examples of GMM estimation with non-smooth moment conditions can be found in Pakes and Pollard (1989) and CLvK.

1. from some known multivariate distribution with mean zero and variance I_{d_θ}, we independently generate B realization $\eta_{\theta,b}$ $(b = 1, ..., B)$;

2. for each realization $\eta_{\theta,b}$ $(b = 1, ..., B)$, calculate $D_{n,\theta}(\eta_{\theta,b}, \widehat{\alpha}_n)$;

3. Γ_1 is then estimated by $\widehat{\Gamma}_{1,B}$ with

$$\widehat{\Gamma}_{1,B} = \frac{1}{B} D'_{n,\theta,B} \underline{\eta_{\theta,B}} = \frac{1}{B} \sum_{b=1}^{B} D_{n,\theta}(\eta_{\theta,b}, \widehat{\alpha}_n)\eta'_{\theta,b} \tag{66}$$

where $\underline{\eta_{\theta,B}} = (\eta_{\theta,1}, ..., \eta_{\theta,B})'$ and $D_{n,\theta,B} = (D_{n,\theta}(\eta_{\theta,1}, \widehat{\alpha}_n), ..., D_{n,\theta}(\eta_{\theta,B}, \widehat{\alpha}_n))'$.

Let $E_\eta \left[D_{n,\theta}(\eta_\theta, \widehat{\alpha}_n)\eta'_\theta\right]$ denote the expectation of $D_{n,\theta}(\eta_\theta, \widehat{\alpha}_n)\eta'_\theta$ with respect to the random vector η_θ, then the consistency of the resampling estimate $\widehat{\Gamma}_{1,B}$ is ensured by the following lemma.

Lemma 6.1 *Let Assumption E.1 in Appendix E hold. Then:* $E_\eta \left[D_{n,\theta}(\eta_\theta, \widehat{\alpha}_n)\eta'_\theta\right] \to_p \Gamma_1$.

When the moment functions $g(Z, \theta, h)$ are pathwise differentiable in h_o, the pathwise derivative $\Gamma_2(\theta_o, h_o)[v]$ can be estimated by

$$\Gamma_{2,n}(\widehat{\theta}_n, \widehat{h}_n)[v] \equiv \frac{1}{n} \sum_{i=1}^{n} \left. \frac{\partial g\left[Z_i, \widehat{\theta}_n, \widehat{h}_n(\cdot) + \tau v(\cdot)\right]}{\partial \tau} \right|_{\tau=0} \quad \text{for any } v \in \mathcal{V}_{k(n)}. \tag{67}$$

When $g(Z, \theta, h)$ are not pathwise differentiable in h_o, we next show that the above resampling technique can be applied to estimate $\Gamma_2(\theta_o, h_o)[v]$. In the following we let $h_o()$ be a d_h-vector valued function.

1. from some known multivariate distribution with mean zero and variance I_{d_h}, we independently generate B realization $\eta_{h,b}$ $(b = 1, ..., B)$;

2. for each realization $\eta_{h,b}$ $(b = 1, ..., B)$, calculate

$$D_{n,h}(\eta_{h,b}, \widehat{\alpha}_n, v) = n^{-\frac{1}{2}} \sum_{i=1}^{n} g(Z_i, \widehat{\theta}_n, \widehat{h}_n + n^{-\frac{1}{2}}\eta_{h,b}v) - n^{-\frac{1}{2}} \sum_{i=1}^{n} g(Z_i, \widehat{\theta}_n, \widehat{h}_n)$$

3. $\Gamma_2(\theta_o, h_o)[v]$ is then estimated by $\widehat{\Gamma}_{2,B}(\widehat{\theta}_n, \widehat{h}_n)[v]$ with

$$\widehat{\Gamma}_{2,B}(\widehat{\theta}_n, \widehat{h}_n)[v] = \frac{1}{B} D'_{n,h,B} \underline{\eta_{h,B}} = \frac{1}{B} \sum_{b=1}^{B} D_{n,h}(\eta_{h,b}, \widehat{\alpha}_n, v) \eta'_{h,b} \tag{68}$$

where $\underline{\eta_{h,B}} = (\eta_{h,1}, ..., \eta_{h,B})'$ and $D_{n,h,B} = [D_{n,h}(\eta_{h,1}, \widehat{\alpha}_n, v), ..., D_{n,h}(\eta_{h,B}, \widehat{\alpha}_n, v)]'$.

The consistency of the resampling estimate $\widehat{\Gamma}_{2,B}(\widehat{\theta}_n, \widehat{h}_n)[\cdot]$ is ensured by the following lemma.

Lemma 6.2 *Let Assumption E.2 in Appendix E hold. Then: the resampling estimate* $\widehat{\Gamma}_{2,B}(\widehat{\theta}_n, \widehat{h}_n)[\cdot]$ *satisfies Assumption A.3.(ii) in Appendix A.*

The resampling estimate defined in (68) can be used to construct LRV estimators proposed in Sections 3 and 4. As an illustration, we consider the simplified example that all moment functions are non-smooth. Let $P_{k(n)}(x) = \left[p_1(x), \cdots, p_{k(n)}(x)\right]'$ be the set of basis functions used to approximate $h_o(x)$. Using the expression in (68), we define a $k(n) \times 1$ vector $\widehat{\Gamma}_{2,B}(\widehat{\theta}_n, \widehat{h}_n)[P_{k(n)}]$ whose j-th entry is defined as $\widehat{\Gamma}_{2,B}(\widehat{\theta}_n, \widehat{h}_n)[p_j]$. Using the expression in (37), we write the empirical Riesz representer as

$$\widehat{\mathbf{v}}_n^*(x) = \widehat{\Gamma}_{2,B}(\widehat{\theta}_n, \widehat{h}_n)[P'_{k(n)}] \left(\widehat{R}_{k(n)}\right)^- P_{k(n)}(x) \tag{69}$$

which together with expressions in (32), (40) and (53) can be used to construct the kernel based and series based LRV estimators.

7 Simulation Studies

This section conducts simulation experiments to investigate the finite sample performances of the inference methods proposed in Sections 3 and 4. We use the following model to simulated data:

$$Y_{1,i} = h_o(X_{1,i}) + u_i, \quad E\left[u_i | X_{1,i}\right] = 0, \tag{70}$$

$$Y_{2,i} = \sum_{j=1}^{4} X_{2,j,i}\theta_{j,o} + h_o(X_{1,i}) + v_i, \tag{71}$$

where $X_{1,i}$, $X_{2,j,i}$ $(j = 1, \ldots 4)$, u_i and v_i are scalar random variables, $h_o(x_1) = x_1^2 \log^2(1 + x_1) + \exp(x_1)$, and $(\theta_{1,o}, \theta_{2,o}, \theta_{3,o}, \theta_{4,o})' = (\theta_o, \theta_o, \theta_o, \theta_o)'$ where θ_o is a real scaler.[6]

To generate the simulated sample, we first generate a set of i.i.d. random vector $(\varepsilon_{1,i}, \ldots, \varepsilon_{7,i})$ from standard multivariate normal distribution $\mathcal{N}(0, I_7)$, where I_7 denotes the 7×7 identity matrix. The error terms u_i and v_i are generated from the first order auto-regressive (AR(1)) model:

$$
\begin{aligned}
u_i &= \rho u_{i-1} + \sqrt{1-\rho^2}\varepsilon_{6,i}, \\
v_i &= \rho v_{i-1} + \sqrt{1-\rho^2}\varepsilon_{7,i},
\end{aligned}
$$

where $u_0 = 0$, $v_0 = 0$ and different values of ρ are considered in this simulation study. To get the regressors $X_{1,i}$ and $X_{2,i}$, we generate 5 random variables $(e_{1,i}, \ldots, e_{5,i})$ using the AR(1) model:

$$
e_{j,i} = \rho e_{j,i-1} + \sqrt{1-\rho^2}\varepsilon_{j,i}, \text{ for } j = 1, \ldots, 5
$$

where $e_{j,0} = 0$ for all j. Using the vector $(e_{1,i}, \ldots, e_{5,i})$, we also generate

$$
e_{6,i} = \frac{e_{1,i} + e_{2,i} + e_{3,i} + e_{4,i}}{2\sqrt{2}} + \frac{e_{5,i}}{\sqrt{2}}.
$$

Given the latent random variables $(e_{1,i}, \ldots, e_{6,i})$, we generate $X_{1,i}$ and $X_{2,j,i}$ $(j = 1, \ldots, 4)$:

$$
\begin{aligned}
X_{1,i} &= \frac{\exp(e_{6,i})}{1 + \exp(e_{6,i})}, \ (X_{2,1,i}, X_{2,2,i}) = (e_{1,i}, e_{2,i}), \\
X_{2,3,i} &= \frac{e_{1,i}e_{4,i} + e_{2,i}e_{4,i} + e_{3,i} + v_i}{4}, \\
X_{2,4,i} &= \frac{e_{1,i}e_{3,i} + e_{2,i}e_{3,i} + e_{4,i} + v_i}{4}.
\end{aligned}
$$

From the data generating mechanism, we see that $X_{2,1,i}$ and $X_{2,2,i}$ are exogenous variables in that they are independent of v_i, while $X_{2,3,i}$ and $X_{2,4,i}$ are endogenous variables. We will assume that four IVs $(R_{1,i}, R_{2,i}, R_{3,i}, R_{4,i})' = R_i$ are available for the empirical researcher, where

$$
R_{1,i} = e_{3,i} + \eta v_i, \ R_{2,i} = e_{1,i}e_{4,i}, \ R_{3,i} = e_{4,i} + \eta v_i \text{ and } R_{4,i} = e_{1,i}e_{3,i}
$$

[6] We take $\theta_{1,o} = \theta_{2,o} = \theta_{3,o} = \theta_{4,o} = \theta_o$ in the DGP to simplify the Monte Carlo analysis of the size and power properties of the proposed tests. $(\theta_{1,o}, \theta_{2,o}, \theta_{3,o}, \theta_{4,o})$ are estimated without imposing these equality restrictions.

where η is a real scaler. The IVs, $R_{1,i}$ and $R_{3,i}$, are valid when $\eta = 0$ and invalid otherwise. Using the IVs, we construct a vector of moment functions $g(Z_i, h, \theta)$:

$$g(Z_i, h, \theta) = \left[Y_{2,i} - X'_{2,i}\theta - h(X_{1,i})\right] \begin{pmatrix} X_{2,1,i} \\ X_{2,2,i} \\ R_i \end{pmatrix},$$

where $X_{2,i} = (X_{2,1,i}, \ldots, X_{2,4,i})'$. Then there are 6 moment conditions for identification and estimation of θ_o. As the moment functions in $g(Z_i, \widehat{h}_n, \theta)$ are linear in θ, the GMM estimator of θ_o has closed form expression.

Two hypotheses will be tested in the simulation study. The first one is the joint hypothesis

$$H_0 : (\theta_{1,o}, \ldots, \theta_{4,o}) = 0 \text{ v.s. } H_1 : (\theta_{1,o}, \ldots, \theta_{4,o}) \neq 0, \tag{72}$$

and the second one is the over-identification test of the moment validity

$$H_0 : E\left[g(Z, h_o, \theta_o)\right] = 0 \text{ v.s. } H_1 : E\left[g(Z, h_o, \theta_o)\right] \neq 0. \tag{73}$$

We consider different values of θ_o and η to investigate both the size and power of the proposed tests. For the joint test (72), we set $\eta = 0$ and $\theta_o = 0.05l$ for $l = 0, 1, \ldots, 10$. While for the over-identification test (73), we set $\theta_o = 0$ and $\eta = 0.05m$ for $m = 0, 1, \ldots, 20$. For each combination (θ_o, η), we will let $\rho \in \{0, 0.25, 0.5, 0.75\}$ in the simulation. We consider different values of ρ to check the performances of our inference methods in scenarios with different data dependence, e.g., zero dependence when $\rho = 0$, weak dependence when $\rho = 0.25$ and strong dependence when $\rho = 0.75$.

Let $P_{k(n)}(\cdot) = [p_1(\cdot), \ldots, p_{k(n)}(\cdot)]'$ be a vector of functions, where $\{p_j(\cdot) : j \geq 1\}$ is a set of basis functions. Given the data on $(Y_{1,i}, X_{1,i})$, we compute the first-step sieve LS estimator of $h_o(x_1)$ as

$$\widehat{h}_n(x_1) = P'_{k(n)}(x_1) \left(\mathbf{P}_n \mathbf{P}'_n\right)^{-1} \mathbf{P}_n \mathbf{Y}_{1,n} \tag{74}$$

where $\mathbf{P}_n = [P_{k(n)}(X_{1,1}), \ldots, P_{k(n)}(X_{1,n})]$ and $\mathbf{Y}_{1,n} = [Y_{1,1}, \ldots, Y_{1,n}]'$. We use the trigonometric polynomials as the basis functions. The sieve dimension $k(n)$ is determined by AIC in the simulation.[7] The first-step estimator (74) is then used in the computation of the second-step GMM estimator of θ_o, which is

$$\widehat{\theta}_n = \left[\mathbf{X}_{2,n}'\mathbf{D}_n W_n \mathbf{D}_n' \mathbf{X}_{2,n}\right]^{-1} \mathbf{X}_{2,n}'\mathbf{D}_n W_n \mathbf{D}_n'(\mathbf{Y}_{2,n} - \mathbf{H}_{1,n}) \tag{75}$$

where

$$\begin{aligned}
\mathbf{Y}_{2,n} &= [Y_{2,1}, \ldots, Y_{2,n}]', \ \mathbf{H}_{1,n} = [\widehat{h}_n(X_{1,1}), \ldots, \widehat{h}_n(X_{1,n})]', \\
\mathbf{D}_n &= [\mathbf{X}_{2,1,n}, \mathbf{X}_{2,2,n}, \mathbf{R}_{1,n}, \mathbf{R}_{2,n}, \mathbf{R}_{3,n}, \mathbf{R}_{4,n}], \\
\mathbf{X}_{2,n} &= [\mathbf{X}_{2,1,n}, \mathbf{X}_{2,2,n}, \mathbf{X}_{2,1,n}, \mathbf{X}_{2,2,n}], \\
\mathbf{X}_{2,j,n} &= [X_{2,j,1}, \ldots, X_{2,j,n}]' \text{ and } \mathbf{R}_{j,n} = [R_{j,1}, \ldots, R_{j,n}]',
\end{aligned}$$

for $j = 1, \ldots, 4$. It is easy to see that $\widehat{\mathbf{v}}_n^*(x) = \mathbf{D}_n'\mathbf{P}_n' \left(\mathbf{P}_n\mathbf{P}_n'\right)^{-1} P_{k(n)}(x)$ and

$$\widehat{S}_{i,n}^* = g(Z_i, \widehat{h}_n, \widehat{\theta}_n) - \mathbf{D}_n'\mathbf{P}_n' \left(\mathbf{P}_n\mathbf{P}_n'\right)^{-1} P_{k(n)}(X_{1,i}) \left[Y_{1,i} - \widehat{h}_n(X_{1,i})\right]. \tag{76}$$

The finite sample performance of the series estimator $\widehat{h}_n$ is evaluated by its integrated mean squared error (IMSE), and the finite sample properties of the two-step GMM estimator $\widehat{\theta}_n$ are measured by its finite sample bias, variance and mean squared error (MSE). These information are summarized in Tables F.1 and F.2 of Appendix F. From these tables, we see that both the unknown function and the finite dimensional parameter are estimated well even the sample size is small, i.e. $n = 200$. Stronger data dependence makes the IMSE of the series LS estimator and the MSE of the two-step

[7] The AIC criterion is defined as

$$AIC_n(k) = \frac{1}{n}\sum_{i=1}^{n}(Y_{1,i} - \widehat{h}_k(X_{1,i}))^2 + \frac{2k}{n}$$

where k denotes the number of sieve functions used in constructing the sieve LS estimator $\widehat{h}_k(\cdot)$. For each simulated sample, we choose $k(n)$ by minimizing $AIC_n(k)$:

$$k(n) = \arg\min_{k\in\{1,2,\ldots,K_n\}} AIC_n(k)$$

where K_n is a predetermined upper bound. In the simulation studies, we set $K_{200} = 20$, $K_{500} = 30$ and $K_{1000} = 40$.

GMM estimator worse. When the sample size is increased from 200 to 1000, the IMSEs and MSEs are reduced significantly even for data with stronger dependence.

Inference based on consistent LRV estimation We first consider the inference based on consistent LRV estimators discussed in Section 3. Using the weight matrix $\mathbf{D}_n'\mathbf{D}_n/n$ and the expression in (75), we compute the initial GMM estimator $\widetilde{\theta}_n$ which is then used to calculate the empirical score in (76) and hence the LRV estimator. The Quadratic Spectral (QS) kernel is used for the LRV estimation.[8] The second step GMM estimator of θ_o is calculated using the inverse of the LRV estimator as weight matrix in (75). We use the test statistic C_n and the asymptotic theory in (48) to test the hypothesis in (72), and the test statistic J_n in (51) and Proposition 3.2 to test the hypothesis in (73). The empirical rejection probabilities of the tests based on C_n and J_n are presented in Table 7.1, Figures 7.1, 7.2, F.1 and F.2.

From Table 7.1, we see that the tests of the joint hypothesis (72) based on the Wald statistic suffer from non-trivial size distortion in small samples with stronger dependent data. The size property of the Wald test is improved when the sample size is increased. On the other hand, the size of the over-identification test based on the consistent LRV estimator is more accurate, and improves with the sample size.

To have an overall assessment on the finite sample performances of the tests based on C_n and J_n, we investigate their empirical power functions for $\theta_o \in [0, 0.5]$ and $\eta \in [0, 1]$ respectively. Without loss of generality, we only consider the cases that $\rho = 0.25$ or 0.5, and the nominal size $\alpha = 0.05$ or 0.1. The empirical power functions of the Wald tests with sample sizes $n = 200$ and $n = 500$ are depicted in Figures 7.1 and F.1 respectively. It is clear that when θ_o approaches to 0.5, the power of the Wald test converges to 1. In Figure 7.1, we see that the size distortion of the Wald test in small

[8] We use the automatic bandwidth selection rule proposed in Andrews (1991) (i.e., equations (6.2) and (6.4) in Andrews (1991)) for the bandwidth determination.

Table 7.1. Empirical Null Rejection Probabilities for Joint Test and Over-identification Test

	$\rho = 0.00$		$\rho = 0.25$		$\rho = 0.50$		$\rho = 0.75$	
	W-Test	J-Test	W-Test	J-Test	W-Test	J-Test	W-Test	J-Test
				$n = 200$				
α=0.1000	.1640	.1040	.1831	.1109	.2450	.1280	.3843	.1653
α=0.0500	.1008	.0521	.1181	.0561	.1689	.0671	.2973	.0950
α=0.0250	.0637	.0256	.0764	.0291	.1183	.0361	.2331	.0543
α=0.0025	.0165	.0027	.0203	.0028	.0400	.0049	.1119	.0091
				$n = 500$				
α=0.1000	.1226	.0988	.1377	.1033	.1740	.1161	.2495	.1353
α=0.0500	.0685	.0495	.0782	.0512	.1069	.0604	.1717	.0731
α=0.0250	.0391	.0245	.0453	.0251	.0668	.0309	.1189	.0391
α=0.0025	.0070	.0023	.0085	.0026	.0157	.0034	.0404	.0048
				$n = 1000$				
α=0.1000	.1109	.0986	.1201	.1027	.1429	.1123	.1424	.1088
α=0.0500	.0579	.0486	.0664	.0518	.0816	.0570	.0809	.0556
α=0.0250	.0311	.0238	.0360	.0265	.0477	.0289	.0467	.0288
α=0.0025	.0047	.0022	.0052	.0025	.0084	.0028	.0086	.0031

Notes: 1. The simulation results are based on 100,000 replications; 2. in each simulated sample, 10 more observations are generated and the first 10 observations are dropped; 3. α denotes the nominal size of the test; 4. the W-Test refers to the test of the hypothesis in (70) using the test statistic C_n and the asymptotic theory in (46); 5. the J-test refers to the test of the hypothesis in (71) using the test statistic J_n in (49) and the asymptotic theory stated in Proposition 3.2.

sample contributes to its power. The sample size improves both the size and the power of the Wald test significantly. For example, when $n = 500$, the power of the Wald test is close to 1 even when θ_o is around 0.2. For the over-identification test, it is clear that when η approaches to 1, the power of the J-test converges to 1. Comparing the power functions of $\rho = 0.25$ with their counterparts in the case of $\rho = 0.5$, we see that the data dependence has nontrivial effect on the power of the tests. The improvement of large sample on the power of J-test is well illustrated in Figure F.2.

Inference based on fixed-bandwidth LRV estimation We next investigate the inference based on the series LRV estimator and fixed-bandwidth asymptotic theory discussed in Section 4. To compare the performances of different inference methods, we consider the test of the same hypotheses specified in (72) and (73). Throughout the simulation, we use $\phi_{2m-1}(x) = \sqrt{2}\cos(2m\pi x)$, $\phi_{2m}(x) =$

$\sqrt{2}\sin(2m\pi x), m = 1, \ldots, M$ as the orthonormal basis functions for the series LRV estimation. The empirical score in (76) together with the orthonormal basis functions are used to construct the series LRV estimators. For the hypothesis in (72), the statistic $F_{R,n}$ and its limiting distribution specified in (55) are used, where we use the weight matrix $\mathbf{S}_n'\mathbf{S}_n/n$ and the expression in (75) to construct the initial GMM estimator for θ_o. We use the statistic $J_{R,n}$ in (57) and its asymptotic distribution stated in Proposition 4.1 to test the hypothesis in (73). To evaluate the robustness of the inference based on the orthonormal series LRV estimator, we consider five different values of M (i.e. $M = 3$, 4, 5, 10 and 20). The empirical rejection probabilities of the tests based on $F_{R,n}$ and $J_{R,n}$ are presented in Tables 7.2 and F.3, Figures 7.1, 7.2, F.1 and F.2.

From Tables 7.2 and F.3, we see that the tests of the joint hypothesis (72) based on $F_{R,n}$ have more accurate size in finite samples, which is in sharp contrast with the tests based on the Wald statistic and the asymptotic Chi-square distribution. The size of the F test is affected by the data dependence. Specifically, the F test becomes slightly over rejecting when the data dependence is strong. The size of the F test is also slightly effected by the number of the orthonormal basis functions used in constructing the series LRV estimator. Moreover, with the growth of the sample size, the actual size of the F test becomes more and more close to the nominal size. On the other hand, the over-identification test based on series LRV estimator is not only size correct in finite samples, but also robust to the strength of the data dependence and the number M of the orthonormal basis functions. These simulation results shows the inference methods based on fixed-bandwidth LRV estimators has better size control than these based on consistent LRV estimators.

The empirical power functions of the tests based on $F_{R,n}$ and $J_{R,n}$ with $M = 3, 5, 10$ and 20 are depicted in Figures 7.1, 7.2, F.1 and F.2. For the F-test based on $F_{R,n}$, it is clear that its power approaches to 1 when θ_o converges to 0.5. Different choices of M lead to different power properties of the F-test. The power improvements are significant when M is increased from 3 to 5 and then from 5 to 10, while the improvement becomes small when M is increased from 10 to 20. On the other hand, increasing M leads to size distortion to the test, although the magnitude is small even when

M is increased from 3 to 20. Figures 7.1 and F.1 show that the Wald test is more powerful than the $F_{R,n}$ test for all the values of M we investigated. However, such a comparison is not fair because the Wald test has nontrivial over-rejection under the null, which contributes to its power. For the over-identification test based on $J_{R,n}$, we see that when η approaches to 1, its power converges to 1. Increasing the value of M from 3 to 5 or from 5 to 10 leads to nontrivial improvement of power with only small effect on the size. When M is increased to 20, the $J_{R,n}$ test becomes almost as powerful as the J_n test based on the consistent LRV estimator. When the sample size is increased, it is clear that the power of the test based on series LRV estimator is improved very quickly.

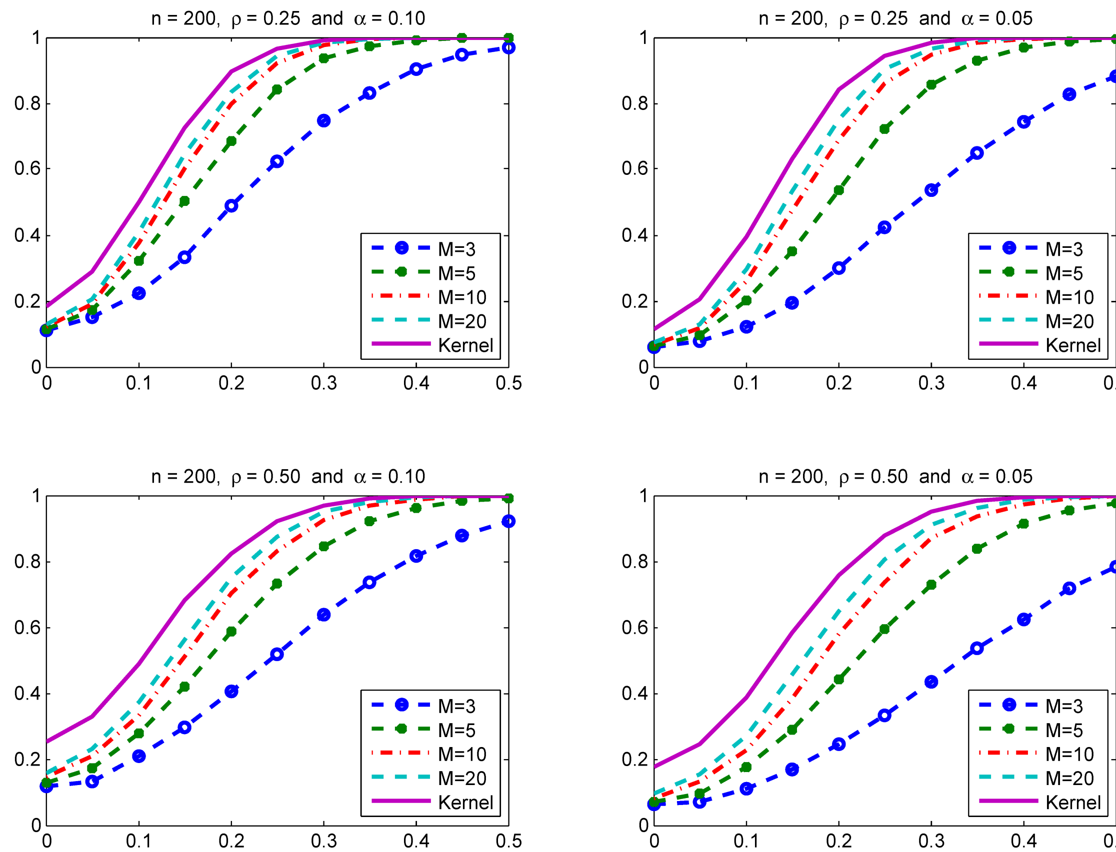
Comparison of two inference procedures From the above discussion, we see that the inference based on the fixed-bandwidth LRV estimator has good size control in all the scenarios we considered. The inference based on consistent LRV estimator has nontrivial size distortion in the joint test of (72), while its size is better in the over-identification test. On the other hand, the empirical power functions of the tests based on the consistent LRV estimator converge to 1 faster than these based on the fixed-bandwidth LRV estimator. The size comparison suggests that one could be more confidently reject the null hypothesis if it is rejected by both the tests based on the consistent and fixed-bandwidth LRV estimators. Otherwise, one should be very careful if the null is only rejected by the test based on the consistent LRV estimator. This is particularly important when the sample size is small and/or the data dependence is strong. Moreover, the power comparison leads to the interesting question of the optimal selection of the number of orthonormal basis functions or the bandwidth in LRV estimation. Sun, Phillips and Jin (2008) investigate this issue in the time series Gaussian location model. Generalizing their results to the semiparametric time series models is an important but challenging problem, which is beyond the scope of this paper.

Table 7.2. Empirical Null Rejection Probabilities for Joint Test and Over-identification Test

	$\rho = 0.00$		$\rho = 0.25$		$\rho = 0.50$		$\rho = 0.75$	
	F-Test	J-Test	F-Test	J-Test	F-Test	J-Test	F-Test	J-Test
	$n = 200$ and $M = 3$							
α=0.1000	.1076	.0952	.1079	.0952	.1175	.0966	.1369	.0965
α=0.0500	.0546	.0458	.0543	.0458	.0603	.0461	.0718	.0459
α=0.0250	.0276	.0223	.0278	.0222	.0298	.0222	.0365	.0219
α=0.0025	.0027	.0022	.0027	.0021	.0030	.0019	.0038	.0022
	$n = 200$ and $M = 4$							
α=0.1000	.1113	.0946	.1129	.0958	.1231	.0937	.1537	.0997
α=0.0500	.0583	.0460	.0588	.0461	.0650	.0447	.0863	.0475
α=0.0250	.0303	.0229	.0299	.0223	.0343	.0216	.0473	.0229
α=0.0025	.0033	.0021	.0032	.0020	.0038	.0018	.0060	.0018
	$n = 200$ and $M = 5$							
α=0.1000	.1134	.0951	.1157	.0951	.1292	.0971	.1672	.0988
α=0.0500	.0597	.0450	.0620	.0456	.0696	.0470	.0973	.0455
α=0.0250	.0316	.0215	.0325	.0217	.0376	.0223	.0557	.0208
α=0.0025	.0036	.0019	.0035	.0020	.0046	.0020	.0081	.0016
	$n = 500$ and $M = 3$							
α=0.1000	.1023	.0980	.1015	.0962	.1058	.0965	.1137	.0970
α=0.0500	.0519	.0480	.0501	.0476	.0541	.0476	.0579	.0470
α=0.0250	.0261	.0238	.0259	.0241	.0274	.0235	.0295	.0231
α=0.0025	.0024	.0021	.0022	.0026	.0027	.0021	.0030	.0023
	$n = 500$ and $M = 4$							
α=0.1000	.1034	.0959	.1046	.0974	.1112	.0973	.1221	.0965
α=0.0500	.0526	.0478	.0535	.0482	.0572	.0481	.0642	.0459
α=0.0250	.0261	.0234	.0274	.0238	.0298	.0238	.0339	.0223
α=0.0025	.0027	.0024	.0029	.0024	.0030	.0022	.0035	.0021
	$n = 500$ and $M = 5$							
α=0.1000	.1029	.0976	.1061	.0976	.1130	.0972	.1280	.0972
α=0.0500	.0525	.0473	.0531	.0473	.0585	.0473	.0688	.0469
α=0.0250	.0271	.0235	.0273	.0237	.0295	.0230	.0366	.0221
α=0.0025	.0030	.0021	.0030	.0020	.0032	.0022	.0045	.0018

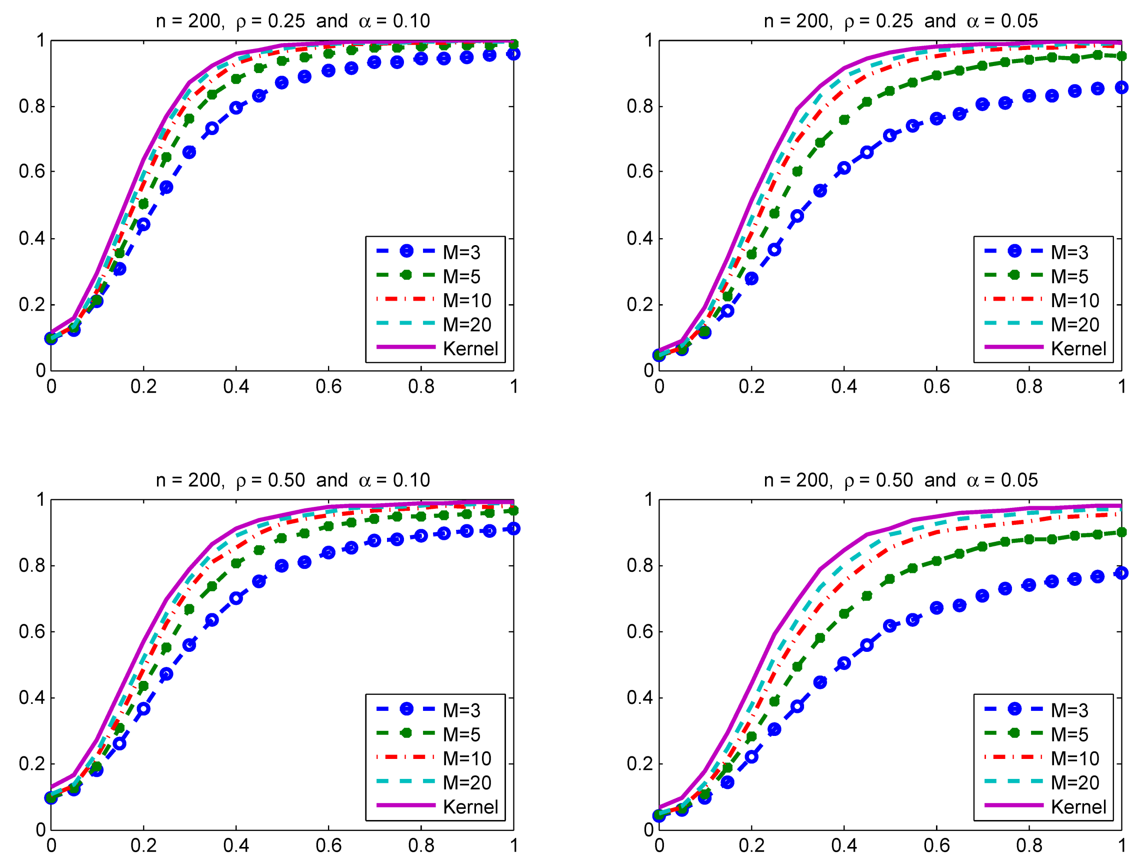
Notes: 1. The simulation results are based on 100,000 replications; 2. α denotes the nominal size of the test; 3. the F-Test refers to the test of the hypothesis in (70) using the test statistic $F_{R,n}$ and the asymptotic theory in (51); 4. the J-test refers to the test of the hypothesis in (71) using the test statistic $J_{R,n}$ in (48) and the asymptotic theory stated in Proposition 4.1.

Figure 7.1. Empirical Power Functions of the Tests of the Joint Hypothesis



Notes: 1. The simulation results are based on 10,000 replications; 2. α denotes the nominal size of the test; 3.the X-axis represents the value of θ_o and the Y-axis represents the rejection probability; 4. the curves denoted by "M=3", "M=5", "M=10" and "M=20" are the power functions of the joint tests based on the series LRV estimators with M=3, M=5, M=10 and M=20 respectively; 5. the curve denoted by "Kernel" is the power function of the Wald test based on kernel LRV estimator.

Figure 7.2. Empirical Power Functions of the Over-identification Tests



Notes: 1. The simulation results are based on 10,000 replications; 2. α denotes the nominal size of the test; 3.the X-axis represents the value of η and the Y-axis represents the rejection probability; 4. the curves denoted by "M=3", "M=5", "M=10" and "M=20" are the power functions of the over-identification test based on the series LRV estimators with M=3, M=5, M=10 and M=20 respectively; 5. the curve denoted by "Kernel" is the power function of the over-identification test based on kernel LRV estimator.

8 Conclusion

In this paper and for weakly dependent data, we first characterize the semiparametric asymptotic variance V_θ of a second-step GMM estimator $\widehat{\theta}_n$, where the unknown nuisance functions are estimated via sieve extremum estimation in the first step. We show that the asymptotic variance V_θ can be well approximated by sieve variances that have simple closed-form expressions. We then provide two different inference procedures for the semiparametric two-step GMM estimation of models with weakly dependent data. The first procedure is based on a kernel HAC estimate of the V_θ, and the corresponding Wald test and the over-identification test are asymptotically chi-square distributed under their respective null. The second procedure uses a robust orthonormal series estimate of the V_θ, and the corresponding Wald test and the over-identification test are asymptotically F distributed under their respective null. A new consistent random-perturbation estimator of the derivative of the expectation of the non-smooth moment function is provided. Finally, we show that the sieve two-step GMM estimation and inference could be implemented using standard softwares as if the first-step were "parametric".

Under Conditions (7) and (8) in Section 2, the condition that the linear functional $\Gamma_{2,j}(\alpha_o)[\cdot] : \mathcal{V} \to \mathbb{R}$ is bounded for all $j = 1, ..., d_g$ is necessary for root-n CAN of the second-step GMM estimator $\widehat{\theta}$. When $\Gamma_{2,j}(\alpha_o)[\cdot] : \mathcal{V} \to \mathbb{R}$ is unbounded for some j, $\Gamma_{2,j}(\alpha_o)[\cdot] : \mathcal{V}_{k(n)} \to \mathbb{R}$ is still bounded for all j for each finite $k(n)$. We could still establish that the sieve semiparametric two-step GMM estimator satisfies

$$\sqrt{n}(V_{\theta,n})^{-1/2}(\widehat{\theta}_n - \theta_o) \to_d \mathcal{N}\left(0, I_{d_\theta}\right), \quad \text{where}$$

$$V_{\theta,n} = \left(\Gamma_1' W \Gamma_1\right)^{-1} \left(\Gamma_1' W V_{1,n}^* W \Gamma_1\right) \left(\Gamma_1' W \Gamma_1\right)^{-1},$$

and $V_{1,n}^*$ is the sieve LRV defined in (26). Moreover, the consistent kernel LRV based inference and the robust orthonormal series LRV based inference results remain valid with the corresponding estimator of the sieve variance $V_{\theta,n}$. In fact, the diverging to infinity sieve variance $V_{\theta,n}$ makes Condition (8) easier to verify. Unfortunately, the results in Khan (2013), Chen, Liao and Sun

(2014), and Chen and Pouzo (2012) for sieve plug-in estimation of slower-than-root-n parameters suggest that the nice SBP of sieve estimators may no longer hold when the second-step GMM estimator $\widehat{\theta}$ converges to θ_o at a slower-than-root-n rate. We leave it to future work for carefully investigating such situations.

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] Ai, C. and X. Chen (2003): "Efficient Estimation of Conditional Moment Restrictions Models Containing Unknown Functions," *Econometrica,* 71, 1795-1843.

[3] Andrews, D. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817-854.

[4] Armstrong, T., M. Bertanha and H. Hong (2014): "A Fast Resample Method for Parametric and Semiparametric Models," *Journal of Econometrics*, 179, 128-133.

[5] Bansal, R. and S. Viswanathan (1993): "No Arbitrage and Arbitrage Pricing: A New Approach," *The Journal of Finance*, 48(4), 1231-1262.

[6] Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," Chp. 76 in *Handbook of Econometrics*, Vol. 6B, eds. J.J. Heckman and E.E. Leamer.

[7] Chen, X. (2013): "Penalized Sieve Estimation and Inference of Semi-nonparametric Dynamic Models: A Selective Review," in *Advances in Economics and Econometrics, 2010 World Congress of the Econometric Society book volumes*, Cambridge University Press.

[8] Chen, X., and S. Ludvigson (2009): "Land of Addicts? An Empirical Investigation of Habit-Based Asset Pricing Models," *Journal of Applied Econometrics*, 24, 1057-1093.

[9] Chen, X., J. Favilukis and S. Ludvigson (2013): "An Estimation of Economic Models with Recursive Preferences," *Quantitative Economics*, 4, 39-83.

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

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

[12] Chen, X. and D. Pouzo (2012): "Sieve Quasi Likelihood Ratio Inference on Semi/nonparametric Conditional Moment Models," *Cowles Foundation Discussion Paper*, no. 1897.

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

[14] Chui, C. (1971): "Concerning Rates of Convergence of Riemann Sums," *Journal of Approximation Theory*, 4, 279–287.

[15] Conley, T. and W. Dupor (2003): "A Spatial Analysis of Sectoral Complementarity," *Journal of Political Economy*, Vol. 111, 311-352.

[16] de Jong, R. (2000): "A Strong Consistency Proof for Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimators," *Econometric Theory*, 16, 262-268.

[17] Engle, R.F. (2010): "Long Term Skewness and Systemic Risk," Presidential Address SoFiE, 2009.

[18] Engle, R. and G. Gonzalez-Rivera (1991): "Semiparametric ARCH Models," *Journal of Business and Economic Statistics*, 9, 345-359.

[19] Engle, R., C. Granger, J. Rice and A. Weiss (1986): "Semiparametric Estimates of the Relation between Weather and Electricity Sales," *Journal of the American Statistical Association*, 81, 310-320.

[20] Engle, R.F and V. Ng (1993): "Measuring and Testing the Impact of News On Volatility," *Journal of Finance*, 48, 1749-1778.

[21] Engle, R.F. and J.G. Rangel (2008): "The Spline-GARCH Model for Unconditional Volatility and its Global Macroeconomic Causes," *Review of Financial Studies*, 21, 1187-1222.

[22] Gallant, A.R. and G. Tauchen (1989): "Semiparametric Estimation of Conditional Constrained Heterogenous Processes: Asset Pricing Applications," *Econometrica*, 57, 1091-1120.

[23] Gallant, A.R. and G. Tauchen (1996): "Which Moments to Match?" *Econometric Theory*, 12, 657-681.

[24] Gallant, A.R., L.P. Hansen, and G. Tauchen (1990): "Using Conditional Moments of Asset Payoffs To Infer the Volatility of Intertemporal Marginal Rates of Substitution," *Journal of Econometrics*, 45, 141-179.

[25] Gallant, A.R., D. Hsieh and G. Tauchen (1991): "On Fitting a Recalcitrant Series: The Pound/Dollar Exchange Rate, 1974-83," in Barnett, W.A., J. Powell and G. Tauchen (eds.), *Non-parametric and Semi-parametric Methods in Econometrics and Statistics,* 199-240, Cambridge: Cambridge University Press.

[26] Hall, A. (2000): "Covariance Matrix Estimation and the Power of the Over-identifying Restrictions Test," *Econometrica*, 68, 1517-1527.

[27] Hansen, B.E. (1992): "Consistent Covariance Matrix Estimation for Dependent Heterogeneous Processes," *Econometrica*, 60, 967-972.

[28] Hansen, L.P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029-1054.

[29] Kawai, K. (2011): "Auction Design and the Incentives to Invest: Evidence from Procurement Auctions," *Working Paper*, NYU Stern.

[30] Khan, S. (2013): "Distribution free estimation of heteroskedastic binary response models using Probit/Logit criterion functions," *Journal of Econometrics*, 172, 168-182.

[31] Lee, J. and P.M. Robinson (2013). Series estimation under cross-sectional dependence. Preprint, London School of Economics.

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

[33] Newey, W.K. and D. F. McFadden (1994): "Large sample estimation and hypothesis testing,", in R.F. Engle III and D.F. McFadden (eds.), *Handbook of Econometrics*, vol. 4. North-Holland, Amsterdam.

[34] Newey, W.K. and K.D. West (1987): "A Simple, Positive Semi-definite, Heteroscedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703-708.

[35] Pakes, A. and Pollard, D. (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027-1057.

[36] Phillips, P.C.B. (2005): "HAC Estimation by Automated Regression," *Econometric Theory*, 21, 116-142.

[37] Sun, Y. (2013): "A Heteroscedasticity and Autocorrelation Robust F Test Using an Orthonormal Series Variance Estimator," *Econometrics Journal*, 16, 1-26.

[38] Sun, Y. and Kim, M.S. (2012): "Simple and Powerful GMM Over-identification Tests With Accurate Size," *Journal of Econometrics*, 166(2), 267-281.

[39] Sun, Y., P.C.B. Phillips, and S. Jin (2008): "Optimal Bandwidth Selection in Heteroskedasticity Autocorrelation Robust Testing," *Econometrica*, 76(1), 175-194.

APPENDIX

A Proof of the Results in Section 2

Under typical conditions imposed on the first-step sieve extremum estimation of unknown functions $h_o()$, we obtain the consistency $||\widehat{h}-h_o||_{\mathcal{H}} = o_p(1)$ and the rate of convergence $||\widehat{h}-h_o|| = O_p(\delta_n)$ with the pseudo-metric $\|\cdot\|$ defined in (13), where $\delta_n = o_p(n^{-1/4})$ is some positive decreasing sequence. See, e.g., Chen and Shen (1998) for sieve M estimation with weakly dependent data, and Chen and Pouzo (2012) for sieve MD estimation with weakly dependent data.

Assumption A.1 *The first-step sieve extremum estimate* $\widehat{h}_n$ *satisfies:*
(i) $\max_{j=1,\ldots,d_g}\left|n^{-1}\sum_{i=1}^{n}\Delta(Z_i,h_o)[v^*_{j,k(n)}]-\langle v^*_{j,k(n)},\widehat{h}_n-h_o\rangle\right| = o_p(n^{-1/2})$;
(ii) (SBP) $||\widehat{h}_n-h_o||\times\max_{j=1,\ldots,d_g}||v^*_{j,k(n)}-v^*_j|| = o_p(n^{-1/2})$.

We note that under low level conditions, Assumption A.1 is satisfied by both sieve M estimation (Chen, Liao and Sun, 2014) and sieve MD estimation (Chen and Pouzo, 2012).

Let $\|\theta\|_E = \sqrt{\theta'\theta}$ and $\|A\|_W = \sqrt{tr\,(A'WA)}$ for any matrix A, where W is a symmetric, positive definite matrix. $G_n(\theta,h) = n^{-1}\sum_{i=1}^{n} g\,(Z_i,\theta,h)$ and $G(\theta,h) = E[g\,(Z_i,\theta,h)]$.

Assumption A.2 *Suppose that* $\theta_o\in int(\Theta)$ *satisfies* $G(\theta_o,h_o)=0$*, that* $\widehat{\theta}_n-\theta_o = o_p(1)$*,* $W_n-W = o_p(1)$*, and that (i)* $\Gamma_1(\theta,h_o)$ *exists in a neighborhood of* θ_o *and is continuous at* θ_o*,* $\Gamma_1'W\Gamma_1$ *is nonsingular; (ii) the pathwise derivative* $\Gamma_2(\theta,h_o)[h-h_o]$ *exists in all directions* $[h-h_o]$ *and satisfies*

$$\|\Gamma_2(\theta,h_o)[h-h_o]-\Gamma_2(\theta_o,h_o)[h-h_o]\|_W \le \|\theta-\theta_o\|_E\times o(1)$$

for all θ *with* $\|\theta-\theta_o\|_E = o(1)$ *and all* h *with* $\|h-h_o\|_{\mathcal{H}} = o(1)$*; either (iii)*

$$\left\|G(\theta,\widehat{h}_n)-G(\theta,h_o)-\Gamma_2(\theta,h_o)[\widehat{h}_n-h_o]\right\|_W = o_p(n^{-\frac{1}{2}})$$

for all θ *with* $\|\theta-\theta_o\|_E = o(1)$*; or (iii)' there are some constants* $c\ge 0$*,* $\epsilon_1>0, \epsilon_2>1$ *such that*

$$\|G(\theta,h)-G(\theta,h_o)-\Gamma_2(\theta,h_o)[h-h_o]\|_W \le c\,\|h-h_o\|_{\mathcal{H}}^{\epsilon_1}\,\|h-h_o\|^{\epsilon_2}$$

for all θ *with* $\|\theta-\theta_o\|_E = o(1)$*, all* h *with* $\|h-h_o\|_{\mathcal{H}} = o(1)$*,* $c||\widehat{h}_n-h_o||_{\mathcal{H}}^{\epsilon_1}||\widehat{h}_n-h_o||^{\epsilon_2} = o_p(n^{-\frac{1}{2}})$*;*
(iv) for all sequences of positive numbers $\{\kappa_n\}$ *with* $\kappa_n = o(1)$

$$\sup_{\|\theta-\theta_o\|<\kappa_n,\|h-h_o\|_{\mathcal{H}}<\kappa_n}\frac{\|G_n(\theta,h)-G(\theta,h)-G_n(\theta_o,h_o)\|_W}{n^{-1/2}+\|G_n(\theta,h)\|_W+\|G(\theta,h)\|_W} = o_p(1);$$

(v) $n^{-\frac{1}{2}}\sum_{i=1}^{n}\{g\,(Z_i,\theta_o,h_o)+\Delta(Z_i,h_o)\,[\mathbf{v}^*_n]\}\to_d\mathcal{N}(0,V_1)$*, where* V_1 *is defined in (28).*

Conditions (iv) and (v) are respectively implied by:

(iv)' for all sequences of positive numbers $\{\kappa_n\}$ *with* $\kappa_n = o(1)$,

$$\sup_{\|\theta-\theta_o\|<\kappa_n, \|h-h_o\|_{\mathcal{H}}<\kappa_n} \|G_n(\theta,h) - G(\theta,h) - G_n(\theta_o,h_o)\|_W = o_p(n^{-1/2});$$

(v)' $n^{-\frac{1}{2}} \sum_{i=1}^{n} \{g(Z_i,\theta_o,h_o) + \Delta(Z_i,h_o)[\mathbf{v}^*]\} \to_d \mathcal{N}(0,V_1)$ *and* $n^{-\frac{1}{2}} \sum_{i=1}^{n} \Delta(Z_i,h_o)[\mathbf{v}^* - \mathbf{v}_n^*] = o_p(1)$.

Assumption A.2 is basically conditions for Theorem 4.1 of Chen (2007), except that we relax Condition (4.1.4)' of Chen (2007) by (iii)'. Lemma 4.2 of Chen (2007) provides low level sufficient conditions for the stochastic equicontinuity condition (iv)' for possibly non-smooth moment with beta mixing data. Condition (v)' is similar to Conditions 4.4 and 4.5 in Chen (2007) for sieve M estimation. We note that Condition (iv) implies Condition (7), and Conditions (ii) + (iii) imply Condition (8). So one could establish the root-n CAN of $\widehat{\theta}_n$ to θ_o under weaker sets of conditions than Assumption A.2.

Proof of Theorem 2.1. Under Assumption A.2.(i)-(iv), we can follow the proof of Theorem 2 of CLvK or the proof of Theorem 4.1 of Chen (2007) to get

$$\widehat{\theta}_n - \theta_o = -\left(\Gamma_1' W \Gamma_1\right)^{-1} \Gamma_1' W \left[G_n(\theta_o,h_o) + \Gamma_2(\theta_o,h_o)[\widehat{h}_n - h_o]\right] + o_p(n^{-\frac{1}{2}}). \tag{A.1}$$

First, note that under the assumption $\max_{j=1,\ldots,d_g} \lim_{k(n)\to\infty} ||v_{j,k(n)}^*|| < \infty$, $\Gamma_{2,j}(\theta_o,h_o)[\cdot]$ ($j = 1,...,d_g$) is a bounded linear functional on $\mathcal{V}$. Hence using the Riesz representation theorem and Assumption A.1.(i)(ii), we have, for $j = 1,...,d_g$,

$$\begin{aligned}\Gamma_{2,j}(\theta_o,h_o)[\widehat{h}_n - h_o] &= \langle v_j^*, \widehat{h}_n - h_o\rangle = \langle v_{j,k(n)}^*, \widehat{h}_n - h_o\rangle + \langle v_j^* - v_{j,k(n)}^*, \widehat{h}_n - h_o\rangle \\ &= \langle v_{j,k(n)}^*, \widehat{h}_n - h_o\rangle + o_p(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \Delta(Z_i,h_o)[v_{j,k(n)}^*] + o_p(n^{-\frac{1}{2}}).\end{aligned} \tag{A.2}$$

Using equations (A.1) and (A.2), we get

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_n - \theta_o) &= \frac{-\left(\Gamma_1' W \Gamma_1\right)^{-1} \Gamma_1' W}{\sqrt{n}} \sum_{i=1}^{n} \{g(Z_i,\theta_o,h_o) + \Delta(Z_i,h_o)[\mathbf{v}_n^*]\} + o_p(1) \\ &\to_d \left(\Gamma_1' W \Gamma_1\right)^{-1} \Gamma_1' W \times \mathcal{N}(0,V_1) \stackrel{d}{=} \mathcal{N}(0,V_\theta),\end{aligned} \tag{A.3}$$

where the weak convergence is by Assumption A.2.(v). Alternatively, by Assumption A.2.(v)' and equation (A.2), we have:

$$\Gamma_{2,j}(\theta_o,h_o)[\widehat{h}_n - h_o] = \frac{1}{n}\sum_{i=1}^{n} \Delta(Z_i,h_o)[v_j^*] + o_p(n^{-\frac{1}{2}}),$$

which, together with equation (A.1) and Assumption A.2.(v)', implies

$$\sqrt{n}(\widehat{\theta}_n-\theta_o) = \frac{-\left(\Gamma_1'W\Gamma_1\right)^{-1}\Gamma_1'W}{\sqrt{n}}\sum_{i=1}^{n}\left\{g\left(Z_i,\theta_o,h_o\right)+\Delta(Z_i,h_o)\left[\mathbf{v}^*\right]\right\}+o_p(1)$$

$$\to_d \quad \left(\Gamma_1'W\Gamma_1\right)^{-1}\Gamma_1'W\times\mathcal{N}(0,V_1)\overset{d}{=}\mathcal{N}(0,V_\theta).$$

■

By the convergence rate of $(\widehat{\theta}_n,\widehat{h}_n)$, we define a local shrinking neighborhood of (θ_o,h_o) as

$$\mathcal{N}_n=\{(\theta,h)\in\Theta\times\mathcal{H}_{k(n)}:||\theta-\theta_o||_E\le n^{-1/2}\log\log n,||h-h_o||_{\mathcal{H}}\le\delta_{s,n},||h-h_o||\le\delta_n\}, \quad \text{(A.4)}$$

where $\delta_{s,n}=o(1)$, $\delta_n=o(n^{-\frac{1}{4}})$ such that $\widehat{\alpha}_n=(\widehat{\theta}_n,\widehat{h}_n)\in\mathcal{N}_n$ w.p.a.1.

Assumption A.3 *Let $\mathcal{W}_n=\{v\in\mathcal{V}_{k(n)}:\|v\|=1\}$ and $\delta_{w,n}=o(1)$ be a positive sequence.*
(i) $\sup_{v_1,v_2\in\mathcal{W}_n}|\langle v_1,v_2\rangle_n-\langle v_1,v_2\rangle|=O_p(\delta_{w,n})$;
(ii) $\sup_{\alpha\in\mathcal{N}_n,v\in\mathcal{W}_n}|\Gamma_{2,j,n}(\theta,h)[v]-\Gamma_{2,j}(\theta_o,h_o)[v]|=O_p(\delta_{w,n})$ *for all* $j=1,...,d_g$;
(iii) $\lim_{k(n)\to\infty}\min_{j=1,\ldots,d_g}||v^*_{j,k(n)}||>0$ *and* $\lim_{k(n)\to\infty}\max_{j=1,\ldots,d_g}||v^*_{j,k(n)}||<\infty$.

Assumption A.3 is mild and allows for any sieve extremum estimation in the first-step.

Proof of Lemma 2.1. Let $v_1=\widehat{v}^*_{j,k(n)}$ and $v_2=v$, we can invoke Assumption A.3.(i) to deduce that

$$\sup_{v\in\mathcal{V}_{k(n)}}\left|\frac{\left\langle\widehat{v}^*_{j,k(n)},v\right\rangle_n-\left\langle\widehat{v}^*_{j,k(n)},v\right\rangle}{||\widehat{v}^*_{j,k(n)}||\,\|v\|}\right|=O_p(\delta_{w,n}). \quad \text{(A.5)}$$

By the triangle inequality and Hölder inequality, we get

$$\begin{aligned}\frac{\left|||\widehat{v}^*_{j,k(n)}||^2-||v^*_{j,k(n)}||^2\right|}{||v^*_{j,k(n)}||^2} &\le \frac{\left|\left\langle\widehat{v}^*_{j,k(n)},\widehat{v}^*_{j,k(n)}\right\rangle-\left\langle\widehat{v}^*_{j,k(n)},v^*_{j,k(n)}\right\rangle\right|}{||v^*_{j,k(n)}||^2}\\ &\quad+\frac{\left|\left\langle\widehat{v}^*_{j,k(n)},v^*_{j,k(n)}\right\rangle-\left\langle v^*_{j,k(n)},v^*_{j,k(n)}\right\rangle\right|}{||v^*_{j,k(n)}||^2}\\ &\le\frac{||\widehat{v}^*_{j,k(n)}||}{||v^*_{j,k(n)}||}\frac{||\widehat{v}^*_{j,k(n)}-v^*_{j,k(n)}||}{||v^*_{j,k(n)}||}+\frac{||\widehat{v}^*_{j,k(n)}-v^*_{j,k(n)}||}{||v^*_{j,k(n)}||}. \end{aligned} \quad \text{(A.6)}$$

By Assumption A.3.(ii) and the result in equation (A.5), we have:

$$\begin{aligned} O_p(\delta_{w,n}) &= \sup_{v\in\mathcal{V}_{k(n)}} \left| \frac{\Gamma_{2,j,n}(\widehat{\theta}_n,\widehat{h}_n)[v]-\Gamma_{2,j}(\theta_o,h_o)[v]}{\|v\|} \right| \\ &= \sup_{v\in\mathcal{V}_{k(n)}} \left| \frac{\left\langle \widehat{v}^*_{j,k(n)}, v\right\rangle_n - \left\langle \widehat{v}^*_{j,k(n)}, v\right\rangle}{||\widehat{v}^*_{j,k(n)}||\,\|v\|} ||\widehat{v}^*_{j,k(n)}|| + \frac{\left\langle \widehat{v}^*_{j,k(n)} - v^*_{j,k(n)}, v\right\rangle}{\|v\|} \right| \\ &= \sup_{v\in\mathcal{V}_{k(n)}} \left| O_p(\delta_{w,n}||\widehat{v}^*_{j,k(n)}||) + \frac{\langle \widehat{v}^*_{j,k(n)} - v^*_{j,k(n)}, v\rangle}{\|v\|} \right|. \end{aligned} \tag{A.7}$$

Let $v = \widehat{v}^*_{j,k(n)} - v^*_{j,k(n)}$ in equation (A.7) and using Assumption A.3.(iii), we get

$$\frac{||\widehat{v}^*_{j,k(n)} - v^*_{j,k(n)}||}{||v^*_{j,k(n)}||} = O_p(\delta_{w,n}) \frac{||\widehat{v}^*_{j,k(n)}||}{||v^*_{j,k(n)}||} + O_p(\delta_{w,n}). \tag{A.8}$$

Plugging the above equation into equation (A.6) and using the triangle inequality, we get

$$\left| \frac{||\widehat{v}^*_{j,k(n)}||^2}{||v^*_{j,k(n)}||^2} - 1 \right| \leq \left| \frac{||\widehat{v}^*_{j,k(n)}||^2}{||v^*_{j,k(n)}||^2} - 1 \right| O_p(\delta_{w,n}) + \left| \frac{||\widehat{v}^*_{j,k(n)}||}{||v^*_{j,k(n)}||} - 1 \right| O_p(\delta_{w,n}) + O_p(\delta_{w,n}),$$

which implies that

$$\frac{\left| ||\widehat{v}^*_{j,k(n)}|| - ||v^*_{j,k(n)}|| \right|}{||v^*_{j,k(n)}||} = O_p(\delta_{w,n}). \tag{A.9}$$

By Assumption A.3.(iii) and the results in equations (A.8) and (A.9), we obtain

$$\left\| v^*_{j,k(n)} - \widehat{v}^*_{j,k(n)} \right\| = O_p(\delta_{w,n}), \text{ for any } j = 1, ..., d_g$$

which finishes the proof. ■

B Proof of the Results in Section 3

For any random vector $S = (S_1, ..., S_d)'$, we define $\|S\|_p = \left(\sum_{j=1}^d E\left[|S_j|^p\right]\right)^{1/p}$. In the following we slightly abuse notation and let $\mathcal{N}_n$ also denote a local shrinking neighborhood of h_o: $\{h \in \mathcal{H}_{k(n)} : ||h - h_o||_{\mathcal{H}} \leq \delta_{s,n}, ||h - h_o|| \leq \delta_n\}$ where $\delta_{s,n} = o(1)$ and $\delta_n = o(n^{-\frac{1}{4}})$. Denote

$$S_i(\alpha)\,[\mathbf{v}] = g(Z_i,\alpha) + \Delta(Z_i,h)[\mathbf{v}] \quad \text{and} \quad \widehat{S}_i(\alpha)\,[\mathbf{v}] = g(Z_i,\alpha) + \widehat{\Delta}(Z_i,h)[\mathbf{v}],$$

$\Delta(Z,h)\,[\mathbf{v}] = \left(\Delta(Z,h)[v_1], \ldots, \Delta(Z,h)[v_{d_g}]\right)'$ and $\widehat{\Delta}(Z,h)\,[\mathbf{v}] = \left(\widehat{\Delta}(Z,h)[v_1], \ldots, \widehat{\Delta}(Z,h)[v_{d_g}]\right)'$.

Assumption B.1 *The kernel function $\mathcal{K}(\cdot)$ is symmetric, continuous at zero, and satisfies $\mathcal{K}(0) = 1$, $\sup_x |\mathcal{K}(x)| \leq 1$, $\int_{\mathbb{R}} |\mathcal{K}(x)|dx < \infty$ and $\int_{\mathbb{R}} |\mathcal{K}(x)|\,|x|\,dx < \infty$.*

Assumption B.2 *(i) $\{Z_i\}$ is a strictly stationary strong mixing process with mixing coefficient α_i satisfying $\sum_{i=0}^{\infty} \alpha_i^{2(1/r-1/p)} < \infty$ for some $r \in (2,4]$ and some $p > r$; (ii) $\sup_{k(n)} \|S_i(\alpha_o)[\mathbf{v}_n^*]\|_p < \infty$ and $\sup_{k(n)} E\left[\sup_{v\in\mathcal{W}_n} |\Delta(Z,h_o)[v]|^2\right] < \infty$; (iii) There is a positive sequence $\delta_n^* = o(1)$ such that*

$$E\left[\sup_{\alpha\in\mathcal{N}_n,\mathbf{v}\in\mathcal{W}_n\times\ldots\times\mathcal{W}_n} \|S_i(\alpha)[\mathbf{v}] - S_i(\alpha_o)[\mathbf{v}]\|_E^2\right] = O(\delta_n^*\delta_{w,n});$$

$$\textit{(iv)} \quad \sup_{h\in\mathcal{N}_n,v\in\mathcal{W}_n} \frac{1}{n}\sum_{i=1}^{n} \left|\widehat{\Delta}(Z_i,h)[v] - \Delta(Z_i,h)[v]\right|^2 = O_p(\delta_n^*\delta_{w,n});$$

(v) $M_n \times \max(\delta_{w,n}, \delta_n^*) = o(1)$ and $n^{-1/2+1/r} M_n = o(1)$.

Assumption B.2.(i) and (ii) are the conditions on the dependence and moments of the data. Assumption B.2.(iii) imposes a local uniform smoothness condition on the score function $S_i(\alpha)[\mathbf{v}]$. Assumption B.2.(iv) imposes a convergence rate of $\widehat{\Delta}(Z_i,h)[v]$ to $\Delta(Z_i,h)[v]$ in the case when the functional form of $\Delta(Z_i,h)[v]$ is unknown. Assumption B.2.(v) is the condition on the bandwidth M_n of the kernel function $\mathcal{K}(\cdot)$. It is clear that all we need is $\delta_n^*\delta_{w,n} = o(1)$ in Assumption B.2.(iii)(iv)(v) when there is zero auto-correlation.

Proof of Theorem 3.1. In this proof, we assume that $S_i(\alpha_o)[\mathbf{v}]$ is a scalar for the ease of notation. As noted in Newey and West (1987), the results established in the scalar case can be directly applied to vector-valued $S_i(\alpha_o)[\mathbf{v}]$. Moreover, we use c to denote some generic positive and finite constant.

For the first result (43), we recall the sieve LRV $V_{1,n}^* = \sum_{i=-n+1}^{n-1} \Upsilon_i(\alpha_o)[\mathbf{v}_n^*, \mathbf{v}_n^*]$ as expressed in (39). Using the kernel function $\mathcal{K}(\cdot)$, we define

$$\widetilde{V}_{1,n} = \sum_{i=-n+1}^{n-1} \mathcal{K}\left(\frac{i}{M_n}\right) \Upsilon_i(\alpha_o)[\mathbf{v}_n^*, \mathbf{v}_n^*] \text{ and } V_{1,n} = \sum_{i=-n+1}^{n-1} \mathcal{K}\left(\frac{i}{M_n}\right) \Upsilon_{n,i}(\alpha_o)[\mathbf{v}_n^*, \mathbf{v}_n^*], \tag{B.1}$$

where

$$\Upsilon_{n,i}(\alpha_o)[\mathbf{v}_n^*, \mathbf{v}_n^*] = \begin{cases} \frac{1}{n}\sum_{l=i+1}^{n} S_{l,n}^* S_{l-i,n}^{*\prime} & \text{for } i \geq 0 \\ \frac{1}{n}\sum_{l=-i+1}^{n} S_{l,n}^* S_{l+i,n}^{*\prime} & \text{for } i < 0 \end{cases}$$

with $S_{i,n}^* = S_i(\alpha_o)[\mathbf{v}_n^*]$ given in (25) for $i = 1, ..., n$. By the triangle inequality, we have

$$\left|\widehat{V}_{1,n} - V_1\right| \leq \left|\widehat{V}_{1,n} - V_{1,n}\right| + \left|V_{1,n} - \widetilde{V}_{1,n}\right| + \left|\widetilde{V}_{1,n} - V_{1,n}^*\right| + \left|V_{1,n}^* - V_1\right|. \tag{B.2}$$

By the definitions of $V_{1,n}^*$ and V_1, we have:

$$\left|V_{1,n}^* - V_1\right| = o(1). \tag{B.3}$$

By the triangle inequality we get

$$\left|\widetilde{V}_{1,n} - V_{1,n}^*\right| \leq \frac{1}{n}\sum_{i=0}^{n-1}\left|\mathcal{K}\left(\frac{i}{M_n}\right) - 1\right|\sum_{l=i+1}^{n}\left|E\left\{S_l(\alpha_o)\left[\mathbf{v}_n^*\right]S_{l-i}(\alpha_o)\left[\mathbf{v}_n^*\right]\right\}\right|$$
$$+\frac{1}{n}\sum_{i=1-n}^{-1}\left|\mathcal{K}\left(\frac{i}{M_n}\right) - 1\right|\sum_{l=1-i}^{n}\left|E\left\{S_l(\alpha_o)\left[\mathbf{v}_n^*\right]S_{l+i}(\alpha_o)\left[\mathbf{v}_n^*\right]\right\}\right|. \qquad \text{(B.4)}$$

Using Assumption B.2.(ii) and the strong mixing inequality, we get

$$\sup_{k(n)}\left|E\left\{S_l(\alpha_o)\left[\mathbf{v}_n^*\right]S_{l-i}(\alpha_o)\left[\mathbf{v}_n^*\right]\right\}\right| \leq 6\alpha_i^{2\left(\frac{1}{2}-\frac{1}{p}\right)}\sup_{k(n)}\left\|S_l(\alpha_o)\left[\mathbf{v}_n^*\right]\right\|_p^2 \leq c\alpha_i^{2\left(\frac{1}{r}-\frac{1}{p}\right)}. \qquad \text{(B.5)}$$

Inequalities (B.4) and (B.5) immediately lead to

$$\left|\widetilde{V}_{1,n} - V_{1,n}^*\right| \leq 2c \times \sum_{i=0}^{n-1}\left|\mathcal{K}\left(\frac{i}{M_n}\right) - 1\right|\alpha_i^{2/r-2/p} \to 0, \qquad \text{(B.6)}$$

where the last result is by Assumptions B.1 and B.2.(i), and the dominated convergence theorem.

For the second term in the right-hand side of inequality (B.2), by Minkowski's inequality we get

$$\left\|V_{1,n} - \widetilde{V}_{1,n}\right\|_{r/2} \leq \sum_{i=0}^{n-1}\left|\mathcal{K}\left(\frac{i}{M_n}\right)\right|\left\|\Upsilon_i(\alpha_o)\left[\mathbf{v}_n^*, \mathbf{v}_n^*\right] - \Upsilon_{n,i}(\alpha_o)\left[\mathbf{v}_n^*, \mathbf{v}_n^*\right]\right\|_{r/2}$$
$$+\sum_{i=-n+1}^{-1}\left|\mathcal{K}\left(\frac{i}{M_n}\right)\right|\left\|\Upsilon_i(\alpha_o)\left[\mathbf{v}_n^*, \mathbf{v}_n^*\right] - \Upsilon_{n,i}(\alpha_o)\left[\mathbf{v}_n^*, \mathbf{v}_n^*\right]\right\|_{r/2}.$$

Under Assumption B.2.(i)-(ii), we can invoke Lemma 2 in Hansen (1992) and the proof of Theorem 2 in de Jong (2000) to deduce that

$$\left\|\Upsilon_i(\alpha_o)\left[\mathbf{v}_n^*, \mathbf{v}_n^*\right] - \Upsilon_{n,i}(\alpha_o)\left[\mathbf{v}_n^*, \mathbf{v}_n^*\right]\right\|_{r/2} \leq c(c+|i|)n^{-1+2/r} \qquad \text{(B.7)}$$

for any $r > 2$. Thus

$$\left\|V_{1,n} - \widetilde{V}_{1,n}\right\|_{r/2} \leq cM_n^2 n^{-1+\frac{2}{r}}\sum_{i=-n+1}^{n-1}\left|\mathcal{K}\left(\frac{i}{M_n}\right)\right|\frac{c+|i|}{M_n}\frac{1}{M_n} = o(1), \qquad \text{(B.8)}$$

where the last equality is by Assumptions B.1 and B.2.(v). Using Markov inequality, from (B.8) we obtain

$$\left|V_{1,n} - \widetilde{V}_{1,n}\right| = o_p(1). \qquad \text{(B.9)}$$

We next deal with the first term in the right-hand side of inequality (B.2). Using the triangle inequality, we have

$$\begin{aligned}\left|\widehat{V}_{1,n}-V_{1,n}\right| &\leq \sum_{i=-n+1}^{n-1}\left|\mathcal{K}\left(\frac{i}{M_n}\right)\right| \left|\Upsilon_{n,i}(\widehat{\alpha}_n)[\widehat{\mathbf{v}}_n^*,\widehat{\mathbf{v}}_n^*]-\Upsilon_{n,i}(\alpha_o)\left[\widehat{\mathbf{v}}_n^*,\widehat{\mathbf{v}}_n^*\right]\right| \\ &+\sum_{i=-n+1}^{n-1}\left|\mathcal{K}\left(\frac{i}{M_n}\right)\right| \left|\Upsilon_{n,i}(\alpha_o)\left[\widehat{\mathbf{v}}_n^*,\widehat{\mathbf{v}}_n^*\right]-\Upsilon_{n,i}(\alpha_o)\left[\widehat{\mathbf{v}}_n^*,\mathbf{v}_n^*\right]\right| \\ &+\sum_{i=-n+1}^{n-1}\left|\mathcal{K}\left(\frac{i}{M_n}\right)\right| \left|\Upsilon_{n,i}(\alpha_o)\left[\widehat{\mathbf{v}}_n^*,\mathbf{v}_n^*\right]-\Upsilon_{n,i}(\alpha_o)[\mathbf{v}_n^*,\mathbf{v}_n^*]\right| \\ &\equiv I_{1,n}+I_{2,n}+I_{3,n}. \end{aligned} \tag{B.10}$$

We first deal with the third term $I_{3,n}$ in equation (B.10). Consider the case that $i \geq 0$ (same bound can be derived when $i < 0$),

$$\begin{aligned} &\sup_{0\leq i\leq n-1}\left|\Upsilon_{n,i}(\alpha_o)\left[\widehat{\mathbf{v}}_n^*,\mathbf{v}_n^*\right]-\Upsilon_{n,i}(\alpha_o)\left[\mathbf{v}_n^*,\mathbf{v}_n^*\right]\right| \\ &\leq \sup_{0\leq i\leq n-1}\frac{1}{n}\sum_{l=i+1}^{n}\left|S_{l-i}(\alpha_o)\left[\mathbf{v}_n^*\right]\Delta(Z_l,h_o)\left[\mathbf{v}_n^*-\widehat{\mathbf{v}}_n^*\right]\right| \\ &\leq \left[\frac{1}{n}\sum_{i=1}^{n}\left|S_i(\alpha_o)\left[\mathbf{v}_n^*\right]\right|^2\right]^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\Delta(Z_i,h_o)\left[\mathbf{v}_n^*-\widehat{\mathbf{v}}_n^*\right]\right|^2\right]^{1/2} \end{aligned} \tag{B.11}$$

where the first inequality is by the triangle inequality and the second inequality is by Cauchy-Schwarz inequality. Using Assumption B.2.(ii) and Markov inequality, we get

$$\frac{1}{n}\sum_{i=1}^{n}\left|S_i(\alpha_o)\left[\mathbf{v}_n^*\right]\right|^2=O_p(1) \text{ and } \frac{1}{n}\sum_{i=1}^{n}\left|\Delta(Z_i,h_o)\left[\mathbf{u}_n^*-\widehat{\mathbf{u}}_n^*\right]\right|=O_p(1) \tag{B.12}$$

where $\mathbf{u}_n^*=\mathbf{v}_n^*||\mathbf{v}_n^*-\widehat{\mathbf{v}}_n^*||^{-1}$ and $\widehat{\mathbf{u}}_n^*=\widehat{\mathbf{v}}_n^*||\mathbf{v}_n^*-\widehat{\mathbf{v}}_n^*||^{-1}$. Under Assumptions B.1 and B.2.(v), from the results in Lemma 2.1, (B.11) and (B.12), we deduce that

$$\begin{aligned} |I_{3,n}| &\leq \left[\frac{1}{n}\sum_{i=1}^{n}\left|S_i(\alpha_o)\left[\mathbf{v}_n^*\right]\right|^2\right]^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\Delta(Z_i,h_o)\left[\mathbf{v}_n^*-\widehat{\mathbf{v}}_n^*\right]\right|^2\right]^{1/2}\sum_{i=-n+1}^{n-1}\left|\mathcal{K}\left(\frac{i}{M_n}\right)\right| \\ &=O_p(M_n||\mathbf{v}_n^*-\widehat{\mathbf{v}}_n^*||)\sum_{i=-n+1}^{n-1}\left|\mathcal{K}\left(\frac{i}{M_n}\right)\right|\frac{1}{M_n}=O_p(M_n\delta_{w,n})\int\left|\mathcal{K}\left(x\right)\right|dx=o_p(1). \end{aligned} \tag{B.13}$$

For the second term $I_{2,n}$ in equation (B.10), using similar arguments, we get

$$
\begin{aligned}
&\sup_{0\leq i\leq n-1} |\Upsilon_{n,i}(\alpha_o)\,[\widehat{\mathbf{v}}_n^*, \widehat{\mathbf{v}}_n^*] - \Upsilon_{n,i}(\alpha_o)\,[\widehat{\mathbf{v}}_n^*, \mathbf{v}_n^*]| \\
&\leq \sup_{0\leq i\leq n-1} \frac{1}{n}\sum_{l=i+1}^{n} |\{S_l(\alpha_o)\,[\mathbf{v}_n^*]\}\,\Delta(Z_{l-i}, h_o)\,[\mathbf{v}_n^* - \widehat{\mathbf{v}}_n^*]| \\
&+ \sup_{0\leq i\leq n-1} \frac{1}{n}\sum_{l=i+1}^{n} |\Delta(Z_l, h_o)\,[\widehat{\mathbf{v}}_n^* - \mathbf{v}_n^*]\,\Delta(Z_{l-i}, h_o)\,[\mathbf{v}_n^* - \widehat{\mathbf{v}}_n^*]| \\
&\leq \left[\frac{1}{n}\sum_{i=1}^{n} |S_i(\alpha_o)\,[\mathbf{v}_n^*]\,|^2\right]^{1/2} \left[\frac{1}{n}\sum_{i=1}^{n} |\Delta(Z_i, h_o)\,[\mathbf{v}_n^* - \widehat{\mathbf{v}}_n^*]\,|^2\right]^{1/2} \\
&+ \frac{1}{n}\sum_{i=1}^{n} |\Delta(Z_i, h_o)\,[\mathbf{v}_n^* - \widehat{\mathbf{v}}_n^*]\,|^2 \\
&= O_p(||\mathbf{v}_n^* - \widehat{\mathbf{v}}_n^*|| + ||\mathbf{v}_n^* - \widehat{\mathbf{v}}_n^*||^2) = O_p(\delta_{w,n}), \qquad \text{(B.14)}
\end{aligned}
$$

where the first two inequalities are by the triangle inequality, the third inequality is by Cauchy-Schwarz inequality and the last equality is by (B.12) and Lemma 2.1. Using (B.14), Assumptions B.1 and B.2.(v), we deduce that

$$
|I_{2,n}| = O_p(M_n\delta_{w,n}) \sum_{i=-n+1}^{n-1} \left|\mathcal{K}\left(\frac{i}{M_n}\right)\right| \frac{1}{M_n} = O_p(M_n\delta_{w,n}) \int_R |\mathcal{K}\,(x)|\,dx = o_p(1). \qquad \text{(B.15)}
$$

For the first term $I_{1,n}$ in equation (B.10), note that when $i \geq 0$ (same result can be derived when $i < 0$),

$$
\begin{aligned}
&\sup_{0\leq i\leq n-1} |\Upsilon_{n,i}(\widehat{\alpha}_n)[\widehat{\mathbf{v}}_n^*, \widehat{\mathbf{v}}_n^*] - \Upsilon_{n,i}(\alpha_o)\,[\widehat{\mathbf{v}}_n^*, \widehat{\mathbf{v}}_n^*]| \\
&\leq \sup_{0\leq i\leq n-1} \frac{1}{n}\sum_{l=i+1}^{n} \left|\widehat{S}_l(\widehat{\alpha}_n)\,[\widehat{\mathbf{v}}_n^*]\,\widehat{S}_{l-i}(\widehat{\alpha}_n)\,[\widehat{\mathbf{v}}_n^*] - S_l(\alpha_o)\,[\widehat{\mathbf{v}}_n^*]\,S_{l-i}(\alpha_o)\,[\widehat{\mathbf{v}}_n^*]\right|.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
|I_{1,n}| &\leq \frac{c}{n}\sum_{i=1}^{n} \left|\widehat{S}_i(\widehat{\alpha}_n)\,[\widehat{\mathbf{v}}_n^*] - S_i(\alpha_o)\,[\widehat{\mathbf{v}}_n^*]\right|^2 \\
&+ \frac{2c}{n}\sqrt{\sum_{i=1}^{n} \left|\widehat{S}_i(\widehat{\alpha}_n)\,[\widehat{\mathbf{v}}_n^*] - S_i(\alpha_o)\,[\widehat{\mathbf{v}}_n^*]\right|^2}\sqrt{\sum_{i=1}^{n} |\Delta(Z_i, h_o)\,[\mathbf{v}_n^* - \widehat{\mathbf{v}}_n^*]\,|^2} \\
&+ \frac{2c}{n}\sqrt{\sum_{i=1}^{n} \left|\widehat{S}_i(\widehat{\alpha}_n)\,[\widehat{\mathbf{v}}_n^*] - S_i(\alpha_o)\,[\widehat{\mathbf{v}}_n^*]\right|^2}\sqrt{\sum_{i=1}^{n} |S_i(\alpha_o)\,[\mathbf{v}_n^*]\,|^2} \qquad \text{(B.16)}
\end{aligned}
$$

where the first inequality is by the triangle inequality, the second inequality is by the triangle inequality and Cauchy-Schwarz inequality. Using the simple inequality $(x+y)^2 \leq 2(x^2+y^2)$, we get

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left|\widehat{S}_i(\widehat{\alpha}_n)\left[\widehat{\mathbf{v}}_n^*\right] - S_i(\alpha_o)\left[\widehat{\mathbf{v}}_n^*\right]\right|^2 \\
&\leq \frac{2}{n}\sum_{i=1}^{n}\left|\widehat{S}_i(\widehat{\alpha}_n)\left[\widehat{\mathbf{v}}_n^*\right] - S_i(\widehat{\alpha}_n)\left[\widehat{\mathbf{v}}_n^*\right]\right|^2 + \frac{2}{n}\sum_{i=1}^{n}\left|S_i(\widehat{\alpha}_n)\left[\widehat{\mathbf{v}}_n^*\right] - S_i(\alpha_o)\left[\widehat{\mathbf{v}}_n^*\right]\right|^2 \\
&= \frac{2}{n}\sum_{i=1}^{n}\left|\widehat{\Delta}(Z_i,\widehat{h}_n)\left[\widehat{\mathbf{v}}_n^*\right] - \Delta(Z_i,\widehat{h}_n)\left[\widehat{\mathbf{v}}_n^*\right]\right|^2 + O_p(\delta_n^*\delta_{w,n}) = O_p(\delta_n^*\delta_{w,n}) \qquad \text{(B.17)}
\end{aligned}
$$

where the first equality is by Assumptions B.2.(iii) and the last equality is by Assumptions B.2.(iv). By equations (B.12), (B.16) and (B.17), and Lemma 2.1, we obtain that

$$
\sup_{0\leq i\leq n-1}\left|\Upsilon_{n,i}(\widehat{\alpha}_n)[\widehat{\mathbf{v}}_n^*,\widehat{\mathbf{v}}_n^*] - \Upsilon_{n,i}(\alpha_o)\left[\widehat{\mathbf{v}}_n^*,\widehat{\mathbf{v}}_n^*\right]\right| = O_p(\sqrt{\delta_n^*\delta_{w,n}}).
$$

This, Assumptions B.1 and B.2.(v) together imply that

$$
|I_{1,n}| = O_p(M_n\sqrt{\delta_n^*\delta_{w,n}})\sum_{i=-n+1}^{n-1}\left|\mathcal{K}\left(\frac{i}{M_n}\right)\right|\frac{1}{M_n} = O_p(M_n\sqrt{\delta_n^*\delta_{w,n}})\int|\mathcal{K}(x)|\,dx = o_p(1). \qquad \text{(B.18)}
$$

By equations (B.10), (B.13), (B.15) and (B.18), we obtain:

$$
\left|\widehat{V}_{1,n} - V_{1,n}\right| = o_p(1),
$$

which, together with the results in (B.3), (B.6) and (B.9), implies that $\widehat{V}_{1,n} - V_1 = o_p(1)$.

For the second result (44), note that by definition, for all $0 \leq i \leq n-1$,

$$
\begin{aligned}
&\overline{\Upsilon}_{n,i}(\widehat{\alpha}_n)\left[\widehat{\mathbf{v}}_n^*,\widehat{\mathbf{v}}_n^*\right] - \Upsilon_{n,i}(\widehat{\alpha}_n)\left[\widehat{\mathbf{v}}_n^*,\widehat{\mathbf{v}}_n^*\right] \\
&= \frac{n-i}{n}\left(\widehat{S}_n^*\right)^2 - \widehat{S}_n^*\left(\frac{1}{n}\sum_{l=1}^{n-i}\widehat{S}_{l,n}^*\right) - \widehat{S}_n^*\left(\frac{1}{n}\sum_{l=i+1}^{n}\widehat{S}_{l,n}^*\right). \qquad \text{(B.19)}
\end{aligned}
$$

Using (B.17), Cauchy-Schwarz inequality and Markov inequality, we get

$$
\begin{aligned}
&\sup_{0\leq i\leq n-1}\left|\frac{1}{n}\sum_{l=i+1}^{n}\left[\widehat{S}_i(\widehat{\alpha}_n)\left[\widehat{\mathbf{v}}_n^*\right] - S_i(\alpha_o)\left[\widehat{\mathbf{v}}_n^*\right]\right]\right| \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\widehat{S}_i(\widehat{\alpha}_n)\left[\widehat{\mathbf{v}}_n^*\right] - S_i(\alpha_o)\left[\widehat{\mathbf{v}}_n^*\right]\right|^2} = O_p(\sqrt{\delta_n^*\delta_{w,n}}). \qquad \text{(B.20)}
\end{aligned}
$$

Using the result in (B.12), Cauchy-Schwarz inequality and Lemma 2.1, we deduce that

$$\begin{aligned}
&\sup_{0\leq i\leq n-1}\left|\frac{1}{n}\sum_{l=i+1}^{n}\left[S_i(\alpha_o)\left[\widehat{\mathbf{v}}_n^*\right]-S_i(\alpha_o)\left[\mathbf{v}_n^*\right]\right]\right| \\
&=\sup_{0\leq i\leq n-1}\frac{1}{n}\left|\sum_{l=i+1}^{n}\left[\Delta(Z_i,h_o)\left[\widehat{\mathbf{v}}_n^*\right]-\Delta(Z_i,h_o)\left[\mathbf{v}_n^*\right]\right]\right| \\
&\leq\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\Delta(Z_i,h_o)\left[\widehat{\mathbf{v}}_n^*-\mathbf{v}_n^*\right]\right|^2}=O_p(\delta_{w,n}).
\end{aligned} \tag{B.21}$$

By Assumption A.2.(v), we have

$$\sup_{0\leq i\leq n-1}\left|\frac{1}{n}\sum_{l=i+1}^{n}S_i(\alpha_o)\left[\mathbf{v}_n^*\right]\right|=O_p(n^{-1/2}). \tag{B.22}$$

Equations (B.21) and (B.22) imply that

$$\sup_{0\leq i\leq n-1}\left|\frac{1}{n}\sum_{l=i+1}^{n}S_i(\alpha_o)\left[\widehat{\mathbf{v}}_n^*\right]\right|=O_p(\delta_{w,n}).$$

This and Equation (B.20) imply that

$$\sup_{0\leq i\leq n-1}\left|\frac{1}{n}\sum_{l=i+1}^{n}\widehat{S}_{l,n}^*\right|=O_p(\sqrt{\delta_n^*\delta_{w,n}})+O_p(\delta_{w,n}).$$

Similarly we can show that

$$\sup_{0\leq i\leq n-1}\left|\frac{1}{n}\sum_{l=1}^{n-i}\widehat{S}_{l,n}^*\right|=O_p(\sqrt{\delta_n^*\delta_{w,n}})+O_p(\delta_{w,n}),$$

$$\left|\widehat{S}_n^*\right|\equiv\left|\frac{1}{n}\sum_{i=1}^{n}\widehat{S}_{i,n}^*\right|=O_p(\sqrt{\delta_n^*\delta_{w,n}})+O_p(\delta_{w,n}).$$

These, together with (B.19), imply that

$$\sup_{0\leq i\leq n-1}\left|\overline{\Upsilon}_{n,i}(\widehat{\alpha}_n)\left[\widehat{\mathbf{v}}_n^*,\widehat{\mathbf{v}}_n^*\right]-\Upsilon_{n,i}(\widehat{\alpha}_n)\left[\widehat{\mathbf{v}}_n^*,\widehat{\mathbf{v}}_n^*\right]\right|=O_p(\max(\delta_{w,n},\delta_n^*)\times\delta_{w,n}).$$

By the definition of $\widehat{V}_{c,1,n}$ and Assumption B.1, we have

$$\begin{aligned}
\left|\widehat{V}_{c,1,n}-\widehat{V}_{1,n}\right|&=\left|\sum_{i=-n+1}^{n-1}\mathcal{K}\left(\frac{i}{M_n}\right)\left[\overline{\Upsilon}_{n,i}(\widehat{\alpha}_n)\left[\widehat{\mathbf{v}}_n^*,\widehat{\mathbf{v}}_n^*\right]-\Upsilon_{n,i}(\widehat{\alpha}_n)\left[\widehat{\mathbf{v}}_n^*,\widehat{\mathbf{v}}_n^*\right]\right]\right| \\
&=O_p(M_n\max(\delta_{w,n},\delta_n^*)\times\delta_{w,n})\sum_{i=-n+1}^{n-1}\left|\mathcal{K}\left(\frac{i}{M_n}\right)\right|\frac{1}{M_n}=o_p(\delta_{w,n}).
\end{aligned}$$

This finishes the proof. ■

Proof of Proposition 3.2. First, applying Theorem 2.1, we have

$$\sqrt{n}(\widetilde{\theta}_n - \theta_o) \to_d \left(\Gamma_1' W \Gamma_1\right)^{-1} \Gamma_1' W \times \mathcal{N}(0, V_1). \tag{B.23}$$

Hence by Theorem 3.1, we obtain:

$$\widetilde{W}_n^{-1} \to_p V_1. \tag{B.24}$$

Equation (B.24) and Theorem 2.1 imply that

$$\sqrt{n}(\widehat{\theta}_n - \theta_o) = -\left(\Gamma_1' V_1^{-1} \Gamma_1\right)^{-1} \Gamma_1' V_1^{-1} \frac{\sum_{i=1}^n S_{i,n}^*}{\sqrt{n}} + o_p(1) \to_d \mathcal{N}(0, V_\theta^o). \tag{B.25}$$

By definition, $E_{z_i}\left[g(Z_i, \widehat{\theta}_n, \widehat{h}_n)\right] = G(\widehat{\theta}_n, \widehat{h}_n)$, then under Assumption A.2.(i)-(iii), Theorem 2.1 and the Riesz Representation Theorem, we deduce that

$$\begin{aligned} G(\widehat{\theta}_n, \widehat{h}_n) &= G(\widehat{\theta}_n, \widehat{h}_n) - G(\widehat{\theta}_n, h_o) - \Gamma_2(\widehat{\theta}_n, h_o)[\widehat{h}_n - h_o] \\ &\quad + \Gamma_2(\widehat{\theta}_n, h_o)[\widehat{h}_n - h_o] - \Gamma_2(\theta_o, h_o)[\widehat{h}_n - h_o] \\ &\quad + \Gamma_2(\theta_o, h_o)[\widehat{h}_n - h_o] + G(\widehat{\theta}_n, h_o) - G(\theta_o, h_o) \\ &= \Gamma_1(\theta_o, h_o)(\widehat{\theta}_n - \theta_o) + \Gamma_2(\theta_o, h_o)[\widehat{h}_n - h_o] + o_p(n^{-\frac{1}{2}}) \\ &= \langle \widehat{h}_n - h_o, \mathbf{v}_n^* \rangle + O_p(n^{-\frac{1}{2}}) = O_p(n^{-\frac{1}{2}}), \end{aligned} \tag{B.26}$$

where the third equality is due to the root-n consistency of $\widehat{\theta}_n$, and the last equality is due to Assumption A.3.(iii). By the consistency of $(\widehat{\theta}_n, \widehat{h}_n)$ and Assumption A.2.(iv), we have

$$\left\| G_n(\widehat{\theta}_n, \widehat{h}_n) - G(\widehat{\theta}_n, \widehat{h}_n) - G_n(\theta_o, h_o) \right\|_W = o_p(n^{-1/2}) + o_p(1) \left\| G(\widehat{\theta}_n, \widehat{h}_n) \right\|_W. \tag{B.27}$$

By equations (B.26) and (B.27), we deduce that

$$\left\| \frac{1}{n} \sum_{i=1}^n \left\{ g(Z_i, \widehat{\theta}_n, \widehat{h}_n) - g(Z_i, \theta_o, h_o) - E_{z_i}\left[g(Z_i, \widehat{\theta}_n, \widehat{h}_n)\right] \right\} \right\|_E = o_p(n^{-1/2}). \tag{B.28}$$

Let Σ be the square root matrix of V_1, i.e. $\Sigma^2 = V_1$. Using equation (B.28), we have

$$\begin{aligned} n^{-\frac{1}{2}} \sum_{i=1}^n g(Z_i, \widehat{\theta}_n, \widehat{h}_n) &= n^{-\frac{1}{2}} \sum_{i=1}^n g(Z_i, \theta_o, h_o) + \sqrt{n} E_{z_i}\left[g(Z_i, \widehat{\theta}_n, \widehat{h}_n)\right] + o_p(1) \\ &= n^{-\frac{1}{2}} \sum_{i=1}^n \left[g(Z_i, \theta_o, h_o) + \Gamma_2(\widehat{h}_n - h_o)\right] + \sqrt{n}\Gamma_1(\widehat{\theta}_n - \theta_o) + o_p(1) \\ &= \left(I_{d_g} - \Gamma_1\left(\Gamma_1' V_1^{-1} \Gamma_1\right)^{-1} \Gamma_1' V_1^{-1}\right) \frac{\sum_{i=1}^n S_{i,n}^*}{\sqrt{n}} + o_p(1) \\ &\to_d \left[I_{d_g} - \Gamma_1\left(\Gamma_1' V_1^{-1} \Gamma_1\right)^{-1} \Gamma_1' V_1^{-1}\right] \Sigma \times \mathcal{N}(0, I_{d_g}), \end{aligned} \tag{B.29}$$

where I_{d_g} denotes a $d_g \times d_g$ identity matrix, the first equality is by equation (B.28), the second equality is by Assumption A.2.(i), (ii) and (iii), the third equality is by equations (B.25) and (A.2) in the proof of Theorem 2.1, and the weak convergence is by Assumption A.2.(v).

Now, using the second result in Theorem 3.1, we deduce that

$$\widehat{W}_{c,n}^{-1} \to_p V_1 \tag{B.30}$$

which, together with equation (B.29), implies that

$$\begin{aligned} J_n &= \frac{\left(\Sigma^{-1}\sum_{i=1}^n S_{i,n}^*\right)'}{\sqrt{n}} Q_{d_g} \frac{\left(\Sigma^{-1}\sum_{i=1}^n S_{i,n}^*\right)}{\sqrt{n}} + o_p(1) \\ &\to_d \ B_{d_g}'(1) Q_{d_g} B_{d_g}(1) \overset{d}{=} \chi_{d_g - d_\theta}^2 \end{aligned} \tag{B.31}$$

under the assumption that all moment conditions are valid, where

$$Q_{d_g} = I_{d_g} - \Sigma^{-1}\Gamma_1 \left(\Gamma_1' V_1^{-1} \Gamma_1\right)^{-1} \Gamma_1' \Sigma^{-1}$$

is an idempotent matrix with rank $d_g - d_\theta$ and $B_{d_g}(1)$ is a d_g dimensional standard Gaussian random vector. ■

C Proof of the Results in Section 4

The following assumptions are useful to derive the asymptotic properties of $\widehat{V}_{R,n}$.

Assumption C.1 *(i) $\widehat{\Gamma}_1 \to_p \Gamma_1$ and $W_n \to_p W$, where W is some nonrandom, $d_g \times d_g$ positive definite matrix; (ii) let $\xi_i \sim$i.i.d. $\mathcal{N}(0, V_1)$, then we have for any $(\mathbf{x}_1, ., , , \mathbf{x}_m) \in \mathbb{R}^{d_g \times m}$,*

$$\begin{aligned} &\Pr\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_m\left(\frac{i}{n}\right) S_i(\alpha_o)[\mathbf{v}_n^*] < \mathbf{x}_m, m = 1, ...M\right) \\ &= \Pr\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_m\left(\frac{i}{n}\right) \xi_i < \mathbf{x}_m, m = 1, ...M\right) + o(1); \end{aligned}$$

(iii) $\{\phi_m\}_{m=0}^\infty$ is a sequence of orthonormal basis functions in $L_2([0,1])$ with $\phi_0(\cdot) \equiv 1$.

Assumption C.2 *The following conditions hold for $m = 1, ..., M$:*

$$(i) \quad \sup_{\alpha \in \mathcal{N}_n, v \in \mathcal{W}_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi_m\left(\frac{i}{n}\right) \{S_i(\alpha)[v] - S_i(\alpha_o)[v] - E_{Z_i}(S_i(\alpha)[v])\} \right| = o_p(1);$$

$$(ii) \quad \delta_{w,n} \times \sup_{v \in \mathcal{W}_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi_m\left(\frac{i}{n}\right) \Delta(Z_i, h_o)[v] \right| = o_p(1);$$

$$\text{(iii)} \quad \left|\frac{1}{n}\sum_{i=1}^{n}\phi_m(\frac{i}{n})\right| \times \sup_{h\in\mathcal{N}_n, v\in\mathcal{W}_n} \sqrt{n}\left|E_{Z_i}\left(\Delta(Z_i,h)\left[v\right] - \Delta(Z_i,h_o)\left[v\right]\right)\right| = o(1);$$

$$\text{(iv)} \quad \sup_{h\in\mathcal{N}_n, v\in\mathcal{W}_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_m\left(\frac{i}{n}\right)\left\{\widehat{\Delta}(Z_i,h)\left[v\right] - \Delta(Z_i,h)\left[v\right]\right\}\right| = o_p(1).$$

Assumption C.1.(ii) is essentially a functional central limit theorem, but it holds under more general data structure (say weakly spatial dependence for example). Assumption C.1.(iii) is about the orthonormal basis. It implies $\int_0^1 \phi_m(r)\,dr = 0$ for all $m \geq 1$ and hence $\frac{1}{n}\sum_{i=1}^{n}\phi_m(\frac{i}{n}) = o(1)$. Assumption C.2.(i) is a stochastic equicontinuity condition that can be verified by applying various empirical process results for weakly dependent data. Assumption C.2.(iii) imposes smoothness condition on the expectation of the first step sieve extremum estimation criterion. Assumption C.2.(iv) is trivially satisfied when the first step is a sieve M estimation (since $\widehat{\Delta}(Z,h)\left[v\right] = \Delta(Z,h)\left[v\right]$). It can be verified when the first step is a sieve MD estimation by checking the following two sufficient conditions:

$$\sup_{h\in\mathcal{N}_n, v\in\mathcal{W}_n}\left|\sum_{i=1}^{n}\phi_m(\frac{i}{n})\left\{\widehat{\Delta}(Z_i,h)\left[v\right] - \Delta(Z_i,h)\left[v\right] - E_{Z_i}[\widehat{\Delta}(Z_i,h)\left[v\right] - \Delta(Z_i,h)\left[v\right]]\right\}\right| = o_p(n^{\frac{1}{2}}) \tag{C.1}$$

$$\text{and} \quad \left|\frac{1}{n}\sum_{i=1}^{n}\phi_m(\frac{i}{n})\right| \times \sup_{h\in\mathcal{N}_n, v\in\mathcal{W}_n} \sqrt{n}\left|E_{Z_i}\left(\widehat{\Delta}(Z_i,h)\left[v\right] - \Delta(Z_i,h)\left[v\right]\right)\right| = o_p(1). \tag{C.2}$$

We note that the smoother the basis function $\phi_m()$ is, the faster $\frac{1}{n}\sum_{i=1}^{n}\phi_m(\frac{i}{n})$ converges to zero. For example, if $\phi_m(\cdot)$ is absolutely continuous on $[0,1]$, then $\left|\frac{1}{n}\sum_{i=1}^{n}\phi_m(\frac{i}{n})\right| = O(n^{-1})$ (see, e.g., Chui, 1971). This makes Assumption C.2 easier to hold. In particular, Assumption C.2.(iii) and the sufficient condition (C.2) could be satisfied even when the first step nonparametric estimator converges to $h_o()$ slowly.

Proof of Lemma 4.1. We write

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_m\left(\frac{i}{n}\right)\widehat{S}_i(\widehat{\alpha}_n)\left[\widehat{\mathbf{v}}_n^*\right] - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_m\left(\frac{i}{n}\right)S_i(\alpha_o)\left[\mathbf{v}_n^*\right] \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_m\left(\frac{i}{n}\right)\left\{\widehat{S}_i(\widehat{\alpha}_n)\left[\widehat{\mathbf{v}}_n^*\right] - S_i(\alpha_o)\left[\widehat{\mathbf{v}}_n^*\right]\right\} \\
&+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_m\left(\frac{i}{n}\right)\left\{S_i(\alpha_o)\left[\widehat{\mathbf{v}}_n^*\right] - S_i(\alpha_o)\left[\mathbf{v}_n^*\right]\right\} \\
&= A_{1,n} + A_{2,n}.
\end{aligned} \tag{C.3}$$

By Lemma 2.1 and Assumption C.2.(ii), we have

$$|A_{2,n}| = \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_m \left(\frac{i}{n}\right) \{\Delta(Z_i, h_o)\,[\widehat{\mathbf{v}}_n^* - \mathbf{v}_n^*]\} \right|$$
$$\leq ||\widehat{\mathbf{v}}_n^* - \mathbf{v}_n^*|| \times \sup_{v \in \mathcal{W}_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_m \left(\frac{i}{n}\right) \Delta(Z_i, h_o)\,[v] \right| = o_p(1). \tag{C.4}$$

We now want to show that $A_{1,n} = o_p(1)$. Note that $A_{1,n} = A_{3,n} + A_{4,n}$ with

$$A_{3,n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_m \left(\frac{i}{n}\right) \{\widehat{S}_i(\widehat{\alpha}_n)\,[\widehat{\mathbf{v}}_n^*] - S_i(\widehat{\alpha}_n)\,[\widehat{\mathbf{v}}_n^*]\},$$
$$A_{4,n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_m \left(\frac{i}{n}\right) \{S_i(\widehat{\alpha}_n)\,[\widehat{\mathbf{v}}_n^*] - S_i(\alpha_o)\,[\widehat{\mathbf{v}}_n^*]\}\,.$$

First, by Lemma 2.1 and Assumption C.2.(iv), we have:

$$|A_{3,n}| \leq ||\widehat{\mathbf{v}}_n^*|| \times \sup_{v \in \mathcal{W}_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_m \left(\frac{i}{n}\right) \left\{\widehat{\Delta}(Z_i, \widehat{h}_n)\,[v] - \Delta(Z_i, \widehat{h}_n)\,[v]\right\} \right| = o_p(1). \tag{C.5}$$

Next Assumption C.1.(iii) implies $\int_0^1 \phi_m\,(r)\,dr = 0$ for all $m \geq 1$ and hence

$$\frac{1}{n} \sum_{i=1}^{n} \phi_m \left(\frac{i}{n}\right) = \int \phi_m\,(r)\,dr + o(1) = o(1). \tag{C.6}$$

By Lemma 2.1, Assumption C.2.(i) and Assumption C.2.(iii), we have

$$|A_{4,n}| \leq \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_m \left(\frac{i}{n}\right) E_{Z_i}\left\{S_i(\widehat{\alpha}_n)\,[\widehat{\mathbf{v}}_n^*] - S_i(\alpha_o)\,[\widehat{\mathbf{v}}_n^*]\right\} \right| + o_p(1)$$
$$= \left| \left\{ \frac{1}{n} \sum_{i=1}^{n} \phi_m \left(\frac{i}{n}\right) \right\} \sqrt{n} E_{Z_i} \left(g(Z_i, \widehat{\alpha}_n) + \Delta(Z_i, \widehat{h}_n)\,[\widehat{\mathbf{v}}_n^*] - \Delta(Z_i, h_o)\,[\widehat{\mathbf{v}}_n^*] \right) \right| + o_p(1)$$
$$\leq \left| \frac{1}{n} \sum_{i=1}^{n} \phi_m \left(\frac{i}{n}\right) \right| \times \left| \sqrt{n} E_{Z_i}\left(g(Z_i, \widehat{\alpha}_n)\right) \right| + o_p(1).$$

By definition of $E_{Z_i}\left(g(Z_i, \widehat{\alpha}_n)\right)$ and following the proof of Theorem 2.1, we have:

$$\sqrt{n} E_{Z_i}\left(g(Z_i, \widehat{\alpha}_n)\right) = \Gamma_1(\theta_o, h_o)\sqrt{n}(\widehat{\theta}_n - \theta_o) + \sqrt{n}\Gamma_2(\theta_o, h_o)[\widehat{h}_n - h_o] + o_p(1) = O_p(1).$$

This and equation (C.6) imply that

$$|A_{4,n}| \leq \left| \frac{1}{n} \sum_{i=1}^{n} \phi_m \left(\frac{i}{n}\right) \right| \times O_p(1) + o_p(1) = o_p(1). \tag{C.7}$$

Equations (C.3), (C.4), (C.5) and (C.7) imply that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_m\left(\frac{i}{n}\right)\widehat{S}_i(\widehat{\alpha}_n)\,[\widehat{\mathbf{v}}_n^*]=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_m\left(\frac{i}{n}\right)S_i(\alpha_o)[\mathbf{v}_n^*]+o_p(1). \tag{C.8}$$

By Assumptions C.1.(i)-(ii) and equation (C.8), we get

$$\begin{aligned}\Omega_\theta^{-1}\hat{\Lambda}_m &= \frac{\Omega_\theta^{-1}(\widehat{\Gamma}_1' W_n\widehat{\Gamma}_1)^{-1}\widehat{\Gamma}_1' W_n}{\sqrt{n}}\sum_{i=1}^{n}\phi_m\left(\frac{i}{n}\right)\widehat{S}_i(\widehat{\alpha}_n)\,[\widehat{\mathbf{v}}_n^*]\\ &= \frac{\Omega_\theta^{-1}(\Gamma_1' W\Gamma_1)^{-1}\Gamma_1' W}{\sqrt{n}}\sum_{i=1}^{n}\phi_m\left(\frac{i}{n}\right)S_i(\alpha_o)[\mathbf{v}_n^*]+o_p(1)\\ &\overset{d}{=} \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_m\left(\frac{i}{n}\right)\xi_{\theta,i}+o_p(1) \end{aligned}\tag{C.9}$$

where $\xi_{\theta,i}\sim i.i.d.\ \mathcal{N}(0,I_{d_\theta})$. Next note that for all $m=1,...,M$,

$$Var\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_m\left(\frac{i}{n}\right)\xi_{\theta,i}\right\}=\frac{1}{n}\sum_{i=1}^{n}\phi_m^2\left(\frac{i}{n}\right)=\int\phi_m^2\,(r)\,dr+o(1)=1+o(1). \tag{C.10}$$

Now, from equations (C.9) and (C.10), we deduce that $\Omega_\theta^{-1}\hat{\Lambda}_m\to_d B_{d_\theta,m}(1)$. The independence of $B_{d_\theta,m}(1)$ and $B_{d_\theta,m'}(1)$ for $m\neq m'$ is by the orthogonality of $\phi_m(\cdot)$ and $\phi_{m'}(\cdot)$. ■

Proof of Proposition 4.1. By equation (C.8) in the proof of Lemma 4.1, we have:

$$n^{-\frac{1}{2}}\sum_{i=1}^{n}\phi_m\left(\frac{i}{n}\right)\widehat{S}_i(\widetilde{\alpha}_n)\,[\widetilde{\mathbf{v}}_n^*]=n^{-\frac{1}{2}}\sum_{i=1}^{n}\phi_m\left(\frac{i}{n}\right)S_i(\alpha_o)[\mathbf{v}_n^*]+o_p(1), \tag{C.11}$$

which together with Assumptions C.1.(ii) implies that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_m\left(\frac{i}{n}\right)\widehat{S}_i(\widetilde{\alpha}_n)\,[\widetilde{\mathbf{v}}_n^*]\to_d\Sigma B_{d_g,m}(1), \tag{C.12}$$

where $\Sigma^2=V_1$. By CMT and equation (C.12), we obtain:

$$\widehat{W}_{R,n}^{-1}\to_d\Sigma\left[\frac{1}{M}\sum_{m=1}^{M}B_{d_g,m}(1)B_{d_g,m}'(1)\right]\Sigma\equiv\Sigma V_{R,\infty}\Sigma\equiv W_{R,\infty}^{-1}. \tag{C.13}$$

As $W_{R,\infty}$ is positive definite with probability one, using equation (C.13) and similar arguments in showing equation (A.3), we get

$$\sqrt{n}(\widehat{\theta}_{R,n}-\theta_o)=-\Gamma_1\left(\Gamma_1' W_{R,\infty}\Gamma_1\right)^{-1}\Gamma_1' W_{R,\infty}\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}S_i(\alpha_o)[\mathbf{v}_n^*]\right]+o_p(1). \tag{C.14}$$

Using equation (C.14) and similar arguments in deriving equation (B.29), we deduce that

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} g(Z_i, \widehat{\theta}_{R,n}, \widehat{h}_n) \to_d \left[I_{d_g} - \Gamma_1 \left(\Gamma_1' W_{R,\infty} \Gamma_1 \right)^{-1} \Gamma_1' W_{R,\infty} \right] \Sigma B_{d_g}(1). \tag{C.15}$$

By equations (C.13) and (C.15), we obtain:

$$\begin{aligned} J_{R,n} \to_d \; & \left[B_{d_g}(1) - \Gamma_{\Sigma 1} \left(\Gamma_{\Sigma 1}' V_{R,\infty}^{-1} \Gamma_{\Sigma 1} \right)^{-1} \Gamma_{\Sigma 1}' V_{R,\infty}^{-1} B_{d_g}(1) \right]' V_{R,\infty}^{-1} \\ & \times \left[B_{d_g}(1) - \Gamma_{\Sigma 1} \left(\Gamma_{\Sigma 1}' V_{R,\infty}^{-1} \Gamma_{\Sigma 1} \right)^{-1} \Gamma_{\Sigma 1}' V_{R,\infty}^{-1} B_{d_g}(1) \right], \end{aligned} \tag{C.16}$$

where $\Gamma_{\Sigma 1} = \Sigma^{-1} \Gamma_1$. Using similar arguments of Theorem 1 in Sun and Kim (2012), we can deduce that

$$J_{R,n}^* \equiv \frac{M - d_g + d_\theta + 1}{M(d_g - d_\theta)} J_{R,n} \to_d \mathcal{F}(d_g - d_\theta, M - d_g + d_\theta + 1).$$

■

D Proof of the Results in Section 5

Proof of Proposition 5.1. The proposition can be established by using a standard proof for parametric two-step GMM estimation such as in Newey and McFadden (1994), and hence its proof is omitted. The proof is available upon request, though. ■

Recall that $\widehat{\Gamma}_{2,P,n} \equiv \frac{1}{n} \sum_{i=1}^{n} \frac{\partial g_P(Z_i, \widehat{\theta}_{n,P}, \widehat{\beta}_{n,P})}{\partial \beta'}$ and $\widehat{R}_{n,P} \equiv -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \varphi_P(Z_i, \widehat{\beta}_{n,P})}{\partial \beta \partial \beta'}$.

Lemma D.1 *Let h_o be real-valued and estimated via a linear sieve M estimation using the sieve $\mathcal{H}_{k(n)} = \left\{ P_{k(n)}(\cdot)' \beta : \beta \in \mathbb{R}^{k(n)} \right\}$. Then we can take $\mathcal{V}_{k(n)} = \{ v(\cdot) = P_{k(n)}(\cdot)' \gamma : \gamma \in \mathbb{R}^{k(n)} \}$. Further, under the parametric specification $h_o(\cdot) = P_K(\cdot)' \beta_{o,P}$ with $K = k(n)$, we have:*

$$\textit{(1)} \quad \widehat{\Delta}(Z, \widehat{h}_n)[\widehat{\mathbf{v}}_n^*(\cdot)] = \widehat{\Gamma}_{2,P,n} \left(\widehat{R}_{n,P} \right)^{-} \frac{\partial \varphi_P(Z, \widehat{\beta}_{n,P})}{\partial \beta},$$

and

$$\textit{(2)} \quad \widehat{S}_{i,n}^* = g_P(Z_i, \widehat{\theta}_{n,P}, \widehat{\beta}_{n,P}) + \widehat{\Gamma}_{2,P,n} \left(\widehat{R}_{n,P} \right)^{-} \frac{\partial \varphi_P(Z_i, \widehat{\beta}_{n,P})}{\partial \beta} = \widehat{S}_{i,P,n}.$$

Proof. Noting that $\widehat{h}_n(\cdot) = P_{k(n)}(\cdot)' \widehat{\beta}$ for some $\widehat{\beta}$, where $\widehat{\beta} = \widehat{\beta}_{n,P}$ as long as $k(n) = K$. Given

$\widehat{\beta} = \widehat{\beta}_{n,P}$, it is clear that we have $\widehat{\theta}_n = \widehat{\theta}_{n,P}$. We see that for $j = 1, ..., d_g$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g_j(Z_i,\widehat{\theta}_n,\widehat{h}_n)}{\partial h}\left[p_a\left(\cdot\right)\right] &= \frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial g_j(Z_i,\widehat{\theta}_n,P_{k(n)}\left(\cdot\right)'\widehat{\beta}+\tau p_a\left(\cdot\right))}{\partial\tau}\right|_{\tau=0}\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial g_j(Z_i,\widehat{\theta}_n,P_{k(n)}\left(\cdot\right)'\widehat{\beta})}{\partial h}p_a\left(\cdot\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial g_{j,P}(Z_i,\widehat{\theta}_{n,P},\widehat{\beta}_{n,P})}{\partial\beta_a}\end{aligned}$$

if we view $g_j\left(Z_i,\theta,h_n\right) = g_j(Z_i,\theta,P_{k(n)}\left(\cdot\right)'\beta) = g_{j,P}(Z_i,\theta,\beta)$ as a function in β instead. It follows that we can write

$$\begin{aligned}\widehat{\Gamma}_{2,j} = \Gamma_{2,j,n}(\widehat{\theta}_n,\widehat{h}_n)\left[P_{k(n)}\left(\cdot\right)\right] &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial g_j(Z_i,\widehat{\theta}_n,\widehat{h}_n)}{\partial h}[P_{k(n)}\left(\cdot\right)]\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial g_{j,P}(Z_i,\widehat{\theta}_{n,P},\widehat{\beta}_{n,P})}{\partial\beta'}.\end{aligned}$$

Denote $\widehat{\Gamma}_2 \equiv \left(\widehat{\Gamma}'_{2,1},\ldots,\widehat{\Gamma}'_{2,d_g}\right)'$. Then

$$\widehat{\Gamma}_2 = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial g_P(Z_i,\widehat{\theta}_{n,P},\widehat{\beta}_{n,P})}{\partial\beta'} = \widehat{\Gamma}_{2,P,n}.$$

For sieve M estimation with $\varphi(Z,P_{k(n)}\left(\cdot\right)'\beta) = \varphi_P(Z,\beta)$, we have from definition (38),

$$\widehat{R}_{k(n)} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\varphi_P(Z_i,\widehat{\beta}_{n,P})}{\partial\beta\partial\beta'} = \widehat{R}_{n,P}.$$

We conclude from definition (37) that

$$\widehat{v}^*_{j,k(n)}\left(\cdot\right) = P_{k(n)}\left(\cdot\right)'(\widehat{R}_{k(n)})^{-}\Gamma_{2,j,n}(\widehat{\theta}_n,\widehat{h}_n)\left[P_{k(n)}\left(\cdot\right)\right] = \widehat{\Gamma}'_{2,j}(\widehat{R}_{n,P})^{-}P_{k(n)}(\cdot).$$

Recall that for sieve M estimation

$$\begin{aligned}\widehat{\Delta}(Z,\widehat{h}_n)\left[p_a\left(\cdot\right)\right] &= \left.\frac{\partial\varphi(Z,P_{k(n)}\left(\cdot\right)'\widehat{\beta}+\tau p_a\left(\cdot\right))}{\partial\tau}\right|_{\tau=0}\\ &= \frac{\partial\varphi(Z,P_{k(n)}\left(\cdot\right)'\widehat{\beta})}{\partial h}p_a\left(\cdot\right) = \frac{\partial\varphi_P(Z,\widehat{\beta}_{n,P})}{\partial\beta_a}.\end{aligned}$$

Thus

$$\widehat{\Delta}(Z,\widehat{h}_n)\left[\widehat{v}^*_{j,k(n)}\right] = \widehat{\Gamma}'_{2,j}(\widehat{R}_{n,P})^{-}\frac{\partial\varphi_P(Z,\widehat{\beta}_{n,P})}{\partial\beta}$$

and

$$\widehat{\Delta}(Z,\widehat{h}_n)\left[\widehat{\mathbf{v}}_n^*\right]=\left(\widehat{\Delta}(Z,\widehat{h}_n)\left[\widehat{v}_{1,k(n)}^*\right],...,\widehat{\Delta}(Z,\widehat{h}_n)\left[\widehat{v}_{d_g,k(n)}^*\right]\right)'$$
$$=\left(\widehat{\Gamma}_{2,1}',...,\widehat{\Gamma}_{2,d_g}'\right)'(\widehat{R}_{n,P})^{-}\frac{\partial\varphi_P(Z,\widehat{\beta}_{n,P})}{\partial\beta}=\widehat{\Gamma}_{2,P,n}(\widehat{R}_{n,P})^{-}\frac{\partial\varphi_P(Z,\widehat{\beta}_{n,P})}{\partial\beta}$$

thus Result (1) holds. Result (2) is trivially implied by Result (1), $(\widehat{\theta}_n,\widehat{\beta})=(\widehat{\theta}_{n,P},\widehat{\beta}_{n,P})$ and the definition of $\widehat{S}_{i,n}^*=g(Z_i,\widehat{\alpha}_n)+\widehat{\Delta}(Z_i,\widehat{h}_n)[\widehat{\mathbf{v}}_n^*]$ in (32). ■

Proof of Theorem 5.2. Since $\widehat{\Gamma}_{1,P}=\frac{1}{n}\sum_{i=1}^n\frac{\partial g_P(Z_i,\widehat{\theta}_{n,P},\widehat{\beta}_{n,P})}{\partial\theta'}=\frac{1}{n}\sum_{i=1}^n\frac{\partial g(Z_i,\widehat{\theta}_n,,\widehat{h}_n)}{\partial\theta'}=\widehat{\Gamma}_1$, the claimed result follows from Result (2) of Lemma D.1. ■

E Proof of the Results in Section 6

Assumption E.1 *(i) η_θ is a random vector with mean zero and variance I_{d_θ}, and independent of data $\{Z_i\}_{i=1}^n$; (ii) $\sup_{(\theta,h)\in\mathcal{N}_n}\|\sqrt{n}\mu_n[g_{\eta_\theta,n}(Z,\theta,h)]\|_E=o_p(1)$ with $g_{\eta_\theta,n}(Z,\theta,h)\equiv E_\eta\left[g(Z,\theta+n^{-\frac{1}{2}}\eta_\theta,h)\eta_\theta'\right]$; (iii) $\sup_{(\theta,h)\in\mathcal{N}_n}\|\Gamma_1(\theta,h)-\Gamma_1\|_E=o(1)$.*

Proof of Lemma 6.1. By definition, we can write

$$\begin{aligned}&E_\eta\left[D_{n,\theta}(\eta_\theta,\widehat{\alpha}_n)\eta_\theta'\right]-E_\eta\left[\Gamma_1\eta_\theta\eta_\theta'\right]\\&=E_\eta\left[\sqrt{n}\mu_n\left\{g(Z_i,\widehat{\theta}_n+n^{-\frac{1}{2}}\eta_\theta,\widehat{h}_n)-g(Z_i,\widehat{\theta}_n,\widehat{h}_n)\right\}\eta_\theta'\right]\\&+E_\eta\left[n^{\frac{1}{2}}\left(G(\widehat{\theta}_n+n^{-\frac{1}{2}}\eta_\theta,\widehat{h}_n)-G(\widehat{\theta}_n,\widehat{h}_n)-n^{-\frac{1}{2}}\Gamma_1\eta_\theta\right)\eta_\theta'\right],\end{aligned}\tag{E.1}$$

where $E_\eta[\Gamma_1\eta_\theta\eta_\theta']=\Gamma_1$ (since $E_\eta[\eta_\theta\eta_\theta']=I_{d_\theta}$). By Assumption E.1.(ii), we have

$$\begin{aligned}&\sup_{(\theta,h)\in\mathcal{N}_n}\left\|E_\eta\left[\sqrt{n}\mu_n\left\{g(Z,\theta+n^{-\frac{1}{2}}\eta_\theta,h)-g(Z,\theta,h)\right\}\eta_\theta'\right]\right\|_E\\&=\sup_{(\theta,h)\in\mathcal{N}_n}\left\|\sqrt{n}\mu_n\left[g_{\eta_\theta,n}(Z,\theta,h)\right]\right\|_E=o_p(1).\end{aligned}\tag{E.2}$$

By the differentiability of $G(\theta,h)$ in the local neighborhood of (θ_o,h_o), Assumptions E.1.(i) and (iii), we deduce that

$$\begin{aligned}&\left\|E_\eta\left[n^{\frac{1}{2}}\left(G(\widehat{\theta}_n+n^{-\frac{1}{2}}\eta_\theta,\widehat{h}_n)-G(\widehat{\theta}_n,\widehat{h}_n)-n^{-\frac{1}{2}}\Gamma_1\eta_\theta\right)\eta_\theta'\right]\right\|_E\\&\leq E_\eta\left[\sup_{(\theta,h)\in\mathcal{N}_n}\left\|\left[\Gamma_1(\theta,h)-\Gamma_1\right]\eta_\theta\eta_\theta'\right\|_E\right]\\&\leq E_\eta\left\|\eta_\theta\eta_\theta'\right\|_E\sup_{(\theta,h)\in\mathcal{N}_n}\|\Gamma_1(\theta,h)-\Gamma_1\|_E=o(1)\end{aligned}\tag{E.3}$$

with probability approaching 1. The result now follows from equations (E.1), (E.2) and (E.3). ■

Assumption E.2 *(i)* η_h *is a zero mean random vector with variance* I_{d_h}*, and independent of data* $\{Z_i\}_{i=1}^n$*; (ii)* $\sup_{\alpha\in\mathcal{N}_n,v\in\mathcal{W}_n}\|\sqrt{n}\mu_n[g_{\eta_h,n}(Z,\alpha,v)]\|_E = o_p(\delta_{w,n})$ *with* $g_{\eta_h,n}(Z,\alpha,v)\equiv E_\eta\left[g(Z,\theta,h+n^{-\frac{1}{2}}\eta_h v)\eta_h'\right]$*; (iii)* $\sup_{\alpha\in\mathcal{N}_n,v\in\mathcal{W}_n}\|\Gamma_2(\alpha)[v]-\Gamma_2(\alpha_o)[v]\|_E = O(\delta_{w,n})$*.*

Proof of Lemma 6.2. Let $E_\eta[D_{n,h}(\eta_h,\widehat{\alpha}_n,v)\eta_h']$ denote the expectation of $D_{n,h}(\eta_h,\widehat{\alpha}_n,v)\eta_h'$ with respect to the random vector η_h. Under Assumption E.2, we can follow the same proof as that of Lemma 6.1 to obtain

$$\sup_{\alpha\in\mathcal{N}_n,v\in\mathcal{W}_n}\left\|E_\eta\left[D_{n,h}(\eta_h,\alpha,v)\eta_h'\right]-\Gamma_2(\alpha_o)[v]\right\|_E = O_p(\delta_{w,n}).$$

This implies that Assumption A.3.(ii) is satisfied by the resampling estimate $\widehat{\Gamma}_{2,B}(\widehat{\theta}_n,\widehat{h}_n)[\cdot]$. ■

F Extra Simulation Results

Table F.1. IMSE of the Series Estimator $\widehat{h}_n$

	$\rho=0.00$	$\rho=0.25$	$\rho=0.50$	$\rho=0.75$
	$n=200$			
IMSE	0.0811	0.0843	0.0930	0.1226
	$n=500$			
IMSE	0.0460	0.0474	0.0513	0.0647
	$n=1000$			
IMSE	0.0296	0.0303	0.0323	0.0323

Notes: The simulation results are based on 100,000 replications.

Table F.2. Finite Sample Properties of the Two-step GMM Estimator

	$\rho = 0.00$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$
		$n = 200$		
Bias	0.0083	0.0092	0.0118	0.0191
Variance	0.0287	0.0314	0.0432	0.0955
MSE	0.0288	0.0315	0.0434	0.0961
		$n = 500$		
Bias	0.0035	0.0032	0.0043	0.0074
Variance	0.0105	0.0115	0.0158	0.0331
MSE	0.0105	0.0115	0.0158	0.0332
		$n = 1000$		
Bias	0.0018	0.0018	0.0024	0.0023
Variance	0.0051	0.0055	0.0075	0.0075
MSE	0.0051	0.0055	0.0075	0.0075

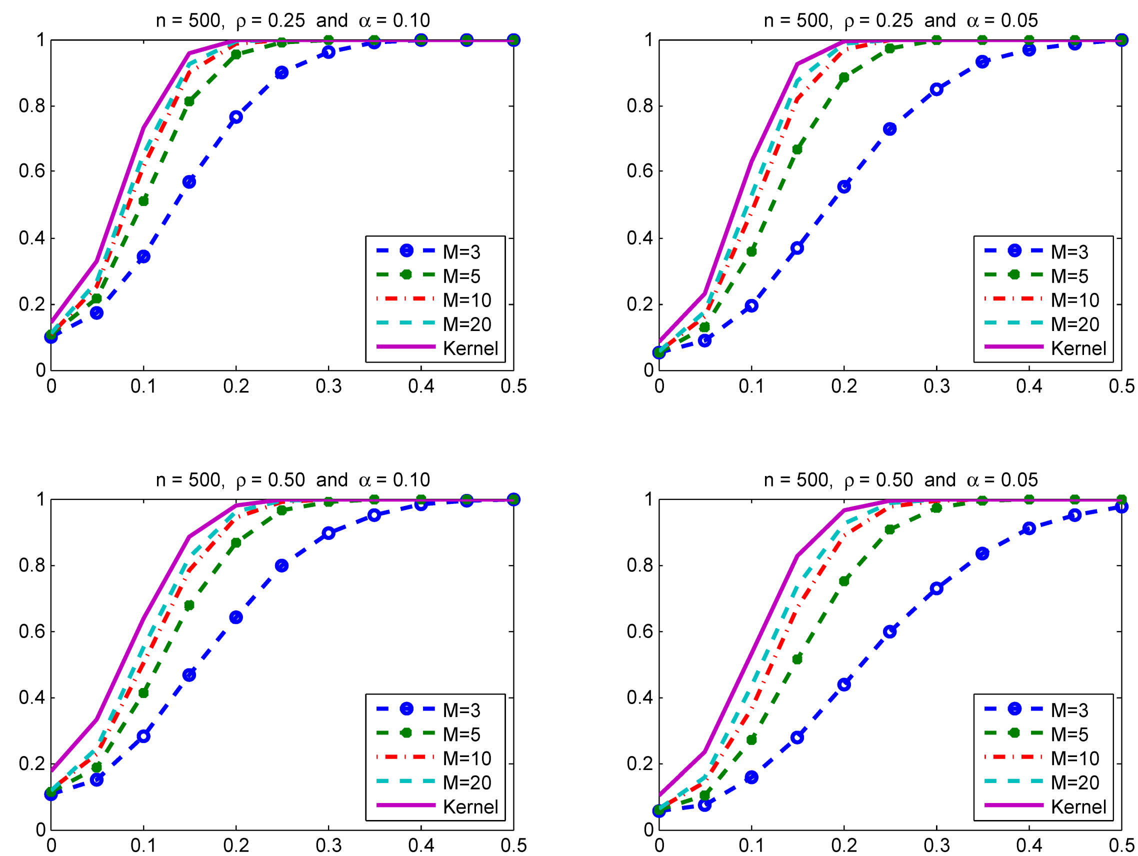
Notes: 1.The simulation results are based on 100,000 replications; 2. we compute the bias, variance and MSE of the GMM estimator of each component in θ_o and then take the averages to get the values of bias, variance and MSE in each row of the table.

Table F.3. Empirical Null Rejection Probabilities for Joint Test and Over-identification Test

	$\rho = 0.00$		$\rho = 0.25$		$\rho = 0.50$		$\rho = 0.75$	
	F-Test	J-Test	F-Test	J-Test	F-Test	J-Test	F-Test	J-Test
				$n = 1000$ and $M = 3$				
$\alpha = .1000$	.0996	.0969	.1015	.0966	.1043	.0968	.1077	.0985
$\alpha = .0500$	.0500	.0487	.0498	.0486	.0524	.0483	.0546	.0476
$\alpha = .0250$	.0246	.0245	.0249	.0246	.0265	.0234	.0273	.0229
$\alpha = .0025$	.0026	.0025	.0026	.0024	.0026	.0022	.0029	.0020
				$n = 1000$ and $M = 4$				
$\alpha = .1000$	.0999	.0964	.1014	.0969	.1046	.0975	.1116	.0973
$\alpha = .0500$	.0492	.0480	.0512	.0479	.0531	.0478	.0579	.0482
$\alpha = .0250$	.0242	.0233	.0265	.0241	.0266	.0240	.0293	.0236
$\alpha = .0025$	.0024	.0022	.0028	.0024	.0026	.0022	.0032	.0022
				$n = 1000$ and $M = 5$				
$\alpha = .1000$	.1004	.0971	.1015	.0969	.1045	.0980	.1160	.0994
$\alpha = .0500$	.0508	.0485	.0512	.0470	.0537	.0477	.0606	.0480
$\alpha = .0250$	.0261	.0238	.0263	.0233	.0273	.0233	.0307	.0232
$\alpha = .0025$	.0027	.0020	.0027	.0019	.0028	.0022	.0038	.0021

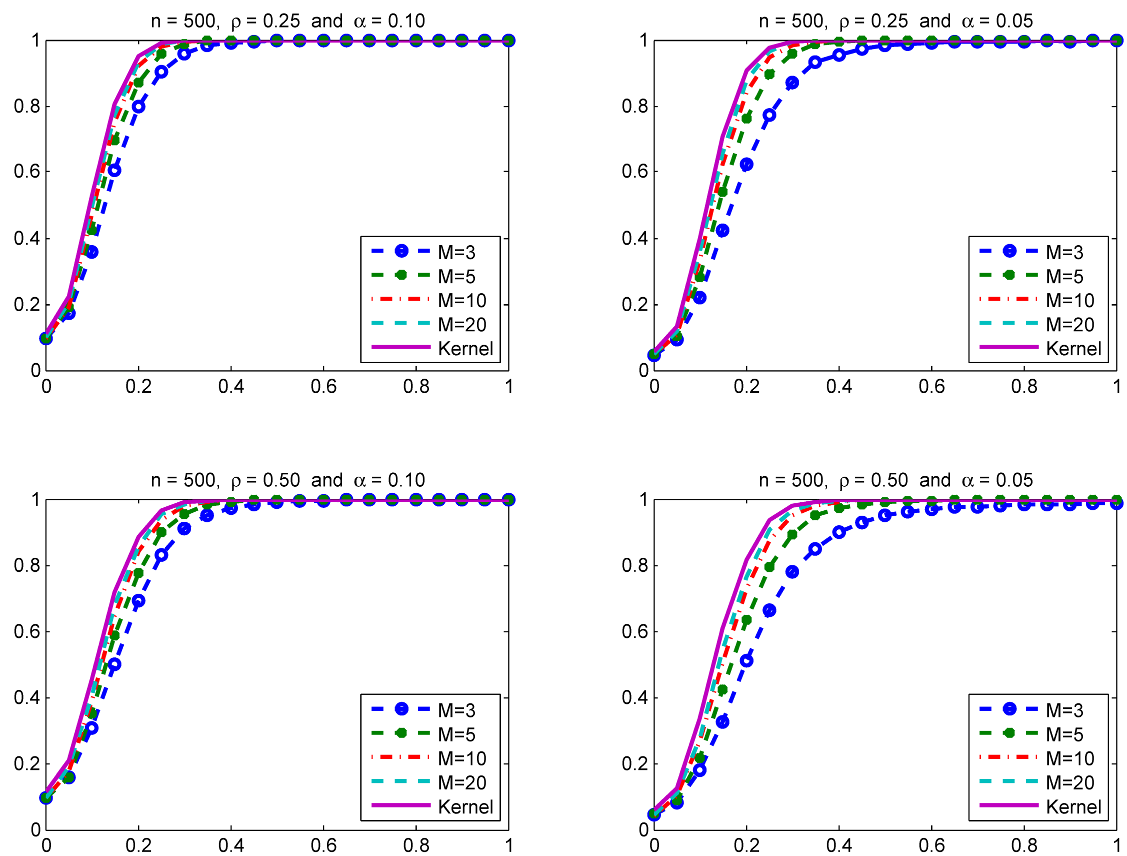
Notes: 1. The simulation results are based on 100,000 replications; 2. α denotes the nominal size of the test; 3. the F-Test refers to the test of the hypothesis in (70) using the test statistic $F_{R,n}$ and the asymptotic theory in (51); 4. the J-test refers to the test of the hypothesis in (71) using the test statistic $J_{R,n}$ in (48) and the asymptotic theory stated in Proposition 4.1.

Figure F.1. Empirical Power Functions of the Tests of Joint Hypothesis



Notes: 1. The simulation results are based on 10,000 replications; 2. α denotes the nominal size of the test; 3.the X-axis represents the value of θ_o and the Y-axis represents the rejection probability; 4. the curves denoted by "M=3", "M=5", "M=10" and "M=20" are the power functions of the joint tests based on the series LRV estimators with M=3, M=5, M=10 and M=20 respectively; 5. the curve denoted by "Kernel" is the power function of the Wald test based on kernel LRV estimator.

Figure F.2. Empirical Power Functions of the Over-identification Tests



Notes: 1. The simulation results are based on 10,000 replications; 2. α denotes the nominal size of the test; 3.the X-axis represents the value of η and the Y-axis represents the rejection probability; 4. the curves denoted by "M=3", "M=5", "M=10" and "M=20" are the power functions of the over-identification test based on the series LRV estimators with M=3, M=5, M=10 and M=20 respectively; 5. the curve denoted by "Kernel" is the power function of the over-identification test based on kernel LRV estimator.