Sieve M Inference on Irregular Parameters *

Xiaohong Chen † Zhipeng Liao ‡

January 2014

Abstract

This paper presents sieve inferences on possibly irregular (i.e., slower than root-n estimable) functionals of semi-nonparametric models with i.i.d. data. We provide a simple consistent variance estimator of the plug-in sieve M estimator of a possibly irregular functionals, which implies that the sieve t statistic is asymptotically standard normal. We show that, even for hypothesis testing of irregular functionals, the sieve likelihood ratio statistic is asymptotically Chi-square distributed. These results are useful in inference on structural parameters that may have singular semiparametric efficiency bounds. The proposed inference methods are investigated in a simulation study and an empirical example.

JEL Classification: C12, C14

Keywords: Irregular Functional, Sieve M Estimation, Sieve t Statistic, Sieve Likelihood Ratio, Zero Information

1 Introduction

The method of sieves (Grenander, 1981) is a general procedure for estimating semiparametric and nonparametric models, and is becoming increasingly popular in estimating complicated semiparametric structural models in economics. See Chen (2007) for a detailed review of the method and some well-known applications.

*We thank the guest Coeditor Norm Swanson, two anonymous referees, Zhiran Wang and the participants at Yale Prospectus Workshop in Econometrics, October 20, 2008 for helpful comments. Chen acknowledges financial support from National Science Foundation grant SES-0838161. 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, University of California, Los Angeles, 8279 Bunche Hall, Mail Stop: 147703, Los Angeles, CA 90095. Tel: 1 310 794 5427; fax: 1 310 825 9528. Email: zhipeng.liao@econ.ucla.edu

In this paper we consider sieve inference on possibly *irregular* (i.e., slower than $\sqrt{n}$ estimable) functionals of semi-nonparametric models with independent and identically distributed (i.i.d.) data. We focus on sieve M estimation of a pseudo true parameter belonging to an infinite dimensional parameter space, which is a procedure that optimizes a sample average of a criterion over a sequence of finite dimensional sieve spaces that becomes dense in the original infinite dimensional parameter space. Different choices of the criterion functions and of the sieve approximations lead to different examples of sieve M estimation, such as sieve Maximum Likelihood (ML), sieve Quasi Maximum Likelihood (QML), sieve Least Square (LS), sieve Generalized Least Square (GLS), sieve Quantile Regression (QR), to name only a few.

Asymptotic properties of general sieve M estimators, such as the consistency and the convergence rate of the nonparametric part, and the asymptotic normality of plug-in sieve estimators of *regular* (i.e., $\sqrt{n}$ estimable) functionals and their consistent variance estimators, are already established for i.i.d. data (see Shen (1997), Chen (2007) and the references therein). However, in many empirical works, applied researchers have difficulties to verify whether a functional of their specific semi-nonparametric model is $\sqrt{n}$ estimable or not. A necessary (but not sufficient) condition for a functional to be estimable at a $\sqrt{n}$ rate is that its semiparametric efficiency bound is positive; see, e.g., Bickel et al (1993), Newey (1990) and van der Vaart (1991). For some functionals (even Euclidean parameters) in complicated semi-nonparametric models, such as semiparametric mixture models, semiparametric nonlinear measurement error models and models containing multiple unknown functions, it is difficult and sometimes impossible to solve the semiparametric efficiency bound in a closed form. Without looking at a closed form expression of its efficiency bound, there is no easy way to verify if a functional is $\sqrt{n}$ estimable or not. It is thus important and desirable to provide simple valid inference procedures that do not require such a prior knowledge.

In a recent work, Chen, Liao and Sun (2014) (CLS) provides a general theory on the asymptotic normality of plug-in sieve M estimators of possibly irregular functionals for time series semi-nonparametric models. Their asymptotic normality result is rate-adaptive in the sense that researchers do not need to know whether a functional of interest is $\sqrt{n}$ estimable or not. For potentially misspecified dynamic semi-nonparametric time series models, CLS provides an *inconsistent* long-run variance estimator and shows that the corresponding sieve t statistic and sieve Wald statistic are asymptotically Student's t and F distributed respectively. Our paper complements CLS in several ways. First, we specialize their general asymptotic normality theorem to models with i.i.d. data, and verify their regularity conditions using two non-trivial semi-nonparametric examples. Second, we provide a simple *consistent* variance estimator of the plug-in sieve M estimator of a possibly irregular functional, and show that the corresponding sieve t statistic is asymptotically standard normal. Third, for correctly specified semi-nonparametric models, we show that the sieve Likelihood Ratio (LR) statistic is asymptotically Chi-square distributed, which provides another

way to construct confidence sets for possibly irregular functionals. Fourth, we apply our results to sieve ML inference on a widely used duration model with nonparametric unobserved heterogeneity of Heckman and Singer (1984), in which a structural parameter is irregular. This application highlights the usefulness of our theory.

There are two leading classes of irregular functionals in the literature on semiparametric and nonparametric models. The first class consists of slower than $\sqrt{n}$ estimable Euclidean parameters, which includes the Euclidean parameters with zero semiparametric information bounds. The second class consists of evaluation functionals of unknown functions (excluding distribution functions of observed random variables). Our paper provides valid inferences for both classes of irregular functionals of semi-nonparametric models via the method of sieve M estimation.

There are already many papers on slower than $\sqrt{n}$ estimable Euclidean parameters of semiparametric models. See, e.g., Chamberlain (1986, 2010), Honoré (1990, 1994), Horowitz (1992), Hahn (1994), Powell (1994), Andrews and Schafgans (1998), Ishwaran (1996, 1999), Ridder and Woutersen (2003), Khan and Tamer (2010), Graham and Powell (2012), Khan (2012), Khan and Nekipelov (2012), to list only a few. However, all of the existing published inference results on this class of irregular functionals are model (or problem) specific, while the sieve M inference results in our paper are applicable to a wide class of semi-nonparametric models.

There are also plenty papers on pointwise asymptotic normality of linear sieve (or series) M estimators of unknown conditional mean and density functions. See, e.g., Andrews (1991), Eastwood and Gallant (1991), Gallant and Souza (1991), Newey (1997), Zhou et al (1998), Huang (2003) and others for their respectively series LS estimators of the conditional mean functions, Stone (1990) on log-spline density estimator, to mention only a few. However, to the best of our knowledge, there is no published work on inference of general sieve M estimators of unknown functions.

The rest of the paper is organized as follows. Section 2 defines the plug-in sieve M estimators and specializes the asymptotic normality theorem of CLS to models with i.i.d. data. It also demonstrates that the regularity conditions for the asymptotic normality are verifiable via a partially additive QR example. Section 3 provides simple consistent variance estimators of the plug-in sieve M estimators. It shows a numerical equivalence result that allows applied researchers to compute the consistent variance estimators as if the model were parametric. Section 4 establishes that the sieve LR statistic is asymptotically Chi-square distributed regardless of whether the functional is regular or not. Section 5 applies the general theory to sieve ML estimation of and inference on the semiparametric duration model of Heckman and Singer (1984). Section 6 presents a simulation study. Section 7 conducts an empirical application of the Heckman and Singer model to the duration analysis of the second child birth in China. Section 8 briefly concludes. Proofs, technical derivations and tables are included in the Appendix.

Notation. We use "$\equiv$" to implicitly define a term or introduce a notation. For any column

vector A, we let A' denote its transpose and $||A||_E$ its Euclidean norm. Denote $L^p(\Omega,\mu)$, $1 \leq p < \infty$, as a space of real-valued measurable functions g with $||g||_{L^p(\mu)} \equiv \{\int_{x\in\Omega} |g(x)|^p d\mu(x)\}^{1/p} < \infty$, where Ω is the support of the sigma-finite positive measure $\mu()$. Let $||g||_{\sup} \equiv ||g||_\infty \equiv \sup_{x\in\Omega} |g(x)|$. Let $(\mathcal{G}, ||\cdot||_{\mathcal{G}})$ be a subset of a metric space of real-valued functions $g : \mathcal{X} \to \mathbb{R}$ on some set. The *covering number with bracketing* $N_{[]}(\varepsilon, \mathcal{G}, ||\cdot||_{\mathcal{G}})$ is the minimal number of N for which there exist $\varepsilon-$brackets $\{[l_j, u_j] : ||l_j - u_j||_{\mathcal{G}} \leq \varepsilon,\ ||l_j||_{\mathcal{G}}, ||u_j||_{\mathcal{G}} < \infty,\ j = 1, \ldots, N\}$ to cover $\mathcal{G}$ (i.e., for each $g \in \mathcal{G}$, there is a $j = j(g) \in \{1, \ldots, N\}$ such that $l_j \leq g \leq u_j$). For any two real numbers a and b, $a \lesssim b$ means $a \leq Cb$ where C denotes a generic positive and finite constant. For any (possibly random) positive sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, $a_n = O_P(b_n)$ means that $\lim_{c\to\infty} \limsup_n \Pr(a_n/b_n > c) = 0$; $a_n = o_P(b_n)$ means that for all $\varepsilon > 0$, $\lim_{n\to\infty} \Pr(a_n/b_n > \varepsilon) = 0$; $a_n \lesssim b_n$ and $a_n \asymp b_n$ respectively mean that there exist two constants $0 < c_1 \leq c_2 < \infty$ such that $c_1 a_n \leq b_n$ and $c_1 a_n \leq b_n \leq c_2 a_n$. For a positive sequence $\{c_n\}_{n=1}^\infty$ we sometimes use $c_n \nearrow c$ ($c_n \searrow c$) to mean that the sequence is increasing (decreasing) and converges to c.

2 Plug-in Sieve M Estimation and Asymptotic Normality

2.1 Plug-in Sieve M Estimators

We assume that the data $\{Z_i\}_{i=1}^n$ is a random sample from the distribution of Z defined on a underlying complete probability space. Let the support of Z be $\mathcal{Z} \subseteq \mathbb{R}^{d_z}, 1 \leq d_z < \infty$. Let $(\mathcal{A}, d_A)$ denote an infinite dimensional metric space. Let $\ell(\cdot,\cdot) : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$ be a measurable function and $E[\ell(Z,\alpha)]$ be a population criterion. For simplicity we assume that there is a unique $\alpha_0 \in (\mathcal{A}, d_A)$ such that $E[\ell(Z,\alpha_0)] > E[\ell(Z,\alpha)]$ for all $\alpha \in (\mathcal{A}, d_A)$ with $d_A(\alpha,\alpha_0) > 0$. Different models in economics correspond to different choices of the criterion function $E[\ell(Z,\alpha)]$ and the parameter space $(\mathcal{A}, d_A)$. A model does not need to be correctly specified and α_0 could be a pseudo-true parameter. In this paper we are interested in estimation of and inference on a functional $f(\alpha_0)$ via the method of sieves.

Let $\mathcal{A}_n$ be a sieve space for the whole parameter space $\mathcal{A}$. Then an *approximate sieve M estimator* $\widehat{\alpha}_n \in \mathcal{A}_n$ of α_0 solves

$$\frac{1}{n}\sum_{i=1}^n \ell(Z_i, \widehat{\alpha}_n) \geq \sup_{\alpha\in\mathcal{A}_n} \frac{1}{n}\sum_{i=1}^n \ell(Z_i,\alpha) - o_P(\frac{1}{n}). \tag{2.1}$$

We call $f(\widehat{\alpha}_n)$ the *plug-in sieve M estimator* of $f(\alpha_0)$.

The method of sieve M estimation includes many special cases. Different choices of criterion functions $\ell(Z_i,\alpha)$ and different choices of sieves $\mathcal{A}_n$ lead to different examples of sieve M estimation; see Chen (2007). In this paper we shall provide two non-trivial examples to demonstrate the usefulness of our new results: a sieve M estimation of a partially additive QR in Subsection 2.3,

and a sieve ML estimation of and inference on a semiparametric duration model of Heckman and Singer (1984) in Section 5.

Rates of Convergence of Sieve M estimators Under very mild conditions, the general sieve M estimator $\widehat{\alpha}_n$ is consistent for α_0: $d_A(\widehat{\alpha}_n, \alpha_0) = o_P(1)$; see, e.g., White and Wooldrige (1991) and Chen (2007, Theorem 3.1 and Remark 3.3). For a small $\epsilon > 0$, let

$$\mathcal{A}(\epsilon) \equiv \{\alpha \in \mathcal{A} : d_A(\alpha, \alpha_0) < \epsilon\} \quad \text{and} \quad \mathcal{A}_n(\epsilon) \equiv \mathcal{A}(\epsilon) \cap \mathcal{A}_n \tag{2.2}$$

denote ϵ-neighborhoods of α_0 under d_A-metric. Let $\|\alpha - \alpha_0\|_s$ be a pseudo metric on $\mathcal{A}$ such that $\|\alpha - \alpha_0\|_s \lesssim d_A(\alpha, \alpha_0)$ and $\|\alpha - \alpha_0\|_s^2 \asymp E[\ell(Z, \alpha_0) - \ell(Z, \alpha)]$ on $\mathcal{A}(\epsilon)$. The convergence rate for general sieve M estimators can be found in Shen and Wong (1994), Birge and Massart (1998), van de Geer (2000), Chen (2007, Theorem 3.2) and the references therein. In particular, suppose that $\mathcal{A}_n$ is a finite dimensional *linear* sieve with $\dim(\mathcal{A}_n) \asymp k_n = o(n)$ and that there is a $\pi_n(\alpha_0) \in \mathcal{A}_n$ with $\|\pi_n(\alpha_0) - \alpha_0\|_s \lesssim d_A(\pi_n(\alpha_0), \alpha_0) = o(1)$. Let $\delta_{s,n}^*$ and $\delta_{A,n}^*$ be positive sequences going to zero as n goes to infinity. Then, under mild additional conditions stated in these papers (see, e.g., Theorem 3.2 in Chen 2007), we obtain:

$$\|\widehat{\alpha}_n - \alpha_0\|_s \lesssim O_P\left(\max\left\{\sqrt{\frac{k_n}{n}}, \|\pi_n(\alpha_0) - \alpha_0\|_s\right\}\right) = O_P(\delta_{s,n}^*) = o_P(1). \tag{2.3}$$

Denote $\xi_n \equiv \sup_{\alpha \in \{\mathcal{A}_n : \|\alpha - \pi_n(\alpha_0)\|_s \neq 0\}} \{d_A(\alpha, \pi_n(\alpha_0)) / \|\alpha - \pi_n(\alpha_0)\|_s\}$. Then we have:

$$d_A(\widehat{\alpha}_n, \alpha_0) \lesssim O_P\left(\max\left\{\xi_n\sqrt{\frac{k_n}{n}}, d_A(\pi_n(\alpha_0), \alpha_0)\right\}\right) = O_P\left(\delta_{A,n}^*\right) = o_P(1). \tag{2.4}$$

For example, if $\mathcal{A}$ is a Hölder, Sobolev or Besov space of functions with bounded supports and $\mathcal{A}_n$ is a linear sieve space consisting of spline, wavelet or cosine bases, then one typically has $d_A(\pi_n(\alpha_0), \alpha_0) = \|\pi_n(\alpha_0) - \alpha_0\|_{\sup} \asymp k_n^{-\gamma}$, $\|\pi_n(\alpha_0) - \alpha_0\|_s = ||\pi_n(\alpha_0) - \alpha_0||_{L^2(P)} \lesssim k_n^{-\gamma}$ for some $\gamma > 0$, and $\xi_n \asymp \sqrt{k_n}$. See, e.g., Newey (1997), Huang (1998) and Chen (2007).

Given the existing results on the convergence rates for sieve M estimators of semi-nonparametric models, we can restrict our attention to a shrinking neighborhood of α_0. Let $\delta_{A,n} = \delta_{A,n}^* \gamma_n$ and $\delta_{s,n} = \delta_{s,n}^* \gamma_n$ where γ_n is a positive sequence that diverges to infinity very slowly (say $\log\log n$) such that $\delta_{A,n} = o(1)$. Then $\widehat{\alpha}_n \in \mathcal{N}_n \subseteq \mathcal{N}_0$ with probability approaching one (w.p.a.1), where

$$\mathcal{N}_0 \equiv \{\alpha \in \mathcal{A} : \|\alpha - \alpha_0\|_s \leq \delta_{s,n}, d_A(\alpha, \alpha_0) \leq \delta_{A,n}\} \subset \mathcal{A}(\epsilon);\ \mathcal{N}_n \equiv \mathcal{N}_0 \cap \mathcal{A}_n. \tag{2.5}$$

2.2 Asymptotic Normality

In this subsection we focus on asymptotic normality of plug-in linear sieve M estimators of possibly irregular functionals.

We suppose that for all $\alpha \in \mathcal{A}(\epsilon)$, $\ell(Z,\alpha) - \ell(Z,\alpha_0)$ can be approximated by $\Delta(Z,\alpha_0)[\alpha - \alpha_0]$ such that $\Delta(Z,\alpha_0)[\alpha - \alpha_0]$ is linear in $\alpha - \alpha_0$. When $\ell(Z,\alpha)$ is pathwise differentiable at α_0, we could let $\Delta(Z,\alpha_0)[\alpha-\alpha_0] = \lim_{\tau\to 0}[(\ell(Z,\alpha_0+\tau[\alpha-\alpha_0]) - \ell(Z,\alpha_0))/\tau]$, which is called the pathwise (directional) derivative of $\ell(Z,\alpha)$ at α_0 in the direction $[\alpha - \alpha_0]$. By the fact that α_0 is the unique maximizer of $E[\ell(Z,\alpha)]$ on $\mathcal{A}$, we could let

$$-\frac{\partial E\left[\Delta(Z,\alpha_0+\tau[\alpha-\alpha_0])[\alpha-\alpha_0]\right]}{\partial\tau}|_{\tau=0} \equiv \|\alpha-\alpha_0\|^2, \tag{2.6}$$

which defines a norm on $\mathcal{A}(\epsilon)$. One could typically choose the pseudo-metric $\|\cdot\|_s$ in such a way that is stronger than the metric $\|\cdot\|$ locally around α_0.

Let $\mathcal{V}$ be the closed linear span of $\mathcal{A}(\epsilon) - \{\alpha_0\}$ under $\|\cdot\|$, which is a Hilbert space under $\|\cdot\|$, with the corresponding inner product $\langle\cdot,\cdot\rangle$ defined as

$$\langle v_1, v_2\rangle = -\frac{\partial E\left[\Delta(Z,\alpha_0+\tau[v_2])[v_1]\right]}{\partial\tau}|_{\tau=0} \text{ for any } v_1, v_2 \in \mathcal{V}.$$

Let $\alpha_{0,n} \equiv \arg\min_{\alpha\in\mathcal{A}_n(\epsilon)} \|\alpha-\alpha_0\|$ and $\mathcal{V}_n$ be the closed linear span of $\mathcal{A}_n(\epsilon) - \{\alpha_{0,n}\}$ under $\|\cdot\|$. Then $\mathcal{V}_n$ is a finite dimensional Hilbert space under $\|\cdot\|$. To simplify presentation, we assume that $\dim(\mathcal{V}_n) = \dim(\mathcal{N}_n) = \dim(\mathcal{A}_n)$, all of which grow to infinity with n.

For any $v \in \mathcal{V}$, we define $\frac{\partial f(\alpha_0)}{\partial\alpha}[v]$ to be the pathwise derivative of $f(\cdot)$ at α_0 and in the direction of $v = \alpha - \alpha_0 \in \mathcal{V}$:

$$\frac{\partial f(\alpha_0)}{\partial\alpha}[v] = \left.\frac{\partial f(\alpha_0+\tau v)}{\partial\tau}\right|_{\tau=0} \text{ for any } v \in \mathcal{V}.$$

By the Riesz Representation Theorem (RRT), $\frac{\partial f(\alpha_0)}{\partial\alpha}[\cdot]$ has a Riesz representer v_n^* on $\mathcal{V}_n$:

$$\frac{\partial f(\alpha_0)}{\partial\alpha}[v] = \langle v_n^*, v\rangle \quad \text{for all } v \in \mathcal{V}_n \tag{2.7}$$

and that

$$\frac{\partial f(\alpha_0)}{\partial\alpha}[v_n^*] = ||v_n^*||^2 = \sup_{v\in\mathcal{V}_n, ||v||\neq 0} \frac{|\frac{\partial f(\alpha_0)}{\partial\alpha}[v]|^2}{||v||^2} < \infty. \tag{2.8}$$

We call v_n^* the *sieve Riesz representer* of the linear functional $\frac{\partial f(\alpha_0)}{\partial\alpha}[\cdot]$ on $\mathcal{V}_n$. It is obvious that $||v_n^*||$ is weakly increasing in $k_n \equiv \dim(\mathcal{V}_n)$. Since $\mathcal{V}_n$ becomes dense in $\mathcal{V}$ as $\dim(\mathcal{V}_n)$ increases to infinity, we have

$$\lim_{k_n\to\infty} ||v_n^*|| = \sup_{v\in\mathcal{V}, ||v||\neq 0} \frac{|\frac{\partial f(\alpha_0)}{\partial\alpha}[v]|}{||v||}.$$

Therefore, $\frac{\partial f(\alpha_0)}{\partial\alpha}[\cdot]$ has a Riesz representer v^* on $\mathcal{V}$ (i.e., $||v^*|| = \sup_{v\in\mathcal{V}, v\neq 0}\left\{|\frac{\partial f(\alpha_0)}{\partial\alpha}[v]|/||v||\right\} < \infty$) if and only if $\lim_{k_n\to\infty} ||v_n^*|| < \infty$.

For any $v \in \mathcal{V}$ we could define another pseudo - norm

$$\|v\|_{sd} \equiv \sqrt{Var\left(\Delta(Z,\alpha_0)[v]\right)} \tag{2.9}$$

whenever it is finite. For a semiparametric model that may not be a correctly specified likelihood model, $\|v\| \neq \|v\|_{sd}$ in general. But, when $\ell(Z,\alpha)$ is a log-likelihood function and $\Delta(Z,\alpha_0)[v] = \frac{\partial \ell(Z,\alpha_0)}{\partial \alpha}[v]$, we have $\|v\| = \|v\|_{sd}$ being the Fisher norm, and in this case

$$\lim_{k_n \to \infty} ||v_n^*|| = \sup_{v \in \mathcal{V}, \|v\|_{sd} \neq 0} \frac{\left|\frac{\partial f(\alpha_0)}{\partial \alpha}[v]\right|}{\|v\|_{sd}} = \|v^*\|_{sd}$$

where the right hand side corresponds to the calculation of semiparametric efficiency bound (or information bound) for a semiparametric likelihood model (see, e.g., van der Vaart (1991), Wong and Severini (1991) and Shen (1997)). In this case, the functional $f : \mathcal{A} \to \mathbb{R}$ is regular (i.e., root-n estimable or positive information bound) iff $\lim_{k_n \to \infty} ||v_n^*|| = \|v^*\| < \infty$ and irregular (i.e., slower than root-n estimable or zero information bound) iff $\lim_{k_n \to \infty} ||v_n^*|| = \infty$. For example, if we are interested in the finite dimensional parameter $\theta_0 \in int(\Theta)$ in a semiparametric likelihood model $\{\ell(Z,\alpha) = \log p(Z,\alpha) : \alpha = (\theta, h) \in \mathcal{A} = \Theta \times \mathcal{H}\}$, then for all $\lambda \neq 0$,

$$\begin{aligned} ||v_n^*||^2 &= \sup_{v=(v_\theta, v_h) \in \mathcal{V}_n, ||v|| \neq 0} \frac{|\lambda' v_\theta|^2}{E\left(\left\{\frac{\partial \log p(Z,\alpha_0)}{\partial \theta'} v_\theta + \frac{\partial \log p(Z,\alpha_0)}{\partial h}[v_h]\right\}^2\right)} \\ &= \lambda' \left(E[\mathcal{S}_{k_n} \mathcal{S}_{k_n}']\right)^{-1} \lambda, \end{aligned} \tag{2.10}$$

and $\lim_{k_n \to \infty} ||v_n^*||^2 = \lim_{k_n \to \infty} \lambda' \left(E[\mathcal{S}_{k_n} \mathcal{S}_{k_n}']\right)^{-1} \lambda = \lambda' \left(E[\mathcal{S}\mathcal{S}']\right)^{-1} \lambda < \infty$ iff the information bound $E[\mathcal{S}\mathcal{S}']$ is positive definite, where $\mathcal{S}$ is the semiparametric efficient score for θ.

The following regularity conditions are directly from CLS.

Assumption 2.1 (local property of functional) *(i) $v \mapsto \frac{\partial f(\alpha_0)}{\partial \alpha}[v]$ is a linear functional from $\mathcal{V}$ to $\mathbb{R}$, and $||v_n^*||/||v_n^*||_{sd} = O(1)$;*

$$\textit{(ii)} \quad \sup_{\alpha \in \mathcal{N}_n} \frac{\left|f(\alpha) - f(\alpha_0) - \frac{\partial f(\alpha_0)}{\partial \alpha}[\alpha - \alpha_0]\right|}{||v_n^*||} = o(n^{-\frac{1}{2}});$$

(iii) either (a) or (b) holds:

$$\begin{aligned} &\textit{(a)} \quad \|v_n^*\| \nearrow \infty \text{ and } \frac{\left|\frac{\partial f(\alpha_0)}{\partial \alpha}[\alpha_{0,n} - \alpha_0]\right|}{||v_n^*||} = o(n^{-\frac{1}{2}}); \\ &\textit{or (b)} \quad \|v_n^*\| \nearrow \|v^*\| < \infty \text{ and } ||v^* - v_n^*|| \times ||\alpha_{0,n} - \alpha_0|| = o(n^{-1/2}). \end{aligned}$$

Let $u_n^* \equiv \frac{v_n^*}{||v_n^*||_{sd}}$ and $\varepsilon_n = o(n^{-1/2})$. Let $\mu_n\{g(Z)\} \equiv n^{-1} \sum_{i=1}^n [g(Z_i) - Eg(Z_i)]$ denote the centered empirical process indexed by function g.

Assumption 2.2 (local behavior of criterion) *(i) The functional $\mu_n\{\Delta(Z,\alpha_0)[v]\}$ is linear in $v \in \mathcal{V}$;*

$$\textit{(ii)} \quad \sup_{\alpha \in \mathcal{N}_n} \mu_n\{\ell(Z, \alpha \pm \varepsilon_n u_n^*) - \ell(Z,\alpha) - \Delta(Z,\alpha_0)[\pm \varepsilon_n u_n^*]\} = O_P(\varepsilon_n^2);$$

(iii) there is a sequence $\{\eta_n \geq 0\}$ *with* $\eta_n = o(1)$ *such that*

$$\sup_{\alpha \in \mathcal{N}_n} \left| E[\ell(Z,\alpha) - \ell(Z,\alpha \pm \varepsilon_n u_n^*)] - \frac{||\alpha \pm \varepsilon_n u_n^* - \alpha_0||^2 - ||\alpha - \alpha_0||^2}{2} (1 + O(\eta_n)) \right| = O(\varepsilon_n^2).$$

Assumption 2.3 (CLT) $\sqrt{n}\mu_n \{\Delta(Z,\alpha_0)[u_n^*]\} \to_d N(0,1)$ *where* $N(0,1)$ *is a standard normal distribution.*

The following lemma specializes Theorem 3.1 of CLS to models with i.i.d. data.

Lemma 2.1 (CLS) *Let Assumptions 2.1, 2.2 and 2.3 hold with i.i.d. data. Then*

$$\sqrt{n}\frac{f(\widehat{\alpha}_n) - f(\alpha_0)}{\|v_n^*\|_{sd}} = \sqrt{n}\mu_n \{\Delta(Z,\alpha_0)[u_n^*]\} + o_P(1) \to_d N(0,1). \tag{2.11}$$

Lemma 2.1 suggests that $\|v_n^*\|_{sd}^2 \equiv \text{Var}(\Delta(Z,\alpha_0)[v_n^*])$ is a sieve variance of the plug-in sieve M estimator $f(\widehat{\alpha}_n)$ of $f(\alpha_0)$. We show it can be consistently estimated in Section 3.

Discussion of Assumptions CLS already provided closed-form expressions of the sieve Riesz representer v_n^* defined in (2.8) and discussed how to verify Assumption 2.1 in details.

As we shall illustrate in the next subsection, the verification of Assumption 2.2 is similar to those in Shen (1997) and Chen (2007, subsection 4.2) for plug-in sieve M estimation of a regular functional. If $\ell(Z,\alpha)$ is pathwise twice differentiable in $\alpha \in \mathcal{N}_0$, then Assumptions 2.2(ii) and (iii) can be respectively replaced by the following Assumptions 2.2(ii)' and (iii)', with $\Delta(Z,\alpha)[v] = \frac{\partial \ell(Z,\alpha)}{\partial \alpha}[v]$:
Assumption 2.2(ii)' $\sup_{\alpha \in \mathcal{N}_n} \mu_n (\Delta(Z,\alpha)[u_n^*] - \Delta(Z,\alpha_0)[u_n^*]) = o_P(n^{-1/2})$.
Assumption 2.2(iii)' *there is a sequence* $\{\eta_n \geq 0\}$ *with* $\eta_n = o(1)$ *such that*[1]

$$\sup_{\alpha \in \mathcal{N}_n} |E(\Delta(Z,\alpha)[u_n^*] - \Delta(Z,\alpha_0)[u_n^*]) + \langle \alpha - \alpha_0, u_n^* \rangle (1 + O(\eta_n))| = o(n^{-1/2}).$$

Assumption 2.2(iii) is also implied by
Assumption 2.2(iii)" *there is a sequence* $\{\eta_n \geq 0\}$ *with* $\eta_n = o(1)$ *such that*

$$\sup_{\alpha \in \mathcal{N}_n} \left| E[\ell(Z,\alpha_0) - \ell(Z,\alpha)] - \frac{||\alpha - \alpha_0||^2}{2} (1 + O(\eta_n)) \right| = o(n^{-1}).$$

The stochastic equicontinuity Assumption 2.2(ii) or (ii)' can be easily verified by applying Theorem 3 in Chen, Linton and van Keilegom (2003) (CLvK) or Lemma 4.2 in Chen (2007) or other empirical process results in van der Vaart and Wellner (1996). Assumption 2.2(iii) or (iii)' or (iii)" can be verified via a (pathwise) second-order Taylor expansion of the centered population criterion function.

[1] Assumption 2.2(iii)' corrects a sign typo in Condition 4.3' of Chen (2007, page 5613). Chen thanks Herman Bierens for spotting the typo.

Assumption 2.3 is implied by the following Lindeberg condition (2.12):

Assumption 2.3' *Let* $\{Z_i\}_{i=1}^n$ *be i.i.d., either Lindeberg condition (1) or Lyapounov condition (2) holds for every* $\varepsilon > 0$:

$$(1) \quad \lim_{n\to\infty} E\left[|\Delta(Z,\alpha_0)[u_n^*]|^2 1\left\{|\{\Delta(Z,\alpha_0)[u_n^*]\}| > \varepsilon\sqrt{n}\right\}\right] = 0; \tag{2.12}$$

$$(2) \quad \lim_{n\to\infty} \frac{E\left[|\Delta(Z,\alpha_0)[u_n^*]|^2 \varrho\left(|\Delta(Z,\alpha_0)[u_n^*]|\right)\right]}{\varrho\left(\sqrt{n}\varepsilon\right)} = 0 \tag{2.13}$$

where $\varrho(\cdot)$ *is some positive strictly increasing function, e.g.,* $\varrho(x) = x^\kappa$ *for some* $\kappa > 0$ *or* $\varrho(x) = \log(1+x)$.

2.3 A Partially Additive Quantile Regression Example

Suppose that the i.i.d. data $\{Y_i, X_i' = (X_{0i}', X_{1i}, ..., X_{qi})\}_{i=1}^n$ is generated according to

$$Y_i = X_{0,i}'\theta_0 + \sum_{j=1}^q h_{j,0}(X_{j,i}) + U_i, \quad E[1\{U_i \le 0\}|X_i] = \tau \in (0,1), \tag{2.14}$$

where $\dim(X_0) = d_\theta$, $\dim(X_j) = 1$ for $j = 1, ..., q$, $\dim(X) = d_\theta + q$, and $\dim(Y) = 1$. Let $\alpha_0 = (\theta_0, h_0)$, where $\theta_0 \in \Theta$ and $h_0 = (h_{1,0}, ..., h_{q,0}) \in \mathcal{H} = \mathcal{H}_1 \times ... \times \mathcal{H}_q$, be the unknown true parameter values. A functional of interest could be: $f(\alpha_0) = \lambda'\theta_0$ for any $\lambda \in \mathbb{R}^{d_\theta}$ with $\lambda \neq 0$, or $f(\alpha_0) = h_{1,0}(\overline{x}_1)$ for some point $\overline{x}_1 \in (0,1)$. This is an extension of the parametric quantile regression model of Koenker and Bassett (1978) to allow for unknown additive functions $\sum_{j=1}^q h_{j,0}(X_{j,i})$. See Koenker (2005) for numerous other extensions.

We can estimate $\alpha_0 = (\theta_0, h_0)$ by the sieve QR estimator $\widehat{\alpha}_n = \left(\widehat{\theta}_n, \widehat{h}_n\right)$ that solves

$$\max_{\theta\in\Theta, h\in\mathcal{H}_n} \sum_{i=1}^n \left(1\left\{Y_i \le X_{0,i}'\theta + \sum_{j=1}^q h_j(X_{j,i})\right\} - \tau\right)\left[Y_i - X_{0,i}'\theta - \sum_{j=1}^q h_j(X_{j,i})\right], \tag{2.15}$$

where $\mathcal{H}_n = \mathcal{H}_{1,n} \times ... \times \mathcal{H}_{q,n}$ is a tensor product sieve space for $\mathcal{H} = \mathcal{H}_1 \times ... \times \mathcal{H}_q$. Then a functional of interest can be estimated by the plug-in sieve QR estimation: $f(\widehat{\alpha}_n) = \lambda'\widehat{\theta}_n$, or $f(\widehat{\alpha}_n) = \widehat{h}_{1,n}(\overline{x}_1)$.

For $j = 1, ..., q$, let $\mathcal{X}_j$, the support of X_j, be a bounded interval. Let $\mathcal{X}_0$ be a bounded set denoting the support of X_0, and $\mathcal{X} = \mathcal{X}_0 \times \mathcal{X}_1 \times ... \times \mathcal{X}_q$ be the support of $X = (X_0', X_1, ..., X_q)'$. Since $h_{1,0}(\cdot)$ can have a constant we assume that X_0 does not contain the constant regressor. For $j = 1, ..., q$, we approximate $h_{j,0}$ by $h_{j,n}(\cdot) \equiv P_{k_{jn}}^j(\cdot)'\beta_{0,k_{jn}}^j$ where $P_{k_{jn}}^j(\cdot)' = \left(p_1^j(\cdot), ..., p_{k_{jn}}^j(\cdot)\right)$ denotes a $1 \times k_{jn}-$vector of known sieve basis functions. Then $P_{k_n}(x)'\beta_{0,k_n}$ is the linear sieve approximation of $\sum_{j=1}^q h_{j,0}(x_j)$, where

$$P_{k_n}(x)' = \left(P_{k_{1n}}^1(x_1)', ..., P_{k_{qn}}^q(x_q)'\right) \text{ and } \beta_{0,k_n}' = (\beta_{0,k_{1n}}^{1\prime}, ..., \beta_{0,k_{qn}}^{q\prime}).$$

Let $\Lambda^\gamma([a,b])$ denote the Hölder space of functions on $[a,b]$ with smoothness γ, and $\Lambda^\gamma_c([a,b])$ denote the following Hölder ball with radius c:

$$\left\{ h \in C^{[\gamma]}([a,b]) : \sup_{j \leq [\gamma]} \sup_{x \in \mathcal{X}} |\nabla^j h(x)| \leq c, \sup_{x,y \in [a,b], x \neq y} \frac{|\nabla^{[\gamma]} h(x) - \nabla^{[\gamma]} h(y)|}{|x-y|^{\gamma - [\gamma]}} \leq c \right\}$$

where $[\gamma]$ denotes the largest non-negative integer that is strictly smaller than γ (i.e., $0 \leq [\gamma] < \gamma$). The Hölder space $\Lambda^\gamma([0,1])$ with $\gamma > 1/2$, of functions can be well approximated in $\|\cdot\|_{\sup}$ norm by various linear sieves such as cosine series $\text{CosPol}(k_n) = \{\beta_0 + \sum_{k=1}^{k_n} \beta_k \cos(k\pi x)\}$, power series $\text{Pol}(k_n) = \{\sum_{k=0}^{k_n} \beta_k x^k\}$, splines $\text{Spl}(s, k_n)$ (with $s > [\gamma]$ denoting the order of the spline and k_n the number of knots), wavelets $\text{Wav}(s, k_n)$ (with $s > [\gamma]$) and others. See Chen (2007) for precise definition of these sieves.

Let $f(\cdot|x)$ (and $F(\cdot|x)$) be the conditional density (and cdf) of U given $X = x$. Denote $k_n = \max\{k_{jn} : j = 1, ..., q\}$.

Condition 2.1 *(i) Θ is a compact subset in $\mathbb{R}^{d_\theta}$ and $\mathcal{H}_j = \Lambda^{\gamma_j}_{c_j}(\mathcal{X}_j)$ with $\gamma_j > \frac{1}{2}$ and finite $c_j > 0$ for $j = 1, ..., q$, and $h_{j,0}(x^*_j) = 0$ for some known $x^*_j \in \mathcal{X}_j$ for $j = 2, ..., q$; (ii) for $j = 1, ..., q$, there is a $\pi_n(h_{j,0}) \in \mathcal{H}_{j,n} \subset \mathcal{H}_j$ such that $||\pi_n(h_{j,0}) - h_{j,0}||_{\sup} = O(k_{jn}^{-\gamma_j})$; (iii) $0 < \inf_{x \in \mathcal{X}} f(0|x) \leq \sup_{x \in \mathcal{X}} f(0|x) < \infty$; (iv) $\sup_{x \in \mathcal{X}} |f(u|x) - f(0|x)| \to 0$ as $|u| \to 0$; (v) the smallest eigenvalue of $E[\overline{P}_{k_n}(X)\overline{P}_{k_n}(X)']$ is bounded away from zero uniformly in k_n; where $\overline{P}_{k_n}(X) \equiv [X_0', P_{k_n}(X)']'$.*

In this example, $\ell(Z, \alpha) = [1\{Y \leq \alpha(X)\} - \tau][Y - \alpha(X)]$ with $\alpha(X) = X_0'\theta + \sum_{j=1}^q h_j(X_j)$. For any $\alpha \in \mathcal{A}$, we define a strong metric $\|\cdot\|_s$ as

$$\|\alpha - \alpha_0\|_s^2 = E\left[\left| X_0'(\theta - \theta_0) + \sum_{j=1}^q [h_j(X_j) - h_{j,0}(X_j)] \right|^2 \right].$$

Following the same proofs as those of Propositions 3.3 and 3.4 in Chen (2007) or that of Example 3.5 in Chen and White (1999), we immediately obtain the convergence rates of the sieve QR estimator.

Remark 2.2 *Let $\widehat{\alpha}_n$ be the sieve QR estimator defined in (2.15). Let Condition 2.1(i)-(iv) hold with $\gamma = \min\{\gamma_1, ..., \gamma_q\}$. Then: (1) $\|\widehat{\alpha}_n - \alpha_0\|_s = O_P\left(\max\left\{\sqrt{\frac{k_n}{n}}, k_n^{-\gamma}\right\}\right) = o_P(1)$, which has the best rate $\|\widehat{\alpha}_n - \alpha_0\|_s = O_P\left(n^{-\gamma/(2\gamma+1)}\right)$ with $k_n \asymp k^*_{n,s} \asymp n^{1/(2\gamma+1)}$. (2) Let $\xi_n = \max_{1 \leq j \leq q} \sup_{x_j \in \mathcal{X}_j} \left\| P^j_{k_{jn}}(x_j) \right\|_E$. If Condition 2.1(v) holds, then*

$$\|\widehat{\alpha}_n - \alpha_0\|_{\sup} = O_P\left(\max\left\{\xi_n\sqrt{\frac{k_n}{n}}, k_n^{-\gamma}\right\}\right) = o_P(1).$$

(3) Further, If the sieve basis $P_{k_n}(X)$ is cosine $\text{CosPol}(k_n)$, or spline $\text{Spl}(s, k_n)$ (with $s > [\gamma]$) or wavelet $\text{Wav}(s, k_n)$ (with $s > [\gamma]$) then Condition 2.1(ii) is satisfied and $\xi_n \asymp \sqrt{k_n}$. Hence $\|\widehat{\alpha}_n - \alpha_0\|_{\sup} = O_P\left(n^{-\gamma/(2\gamma+2)}\right)$ with $k_n \asymp k^*_{n,\sup} \asymp n^{1/(2\gamma+2)}$

Condition 2.3 *(i) There is a finite constant C such that $\sup_{x\in\mathcal{X}}|f(u_1|x)-f(u_2|x)|\leq C|u_1-u_2|$ for all u_1 and u_2 in a neighborhood of zero and $\gamma>1$; (ii) $n\times\max\left\{\frac{k_n}{n},k_n^{-2\gamma}\right\}\times\max\left\{\xi_n\sqrt{\frac{k_n}{n}},k_n^{-\gamma}\right\}=o(1)$.*

Let $\Delta(Z,\alpha_0)[\alpha-\alpha_0]=-\left(1\{U\leq 0\}-\tau\right)\left(\alpha(X)-\alpha_0(X)\right)$. By the definitions of the metrics (2.6) and (2.9), we have, for any $\alpha\in\mathcal{A}$,

$$\|\alpha-\alpha_0\|^2=E\left[f(0|X)\left|\alpha(X)-\alpha_0(X)\right|^2\right]\text{ and }\|\alpha-\alpha_0\|_{sd}^2=\tau(1-\tau)E\left[\left|\alpha(X)-\alpha_0(X)\right|^2\right].$$

Hence, for all $\alpha\in\mathcal{A}$, $\|\alpha-\alpha_0\|_{sd}\asymp\|\alpha-\alpha_0\|_s$, and $\|\alpha_1-\alpha_2\|\asymp\|\alpha-\alpha_0\|_s$ under Condition 2.1(iii).

Let $v_n^*=\left(v_{\theta,n}^*,v_{h,n}^*\right)$ be the sieve Riesz representer (2.8) of the functional $\frac{\partial f(\alpha_0)}{\partial\alpha}[v]$ on $\mathcal{V}_n$. Denote $v_n^*(X)=X_0'v_{\theta,n}^*+\sum_{j=1}^q v_{h_j,n}^*(X_j)$. Then the variance of the plug-in sieve M estimator $f(\widehat{\alpha}_n)$ of $f(\alpha_0)$ is

$$\|v_n^*\|_{sd}^2=\tau(1-\tau)E\left[|v_n^*(X)|^2\right]=\tau(1-\tau)\|v_n^*\|_s^2\asymp\|v_n^*\|^2.$$

Proposition 2.4 *Let Assumption 2.1(ii)(iii), Condition 2.1 and Condition 2.3 hold. Then:*

$$\sqrt{n}\frac{f(\widehat{\alpha}_n)-f(\alpha_0)}{\sqrt{\tau(1-\tau)E\left[\left|X_0'v_{\theta,n}^*+\sum_{j=1}^q v_{h_j,n}^*(X_j)\right|^2\right]}}\to_d N(0,1).$$

For ease of notation, we assume that there is only one real-valued unknown function in α_0, i.e. $\alpha_0=(\theta_0,h_0)$ with $h_0=h_{1,0}$ and $P_{k_n}(X)=P_{k_n}^1(X_1)$. Denote $I_{n,11}=E\left[f(0|X)X_0X_0'\right]$, $I_{n,22}=E\left[f(0|X)P_{k_n}(X_1)P_{k_n}(X_1)'\right]$ and $I_{n,12}=E\left[f(0|X)X_0P_{k_n}(X_1)'\right]=I_{n,21}'$. For the *evaluation* functional $f(\alpha)=h(\overline{x}_1)$ with $\overline{x}_1\in int(\mathcal{X}_1)$, its sieve Riesz representer v_n^* is

$$v_n^*=\left(v_{\theta,n}^*,v_{h,n}^*\right)=\left(-I_{n,11}^{-1}I_{n,12},P_{k_n}(\cdot)'\right)\beta_{h,n}^*$$

where $\beta_{h,n}^*=I_n^{22}P_{k_n}(\overline{x}_1)$ and $I_n^{22}=\left[I_{n,22}-I_{n,21}I_{n,11}^{-1}I_{n,12}\right]^{-1}$. Then

$$\begin{aligned}V_{\overline{x}_1}&\equiv E[|X_0'v_{\theta,n}^*+v_{h,n}^*(X_1)|^2]\\&=\beta_{h,n}^{*\prime}E\left[\left(P_{k_n}(X_1)-I_{n,21}I_{n,11}^{-1}X_0\right)\left(P_{k_n}(X_1)-I_{n,21}I_{n,11}^{-1}X_0\right)'\right]\beta_{h,n}^*\end{aligned}$$

As $f(\alpha)=h(\overline{x}_1)$ is a linear functional, Assumption 2.1(ii) is automatically satisfied. Assumption 2.1(iii) is satisfied given Condition 2.1(ii) and $nk_n^{-2\gamma}/V_{\overline{x}_1}=o(1)$. These and Proposition 2.4 immediately lead to the following result:

$$\sqrt{n}\frac{\widehat{h}_n(\overline{x}_1)-h_0(\overline{x}_1)}{\sqrt{\tau(1-\tau)V_{\overline{x}_1}}}\to_d N(0,1).\tag{2.16}$$

3 Consistent Sieve Variance Estimators

To apply Lemma 2.1 for Wald type inference on $f(\alpha_0)$, we want to estimate the sieve variance $||v_n^*||_{sd}^2$ *consistently*. In this section, we show that an empirical normalizer $\|\widehat{v}_n^*\|_{sd,n}$, constructed using an empirical sieve Riesz representer $\widehat{v}_n^*$ and an empirical norm $||\cdot||_{sd,n}$, is a consistent estimator of $||v_n^*||_{sd}$.

For any $v_1, v_2 \in \mathcal{V}$, we define $r(Z,\alpha)[v_1,v_2] \equiv \frac{\partial \Delta(Z,\alpha+\tau[v_1])[v_2]}{\partial \tau}$. Then it is clear that $\langle v_1, v_2\rangle = E\left[-r(Z,\alpha_0)[v_1,v_2]\right]$ for any $v_1, v_2 \in \mathcal{V}$. Denote an empirical norm $\|\cdot\|_n$ on $\mathcal{V}_n$ as

$$\|v\|_n^2 = -\frac{1}{n}\sum_{i=1}^n r(Z_i,\widehat{\alpha}_n)[v,v], \text{ for any } v \in \mathcal{V}_n. \tag{3.1}$$

We define an empirical sieve Riesz representer $\widehat{v}_n^*$ of the functional $\frac{\partial f(\widehat{\alpha}_n)}{\partial \alpha}[\cdot]$ as

$$\frac{\partial f(\widehat{\alpha}_n)}{\partial \alpha}[\widehat{v}_n^*] = \|\widehat{v}_n^*\|_n^2 = \sup_{v\in\mathcal{V}_n, \|v\|_n>0} \frac{\left|\frac{\partial f(\widehat{\alpha}_n)}{\partial \alpha}[v]\right|^2}{\|v\|_n^2} < \infty. \tag{3.2}$$

An empirical sieve variance $\|\widehat{v}_n^*\|_{sd,n}^2$ is then defined as

$$\|\widehat{v}_n^*\|_{sd,n}^2 = \frac{1}{n}\sum_{i=1}^n \left|\Delta(Z_i,\widehat{\alpha}_n)\left[\widehat{v}_n^*\right]\right|^2, \tag{3.3}$$

or

$$\|\widehat{v}_n^*\|_{sd,n}^2 = \frac{1}{n}\sum_{i=1}^n \left|\Delta(Z_i,\widehat{\alpha}_n)\left[\widehat{v}_n^*\right] - \frac{1}{n}\sum_{j=1}^n \Delta(Z_j,\widehat{\alpha}_n)\left[\widehat{v}_n^*\right]\right|^2.$$

Assumption 3.1 *Let* $\mathcal{W}_n \equiv \{v \in \mathcal{V}_n : \|v\| = 1\}$, *then:*
(i) $\sup_{\alpha\in\mathcal{N}_n, v_1,v_2\in\mathcal{W}_n} |E\left\{r(Z,\alpha)[v_1,v_2] - r(Z,\alpha_0)[v_1,v_2]\right\}| = o(1)$;
(ii) $\sup_{\alpha\in\mathcal{N}_n, v_1,v_2\in\mathcal{W}_n} |\mu_n\left\{r(Z,\alpha)[v_1,v_2]\right\}| = o_P(1)$;
(iii) $\sup_{\alpha\in\mathcal{N}_n, v\in\mathcal{W}_n} \left|\frac{\partial f(\alpha)}{\partial \alpha}[v] - \frac{\partial f(\alpha_0)}{\partial \alpha}[v]\right| = o(1)$.

Assumptions 3.1(i) imposes local smoothness restriction on $E\left\{r(Z,\alpha)[v_1,v_2]\right\}$ in α uniformly over $v_1, v_2 \in \mathcal{W}_n$. Assumptions 3.1(ii) is implied by the class $\mathcal{F}_n^{sd} = \{r(Z,\alpha)[v_1,v_2] : \alpha \in \mathcal{N}_n, v_1, v_2 \in \mathcal{W}_n\}$ being Glivenko-Cantelli, which is in turn implied by $N_{[]}(\varepsilon, \mathcal{F}, ||\cdot||_{L^1(P)}) < \infty$ (see, e.g., theorem 2.4.1 in van der Vaart and Wellner, 1996). Assumption 3.1(iii) assumes that the functional $\frac{\partial f(\alpha)}{\partial \alpha}[v]$ is smooth in $\alpha \in \mathcal{N}_n$ uniformly over $v \in \mathcal{W}_n$. It becomes a condition of $\sup_{\alpha\in\mathcal{N}_n} |f(\alpha) - f(\alpha_0)| = o(1)$ for a linear functional $f(\cdot)$.

Assumption 3.2 *(i)* $\|v\|_{sd} \asymp \|v\|$ *for all* $v \in \mathcal{V}_n$;
(ii) $\sup_{\alpha\in\mathcal{N}_n, v\in\mathcal{W}_n} \left|E\left[(\Delta(Z,\alpha)[v])^2 - (\Delta(Z,\alpha_0)[v])^2\right]\right| = o(1)$;
(iii) $\sup_{\alpha\in\mathcal{N}_n, v\in\mathcal{W}_n} \left|\mu_n\left[(\Delta(Z,\alpha)[v])^2\right]\right| = o_P(1)$.

Assumption 3.2(i) is stronger than Assumptions 2.1(i). In the sieve ML estimation, we have $\|v\|_{sd} = \|v\|$ for any $v \in \mathcal{V}_n$ and hence Assumption 3.2(i) is trivially satisfied. Assumptions 3.2(ii) and (iii) are similar to Assumptions 3.1(i) and (ii) respectively.

Lemma 3.1 *Under Assumptions 3.1 and 3.2, we have:* $\left| \|\widehat{v}_n^*\|_{sd,n} \Big/ \|v_n^*\|_{sd} - 1 \right| = o_P(1)$.

The following result directly follows from Lemma 2.1, Lemma 3.1 and the Slutsky Theorem.

Theorem 3.1 *Under Assumptions 2.1, 2.2, 2.3, 3.1 and 3.2, we have*

$$\frac{\sqrt{n}\left[f(\widehat{\alpha}_n) - f(\alpha_0)\right]}{\|\widehat{v}_n^*\|_{sd,n}} \to_d N(0,1). \tag{3.4}$$

Remark 3.2 *In the first draft of this paper, we show that in the nonparametric LS regression model*

$$Y = h_0(X) + U \text{ with } E[U|X] = 0, \tag{3.5}$$

the following conditions imply Assumptions 2.2, 2.3, 3.1(i)(ii) and 3.2: (i) $h_0(\cdot) \in \Lambda^\gamma(\mathcal{X})$ *for some* $\gamma > \frac{d}{2}$ *with* $d = \dim(X)$ *and there exists* $\pi_n(h_0)$ *such that* $||h_0 - \pi_n(h_0)||_{\sup} = O(k_n^{-\frac{\gamma}{d}})$*; (ii)* $Var[Y|X]$ *is bounded and bounded away from zero; (iii) the smallest eigenvalue of* $E[P_{k_n}(X)P_{k_n}(X)']$ *is bounded and bounded away from zero for all* k_n*; (iv)* $\xi_n\sqrt{\frac{k_n}{n}} + k_n^{-\frac{\gamma}{d}} = o(1)$*; where* $\xi_n = \sup_{x\in\mathcal{X}} ||P_{k_n}(x)||_E$*; (v) for any* $\varepsilon > 0$*,* $E\left[\frac{U^2 \varrho(|U|)}{\varrho(\varepsilon\sqrt{n})}\middle| X\right] \to 0$ *as* $n \to \infty$*, where* $\varrho(\cdot)$ *is some positive strictly increasing function. Hence, if the functional* $f(\cdot)$ *satisfies Assumptions 2.1 and 3.1(iii), then we can invoke Theorem 3.1 to deduce that*

$$\sqrt{n}\frac{f(\widehat{h}_n) - f(h_0)}{\sqrt{\widehat{A}_{k_n}'\widehat{Q}_{k_n}^{-1}\frac{\sum_{i=1}^n[\widehat{\sigma}_{n,i}^2 P_{k_n}(X_i)P_{k_n}(X_i)']}{n}\widehat{Q}_{k_n}^{-1}\widehat{A}_{k_n}}} \to_d N(0,1), \tag{3.6}$$

where $\widehat{A}_{k_n} = \frac{\partial f(\widehat{h}_n)}{\partial h}[P_{k_n}(\cdot)]$ *is a* $k_n \times 1$ *vector with its a-th element being* $\frac{\partial f(\widehat{h}_n)}{\partial h}[p_a(\cdot)]$ *(*$a = 1, ..., k_n$*),* $\widehat{\sigma}_{n,i}^2 = [Y_i - \widehat{h}_n(X_i)]^2$ *and* $\widehat{Q}_{k_n} = n^{-1}\sum_{i=1}^n [P_{k_n}(X_i)P_{k_n}(X_i)']$*. Compared with Newey (1997), the above sufficient conditions for the asymptotic normality (3.6) are weaker. For example, we do not impose the conditions that* $E[U^4|X]$ *being bounded and that* $\frac{\xi_n^4 k_n^2}{n} = o(1)$*. Also, Assumption 3.1(iii) is weaker than Newey (1997)'s condition* $\left|\frac{\partial f(h)}{\partial h}[v] - \frac{\partial f(h_0)}{\partial h}[v]\right| \lesssim ||v||_{\sup}||h - h_o||_{\sup}$ *for all* $v \in \mathcal{A}_n$*.*

In the sieve ML estimation, one has $\|v\|_{sd} = \|v\|$ for all $v \in \mathcal{V}_n$. We can compute an alternative sieve variance estimator as follows. Using the empirical norm $\|v\|_{sd,n} = \sqrt{\frac{1}{n}\sum_{i=1}^n \left(\Delta(Z_i, \widehat{\alpha}_n)[v]\right)^2}$, we define an empirical sieve Riesz representer $\widetilde{v}_n^*$ of the functional $\frac{\partial f(\widehat{\alpha}_n)}{\partial \alpha}[\cdot]$ as

$$\frac{\partial f(\widehat{\alpha}_n)}{\partial \alpha}[\widetilde{v}_n^*] = \|\widetilde{v}_n^*\|_{sd,n}^2 = \sup_{v\in\mathcal{V}_n, \|v\|_{sd,n}>0} \frac{\left|\frac{\partial f(\widehat{\alpha}_n)}{\partial \alpha}[v]\right|^2}{\|v\|_{sd,n}^2} < \infty. \tag{3.7}$$

We next show that in sieve ML estimation, under Assumption 3.3 (that is akin to Assumption 3.1), $\|\widetilde{v}_n^*\|_{sd,n}$ is also a consistent estimator of $||v_n^*||_{sd}$.

Assumption 3.3 *The following conditions are satisfied:*

(i) $\sup_{\alpha\in\mathcal{N}_n,v_1,v_2\in\mathcal{W}_n}|E\left[\Delta(Z,\alpha)[v_1]\Delta(Z,\alpha)[v_2]-\Delta(Z,\alpha_0)[v_1]\Delta(Z,\alpha_0)[v_2]\right]|=o(1)$;

(ii) $\sup_{\alpha\in\mathcal{N}_n,v_1,v_2\in\mathcal{W}_n}|\mu_n\left[\Delta(Z,\alpha)[v_1]\Delta(Z,\alpha)[v_2]\right]|=o_P(1)$;

(iii) $\sup_{\alpha\in\mathcal{N}_n,v\in\mathcal{W}_n}\left|\frac{\partial f(\alpha)}{\partial\alpha}[v]-\frac{\partial f(\alpha_0)}{\partial\alpha}[v]\right|=o(1)$.

Corollary 3.3 *In the sieve ML estimation, (1) if Assumption 3.1 holds, then:* $|\|\widehat{v}_n^*\|_n/\,||v_n^*||_{sd}-1|=o_P(1)$; *(2) if Assumption 3.3 holds, then:* $\left|\|\widetilde{v}_n^*\|_{sd,n}\Big/\,||v_n^*||_{sd}-1\right|=o_P(1)$.

3.1 Numerical equivalence of series variance estimator

For the linear sieve (i.e., series) M estimators $\widehat{\alpha}_n$ of $\alpha_0=(\theta_0,h_0(\cdot))\in\mathcal{A}$, we now establish a numerical equivalence result for two variance estimators of $f(\widehat{\alpha}_n)$ that are derived under different specifications. The first is $||\widehat{v}_n^*||_{sd,n}^2$ used in the asymptotic normality (3.4). This variance estimator is derived from a nonparametric perspective, i.e., the model satisfies the following restriction

$$E\{\Delta(Z,\alpha_0)[v]\}=0 \text{ for any } v\in\mathcal{V} \tag{3.8}$$

where $\alpha_0\in\mathcal{A}$ is an infinite dimensional parameter. Next, we assume that the econometrician specifies the following possibly misspecified model

$$E\{\Delta(Z,\alpha_{0,n^*})[v]\}\stackrel{?}{=}0 \text{ for any } v\in\mathcal{V}_n \text{ and some fixed integer } n^* \tag{3.9}$$

where "$\stackrel{?}{=}$" signifies that the moment conditions in (3.9) are misspecified though the econometrician takes them as if they were correctly specified, $\alpha_{0,n^*}=(\theta_0,h_{0,n^*}(\cdot))$ denotes the projection of α_0 on $\mathcal{V}_{n^*}$ where $h_{0,n^*}(\cdot)=P_{k_{n^*}}(\cdot)'\beta_{0,n^*}$ is a series approximator of $h_0(\cdot)$.

The restriction in (3.9) implies the following set of possibly misspecified moment conditions on the sieve space $\mathcal{A}_n$

$$E\left[g(Z,\alpha_{0,n^*})\right]\equiv E\left(\begin{array}{c}\Delta(Z,\alpha_{0,n^*})[P_{k_{n^*}}(\cdot)]\\ \Delta(Z,\alpha_{0,n^*})[P_{d_\theta}(\cdot)]\end{array}\right)=E\{\Delta(Z,\alpha_{0,n^*})[\overline{P}_{k_{n^*}}(\cdot)]\}\stackrel{?}{=}0, \tag{3.10}$$

where $P_{d_\theta}(\cdot)=(1,...,1)'_{1\times d_\theta}$ and $\overline{P}_{k_{n^*}}(\cdot)'=\left[p_1(\cdot),...,p_{k_{n^*}}(\cdot),P_{d_\theta}(\cdot)'\right]_{1\times(k_{n^*}+d_\theta)}$. Based on the pseudo moment conditions in (3.10), (β_{0,n^*},θ_0) can be estimated using GMM. The standard formula gives the following estimator of the asymptotic variance-covariance matrix of the optimally weighted GMM estimator $\widehat{\alpha}_{n^*,n}\equiv(\widehat{\beta}_{n^*,n},\widehat{\theta}_{n^*,n})$

$$\widehat{V}_{(k_{n^*}+d_\theta)\times(k_{n^*}+d_\theta)}=M_n^{-1}W_nM_n^{-1}, \tag{3.11}$$

where

$$M_n=\frac{1}{n}\sum_{i=1}^n r(Z_i,\widehat{\alpha}_{n^*,n})[\overline{P}_{k_{n^*}}(\cdot),\overline{P}_{k_{n^*}}(\cdot)]$$

and

$$W_n = \frac{1}{n}\sum_{i=1}^{n}\left\{\Delta(Z_i,\widehat{\alpha}_{n^*,n})[\overline{P}_{k_{n^*}}(\cdot)]\Delta(Z_i,\widehat{\alpha}_{n^*,n})[\overline{P}_{k_{n^*}}(\cdot)]'\right\}.$$

Let $\overline{p}_a(\cdot)$ denote the a-th ($a = 1,\ldots,k_{n^*}+d_\theta$) element in $\overline{P}_{k_{n^*}}(\cdot)$. Then M_n and W_n are $(k_{n^*}+d_\theta)\times(k_{n^*}+d_\theta)$ matrices with the a-th elements in the b-th column being $\frac{1}{n}\sum_{i=1}^{n} r(Z_i,\widehat{\alpha}_{n^*,n})[\overline{p}_a(\cdot),\overline{p}_b(\cdot)]$ and $\frac{1}{n}\sum_{i=1}^{n}\{\Delta(Z_i,\widehat{\alpha}_{n^*,n})[\overline{p}_a(\cdot)]\Delta(Z_i,\widehat{\alpha}_{n^*,n})[\overline{p}_b(\cdot)]'\}$ respectively. Using the parametric GMM estimator $\widehat{\beta}_{n^*,n}$ of β_{0,n^*}, we get an estimator $\widehat{h}_{n^*,n}(\cdot) = P_{k_{n^*}}(\cdot)'\widehat{\beta}_{n^*,n}$ for $h_{0,n^*}(\cdot)$. Using the expression in (3.11), we get an estimator for the variance of $\widehat{h}_{n^*,n}(\overline{x})$, where $\overline{x}$ is a point in the interior of the support of X

$$\widehat{Var}[\widehat{h}_{n^*,n}(\overline{x})] = P_{k_{n^*}}(\overline{x})'\widehat{V}_{k_{n^*}\times k_{n^*}}P_{k_{n^*}}(\overline{x}), \tag{3.12}$$

where $\widehat{V}_{k_{n^*}\times k_{n^*}}$ is the leading $k_{n^*}\times k_{n^*}$ submatrix of $\widehat{V}_{(k_{n^*}+d_\theta)\times(k_{n^*}+d_\theta)}$. Denote the last $d_\theta\times d_\theta$ submatrix of $\widehat{V}_{(k_{n^*}+d_\theta)\times(k_{n^*}+d_\theta)}$ as $\widehat{V}_{d_\theta\times d_\theta}$, which gives an estimator of the variance-covariance matrix of $\widehat{\theta}_{n^*,n}$.

Denote $f_h(\cdot):\mathcal{V}\to\mathbb{R}$ as $f_h(\alpha) = h(\overline{x})$ and similarly $f_\theta(\cdot):\mathcal{V}\to\mathbb{R}$ as $f_\theta(\alpha) = \lambda'\theta$ for some $\lambda\in\mathbb{R}^{d_\theta}$, then both functionals are linear in the Hilbert space $\mathcal{V}$. We can compute the empirical Riesz representer $\widehat{v}^*_{h,n}$ for the functional $f_h(\cdot)$ as

$$||\widehat{v}^*_{h,n}||_n^2 = \sup_{v\in\mathcal{V}_n,\ ||v||_n>0}\frac{[f_h(v)]^2}{-E_n\{r(Z_i,\widehat{\alpha}_n)[v,v]\}},$$

where $E_n\{\}$ denotes the expectation with respect to the empirical distribution. To get an explicit form for $\widehat{v}^*_{h,n}$, we write

$$-E_n\{r(Z_i,\widehat{\alpha}_n)[\overline{P}_{k_n}(\cdot),\overline{P}_{k_n}(\cdot)]\} = \begin{pmatrix} I_{11,n} & I_{12,n}\\ I_{21,n} & I_{22,n}\end{pmatrix},$$

where $I_{11,n}$ and $I_{22,n}$ are the leading $k_n\times k_n$ and last $d_\theta\times d_\theta$ submatrices of $-E_n\{r(Z_i,\widehat{\alpha}_n)[\overline{P}_{k_n}(\cdot),\overline{P}_{k_n}(\cdot)]\}$ respectively. Then, it is easy to see that

$$\widehat{v}^*_{h,n} = \left[-I_{22,n}^{-1}I_{21,n}\beta^*_{h,n}, P_{k_n}(\cdot)'\beta^*_{h,n}\right]$$

where $\beta^*_{h,n} = I_n^{-1}P_{k_n}(\overline{x})$ and $I_n = I_{11,n} - I_{12,n}I_{22,n}^{-1}I_{21,n}$. Similarly, we get

$$\widehat{v}^*_{\theta,n} = \left[\theta^*_{\lambda,n}, -P_{k_n}(\cdot)'I_{n,11}^{-1}I_{n,12}\theta^*_{\lambda,n}\right]$$

where $\theta^*_{\lambda,n} = J_n^{-1}\lambda$ and $J_n = I_{22,n} - I_{21,n}I_{22,n}^{-1}I_{21,n}$.

Lemma 3.2 *If $n^* = n$, then: (1) $||\widehat{v}^*_{h,n}||^2_{sd,n} = \widehat{Var}[\widehat{h}_{n^*,n}(\overline{x})]$, and (2) $||\widehat{v}^*_{\theta,n}||^2_{sd,n} = \lambda'\widehat{V}_{d_\theta\times d_\theta}\lambda$, where $\lambda\neq 0$ is any vector in $\mathbb{R}^{d_\theta}$.*

The above lemma simplifies the computation of the sieve variance estimator $||\widehat{v}_n^*||_{sd,n}^2$. Indeed, it implies that to compute $||\widehat{v}_n^*||_{sd,n}^2$, we can view the infinite dimensional parameter α_0 to be parametrically specified as $\alpha_{0,n}$ in finite samples. Then we can use the standard formula in the parametric estimation to get $||\widehat{v}_n^*||_{sd,n}^2$.

Notice that the above numerical equivalence results are established when the same finite dimensional linear sieves (or series) are used to approximate the unknown function $h_0()$ and the sieve Riesz representer $v_n^*()$. The numerical equivalence results may no longer hold when the unknown function $h_0()$ is approximated by nonlinear sieves such as neural networks (e.g., White and Wooldridge (1991), Gallant and White (1992), Chen and White (1999)) and radial basis networks (e.g., Chen, Racine and Swanson, 2001).

4 Sieve Likelihood Ratio Inference

Previously, Shen and Shi (2005) establishes that the sieve LR test of a regular functional is asymptotically chi-square distributed. In this section, we extend their result to the sieve LR test of an irregular functional. This result allows us to construct confidence sets for a general functional $f(\alpha_0)$ without the need to estimate the sieve variance $||v_n^*||_{sd}^2$ in Lemma 2.1.

Let $\mathbf{f}(\alpha) = [f_1(\alpha), ..., f_m(\alpha)]$ be a m-dimensional vector-valued functionals on the parameter space $\mathcal{A}$. We are interested in testing the null hypothesis of $\mathbf{f}(\alpha_0) = 0$ based on a LR statistic. Let $L_n(\alpha) \equiv n^{-1} \sum_{i=1}^n \ell(Z_i, \alpha)$ be the sample log-likelihood function. Let $\widehat{\alpha}_n$ be the unconstrained sieve ML estimate: $\widehat{\alpha}_n = \arg\max_{\alpha \in \mathcal{A}_n} L_n(\alpha)$. We define the constrained sieve ML estimate $\widetilde{\alpha}_n$ as

$$\widetilde{\alpha}_n = \arg \max_{\{\alpha \in \mathcal{A}_n:\ \mathbf{f}(\alpha)=0\}} L_n(\alpha). \tag{4.1}$$

For $j = 1, ..., m$, we assume that the pathwise derivative $\frac{\partial f_j(\alpha_0)}{\partial \alpha}[\cdot]$ is a linear functional on $\mathcal{V}$. By the RRT, there exists a sieve Riesz representer $v_{j,n}^* \in \mathcal{V}_n$ for $\frac{\partial f_j(\alpha_0)}{\partial \alpha}[\cdot]$. Denote $u_{j,n}^* \equiv v_{j,n}^*/||v_{j,n}^*||_{sd}$.

Assumption 4.1 *(i) For $j = 1, ..., m$, $f_j(\cdot)$ and its sieve Riesz representer $v_{j,n}^*$ satisfy Assumption 2.1 with $||v_{j,n}^*|| = ||v_{j,n}^*||_{sd}$; and $\left(\frac{\partial f_1(\alpha_0)}{\partial \alpha}[\cdot], ..., \frac{\partial f_m(\alpha_0)}{\partial \alpha}[\cdot]\right)$ is linearly independent; (ii) Denote $\alpha^*(\alpha) \equiv \alpha \pm \sum_{j=1}^m \langle u_{j,n}^*, \widehat{\alpha}_n - \alpha_0 \rangle u_{j,n}^*$.*

$$(a) \quad \sup_{\alpha \in \mathcal{N}_n} \mu_n \left\{ \ell(Z, \alpha^*(\alpha)) - \ell(Z, \alpha) - \Delta(Z, \alpha_0)[\alpha^*(\alpha) - \alpha] \right\} = o_P(n^{-1});$$

$$(b) \quad \sup_{\alpha \in \mathcal{N}_n} \left\{ E[\ell(Z, \alpha) - \ell(Z, \alpha^*(\alpha))] - \frac{||\alpha^*(\alpha) - \alpha_0||^2 - ||\alpha - \alpha_0||^2}{2} \right\} = o_P(n^{-1});$$

(iii) for $j = 1, ..., m$, $\sqrt{n}\mu_n \left\{ \Delta(Z, \alpha_0)[u_{j,n}^*] \right\} \to_d N(0, 1)$.

Assumption 4.1(ii) is similar to Assumption 2.2, except that the local perturbation $\alpha^*(\alpha)$ is defined slightly different here.

Theorem 4.1 *Let Assumptions 2.2 and 4.1 hold. Suppose that* $||\widetilde{\alpha}_n - \alpha_0||_s = O_P(\delta^*_{s,n})$ *under the null of* $\mathbf{f}(\alpha_0) = 0$*. Then under the null hypothesis of* $\mathbf{f}(\alpha_0) = 0$*, we have*

$$2n\left[L_n(\widehat{\alpha}_n) - L_n(\widetilde{\alpha}_n)\right] \to_d \mathcal{X}^2(m) \tag{4.2}$$

where $\mathcal{X}^2(m)$ *is a Chi-square random variable with degree of freedom* m*.*

Theorem 4.1 is valid regardless if the functional $\mathbf{f}(\cdot) : \mathcal{A} \to \mathbb{R}^m$ is regular or not. Previously Fan, Zhang and Zhang (2001) established a result on generalize LR statistics with slowly growing dimension m for a consistent specification test based on local linear regression of a nonparametric regression model. It would be interesting to see if Theorem 4.1 remains valid when m grows slowly with the sample size.

Let $\alpha_0 = (\theta_0, h_0)$ with $\theta_0 \in int(\Theta)$. We next illustrate how to construct confidence intervals for θ_0. Let $L^*_n(\theta) \equiv \sup_{h \in \mathcal{H}_n} L_n(\theta, h)$. Theorem 4.1 implies that

$$2n\left[L_n(\widehat{\alpha}_n) - L^*_n(\theta_0)\right] \to_d \mathcal{X}^2(d_\theta). \tag{4.3}$$

Let $\mathcal{X}^2_\tau(d_\theta)$ denote the critical value of the Chi-square random variable with degree of freedom d_θ and confidence level τ, then using (4.3), we can see that the $1-\tau$ confidence region of θ_0 can be constructed as

$$\left\{\theta \in \Theta : L_n(\widehat{\alpha}_n) - \frac{\mathcal{X}^2_\tau(d_\theta)}{2n} \leq L^*_n(\theta) \leq L_n(\widehat{\alpha}_n)\right\}. \tag{4.4}$$

5 Application to Semiparametric Duration Models with Unobserved Heterogeneity

Duration Model with Unobserved Heterogeneity. Let $g(T|U, X, \theta_0)$ be the conditional density function of an observed duration T given a scalar unobserved heterogeneity U and a vector of observed control variables X. For example, $g(T|U, X, \theta_0)$ can be the Weibull density assumed in Heckman and Singer (1984), i.e.

$$g(T|U, X, \theta_0) = \theta_{1,0} T^{\theta_{1,0}-1} \exp\left[\theta'_{2,0} X + U - T^{\theta_{1,0}} \exp\left(\theta'_{2,0} X + U\right)\right]. \tag{5.1}$$

Let $Z = (T, X')'$ with support $\mathcal{Z}$. Let $h_0(\cdot) \in \mathcal{H}$ denote the square-root density function of U with support $\mathbb{R}$. Let $\alpha_0 = (\theta_0, h_0)$. We can write the conditional density of the observed duration T given X as

$$p(Z, \alpha_0) = p(T|X, \theta_0, h_0) = \int_{\mathbb{R}} g(T|u, X, \theta_0) h_0^2(u) du. \tag{5.2}$$

A functional of interest could be: $f(\alpha_0) = \lambda' \theta_0$ for any $\lambda \in \mathbb{R}^{d_\theta}$ with $\lambda \neq 0$, or $f(\alpha_0) = h_0^2(\overline{u})$ for some point $\overline{u}$.

With a random sample of $\{Z_i\}_{i=1}^n = \{(T_i, X_i')'\}_{i=1}^n$, we can estimate the unknown parameters $\alpha_0 = (\theta_0, h_0)$ by sieve ML estimator $\widehat{\alpha}_n = \left(\widehat{\theta}_n, \widehat{h}_n\right)$ that solves

$$\max_{\theta\in\Theta,\, h\in\mathcal{H}_n} L_n(\alpha), \quad \text{with } L_n(\alpha) = \frac{1}{n}\sum_{i=1}^n \log p(Z_i, \alpha), \tag{5.3}$$

where $\mathcal{H}_n$ denotes the sieve space that becomes dense in $\mathcal{H}$ as $n \to \infty$.

Let $\mathcal{H} = \Lambda_c^\gamma(\mathbb{R})$. Then $\mathcal{H}_n$ could be a Hermite polynomial sieve

$$\left\{h(u) = (2\pi)^{-1/2}\sum\nolimits_{j=1}^{k_n} \beta_j u^j \exp(-u^2/2),\ \int h^2(u)du = 1\right\}, \tag{5.4}$$

or a B-spline wavelet sieve

$$\left\{h(u) = \sum_{k=0}^{k_n}\sum_{l\in\mathcal{K}_n} \beta_{kl} 2^{k/2} B_s(2^k x - l),\ \int h^2(u)du = 1\right\} \tag{5.5}$$

where $B_s(\cdot)$ denotes the cardinal B-spline of order s (with $s > [\gamma]$):

$$B_s(y) = \frac{1}{(s-1)!}\sum_{i=0}^{s}(-1)^i \begin{pmatrix} s \\ i \end{pmatrix} [\max(0, y-i)]^{s-1}. \tag{5.6}$$

Define a norm on $\mathcal{A} = \Theta\times\Lambda_c^\gamma(\mathbb{R})$ as: $\|\alpha\|_A = \|\theta\|_E + \|h\|_{\infty,\omega}$ in which $\|h\|_{\infty,\omega} \equiv \sup_u |h(u)\omega(u)|$ with $\omega(u) = \left(1+u^2\right)^{-\varsigma/2}$, $\varsigma > 0$. Let $\pi_n\alpha_0 = (\theta_0, \pi_n h_0)$, where $\pi_n h_0 \in \mathcal{H}_n$.

Condition 5.1 *(i) $\alpha_0 = (\theta_0, h_0) \in \mathcal{A} = \Theta\times\Lambda_c^\gamma(\mathbb{R})$ with $\gamma > 1/2$ and there is $p(z,\alpha_0) \neq p(z,\alpha)$ a.e. for any $\alpha \neq \alpha_0$, where Θ is a compact set in $\mathbb{R}^{d_\theta}$ and $\int h_0^2(u)du = 1$; (ii) $E[\log p(Z,\alpha)]$ is continuous in α_0 under the norm $\|\cdot\|_A$; (iii) there is a $\pi_n h_0 \in \mathcal{H}_n$ such that $\|h_0 - \pi_n h_0\|_{\infty,\omega} = O(k_n^{-\gamma})$; (iv) $E[|\log p(Z, \pi_n\alpha_0)|] \leq const. < \infty$; there are a finite $\kappa_1 > 0$ and a measurable function $D_{1,n}(Z)$ with $E[D_{1,n}(Z)] \leq const. < \infty$ such that*

$$\sup_{\alpha',\alpha\in\mathcal{A}_n:\|\alpha'-\alpha\|_A\leq\delta} \left|\log p(Z,\alpha) - \log p(Z,\alpha')\right| \leq D_{1,n}(Z)\times\delta^{\kappa_1} \text{ for all } \delta > 0.$$

Under Condition 5.1 and $\frac{k_n}{n} = o(1)$, Remark 3.3 in Chen (2007) immediately implies the consistency of the sieve MLE $\widehat{\alpha}_n$ under the $\|\cdot\|_A$−metric: $\|\widehat{\alpha}_n - \alpha_0\|_A = o_P(1)$. Let $\mathcal{A}(\epsilon)$ and $\mathcal{A}_n(\epsilon)$ be given in (2.2) for some $\epsilon > 0$.

By the information identity, we have $\|v\| = \|v\|_{sd}$, where

$$\|v\|_{sd}^2 = E\left[\left(\frac{\partial \log p(Z,\alpha_0)}{\partial\alpha}[v]\right)^2\right] = E\left[\frac{\left|\left(\int \frac{\partial g(T|u,X,\theta_0)h_0^2(u)}{\partial\theta'}du\right)v_\theta + 2\int g(T|u,X,\theta_0)h_0(u)v_h(u)du\right|^2}{p^2(Z,\alpha_0)}\right]$$

for any $v = (v_\theta, v_h) \in \mathcal{V}$. For ease of notation, in this example, we use $\|v\|_{sd}$ and $\|v\|_{sd,n}$ to define the theoretical and empirical Riesz representers of related functionals respectively.

Condition 5.2 *(i)* $(\theta_0, h_0) \in int(\Theta) \times \Lambda^{\gamma}_{c_0}(\mathbb{R})$ *for* $c_0 \in (0, c)$*;*

$$(ii) \quad E\left[\sup_{\widetilde{\alpha}\in\mathcal{A}(\epsilon),\alpha\in\mathcal{A}_n(\epsilon)}\left(\frac{\partial \log p(Z,\widetilde{\alpha})}{\partial\alpha}\left[\frac{\alpha-\alpha_0}{\|\alpha-\alpha_0\|_{sd}}\right]\right)^2\right] \leq const. < \infty;$$

$$(iii) \quad \sup_{\widetilde{\alpha}\in\mathcal{A}(\epsilon),v\in\mathcal{A}(\epsilon):\|v\|_{sd}=1} E\left[\left|\frac{\partial^2 \log p(Z,\widetilde{\alpha})}{\partial\alpha\partial\alpha}[v,v] - \frac{\partial^2 \log p(Z,\alpha_0)}{\partial\alpha\partial\alpha}[v,v]\right|\right] = o(1).$$

Let $\xi_n = \sup_{\alpha\in\{\mathcal{A}_n:\|\alpha-\pi_n(\alpha_0)\|_{sd}\neq 0\}} \{d_A(\alpha, \pi_n(\alpha_0))/\|\alpha-\pi_n(\alpha_0)\|_{sd}\}$. The following remark on the convergence rates are direct applications of Theorem 3.2 in Chen (2007).

Remark 5.3 *Let* $\frac{k_n}{n} = o(1)$*. Under Conditions 5.1 and 5.2,* $\|\widehat{\alpha}_n - \alpha_0\|_{sd} = O_P\left(\max\left\{\sqrt{\frac{k_n}{n}}, k_n^{-\gamma}\right\}\right)$*, and* $\|\widehat{\alpha}_n - \alpha_0\|_A = O_P\left(\max\left\{\xi_n\sqrt{\frac{k_n}{n}}, k_n^{-\gamma}\right\}\right) = o_P(1)$.

Let v_n^* be the sieve Riesz representer of the functional $\frac{\partial f(\alpha_0)}{\partial\alpha}[\cdot]$ on $\mathcal{V}_n$ and $u_n^* = v_n^*/\|v_n^*\|_{sd}$. The following condition is for the limiting distribution of the plug-in sieve estimator $f(\widehat{\alpha}_n)$.

Condition 5.4 *(i) For all small positive value* $\delta = o(1)$*,*

$$E\left[\sup_{\alpha',\alpha\in\mathcal{N}_n:||\alpha'-\alpha||_A<\delta}\left|\frac{\partial \log p(Z,\alpha')}{\partial\alpha}[u_n^*] - \frac{\partial \log p(Z,\alpha)}{\partial\alpha}[u_n^*]\right|^2\right] \lesssim \delta^2;$$

(ii) uniformly over $\alpha \in \mathcal{N}_n, \widetilde{\alpha} \in \mathcal{N}_0$*,*

$$\|\alpha-\alpha_0\|_{sd}^2 \sup_{v\in\mathcal{N}_0:\|v\|_{sd}=1} E\left[\left|\frac{\partial^2 \log[p(Z,\widetilde{\alpha})]}{\partial\alpha\partial\alpha}[v,v] - \frac{\partial^2 \log[p(Z,\alpha_0)]}{\partial\alpha\partial\alpha}[v,v]\right|\right] = o(n^{-1});$$

(iii) there is $\kappa_2 > 0$ *such that* $\lim_{n\to\infty} n^{-\frac{\kappa_2}{2}} E\left[\left|\frac{\partial \log p(Z,\alpha_0)}{\partial\alpha}[u_n^*]\right|^{2+\kappa_2}\right] = 0.$

Let $\mathcal{W}_n = \{v \in \mathcal{V}_n : \|v\|_{sd} = 1\}$. The following condition is for the consistent sieve variance estimation:

Condition 5.5 *(i)* $\sup_{\alpha\in\mathcal{N}_n,v\in\mathcal{W}_n} E\left[\left(\frac{\partial \log[p(Z,\alpha)]}{\partial\alpha}[v] - \frac{\partial \log[p(Z,\alpha_0)]}{\partial\alpha}[v]\right)^2\right] = o(1)$*; (ii) there is a measurable function* $D_{2,n}(Z)$ *such that for all small positive value* $\delta = o(1)$*,*

$$\sup_{\alpha',\alpha\in\mathcal{N}_n:||\alpha'-\alpha||_A<\delta,v\in\mathcal{W}_n}\left|\frac{\partial \log[p(Z,\alpha')]}{\partial\alpha}[v] - \frac{\partial \log[p(Z,\alpha)]}{\partial\alpha}[v]\right| \leq D_{2,n}(Z)\times\delta;$$

(iii) there is a measurable function $D_{3,n}(Z)$ *such that* $\sup_{\alpha\in\mathcal{N}_n,v\in\mathcal{W}_n}\left|\frac{\partial \log p(Z,\alpha)}{\partial\alpha}[v]\right| \leq D_{3,n}(Z)$ *and*

$$E\left[D_{3,n}(Z)\times(D_{2,n}(Z)+D_{3,n}(Z))\right] \leq const. < \infty.$$

Recall that $\widetilde{v}_n^*$ is the empirical Riesz representer of the functional $\frac{\partial f(\widehat{\alpha}_n)}{\partial\alpha}[\cdot]$ w.r.t. the empirical norm $\|\cdot\|_{sd,n}$.

Proposition 5.6 *Let Assumptions 2.1 and 3.3(iii), and Conditions 5.1, 5.4 and 5.5 hold. Then:*

$$\sqrt{n}\frac{f(\widehat{\alpha}_n)-f(\alpha_0)}{\|\widetilde{v}_n^*\|_{sd,n}} \to_d N(0,1)$$

where $\|\widetilde{v}_n^*\|_{sd,n}^2 = n^{-1}\sum_{i=1}^n \left(\frac{\partial \log[p(Z_i,\widehat{\alpha}_n)]}{\partial \alpha}[\widetilde{v}_n^*]\right)^2$.

For the functional $f(\alpha)=\lambda'\theta$ for any $\lambda \in \mathbb{R}^{d_\theta}$ with $\lambda \neq 0$ we have $||v_n^*||^2 = \lambda'\left(E[\mathcal{S}_{k_n}\mathcal{S}_{k_n}']\right)^{-1}\lambda$ as computed in (2.10). Previously Hahn (1994) and Ishwaran (1996, 1999) have shown that if $g(T|u,X,\theta)$ is in the exponential family then θ_0 could have zero information bound (i.e., $f(\alpha_0)=\lambda'\theta_0$ has $\lim_{k_n\to\infty}||v_n^*||^2 = \lim_{k_n\to\infty}\lambda'\left(E[\mathcal{S}_{k_n}\mathcal{S}_{k_n}']\right)^{-1}\lambda = \infty$, where $\mathcal{S}_{k_n}$ is defined in (2.10)). Let $||\widetilde{v}_n^*||_{sd,n}^2 = \lambda'(E_n[\widetilde{\mathcal{S}}_{k_n}\widetilde{\mathcal{S}}_{k_n}'])^{-1}\lambda$ denote the empirical counterpart of $||v_n^*||^2$. As $f(\alpha)=\lambda'\theta$ satisfies Assumptions 2.1 and 3.3(iii) trivially, Proposition 5.6 immediately leads to

$$\sqrt{n}\frac{\lambda'(\widehat{\theta}_n-\theta_0)}{||\widetilde{v}_n^*||_{sd,n}} \to N(0,1). \tag{5.7}$$

The alternative way of constructing the confidence bands is to use the Chi-square approximation established in Section 4. Let $\theta_{S,0}$ (with dimension d_{θ_S}) be a subset of θ_0 and $\theta_{S^c,0}$ be its complement, i.e., $\theta_0=(\theta_{S,0},\theta_{S^c,0})$. Then under Conditions 5.1 and 5.4 we can invoke Theorem 4.1 to deduce that

$$2\left[\sum_{i=1}^n \log p(Z_i,\widehat{\alpha}_n) - \max_{\{\alpha\in\mathcal{A}_n:\ \theta_S=\theta_{S,0}\}}\sum_{i=1}^n \log p(Z_i,\alpha)\right] \to_d \mathcal{X}^2(d_{\theta_S}).$$

6 Simulation Studies

We assess the finite sample performance of the sieve M estimate, the inferences using the Gaussian approximation and the Chi-square approximation in a simulation study. We use the semiparametric mixture model (5.1)-(5.2) studied in the previous section to generate the data, where the control variable X is a standard normal random variable and the unobserved heterogeneity U has the following log-gamma density

$$h_0^2(u) = 2\pi^{-\frac{1}{2}}\exp(3u/2)\exp\left[-\exp(u)\right],\ u\in(-\infty,\infty).$$

The unknown structural parameter θ_0 is specified as $\theta_0'=(\theta_{1,0},\theta_{2,0})=(1,1)$.

The unknown parameter $\alpha_0=(\theta_0,h_0)$ is estimated by the sieve ML estimation (5.3), in which we use the Hermite polynomial sieve

$$\left\{h_{k_n}^2(u) = (2\pi)^{-1}\left[\sum\nolimits_{j=1}^{k_n}\beta_j u^j\exp(-u^2/2)\right]^2 \text{ with } \int h_{k_n}^2(u)du = 1\right\} \tag{6.1}$$

to approximate the unknown density function $h^2(u)$. In this simulation study, we are also interested in the finite sample properties of the sieve ML estimate with basis function selected by some

information criteria (IC), i.e. $k_n = \widehat{k}_n$ is selected to minimizing

$$\min_k \left[-2nL_n(\widehat{\theta}_{n,k}, \widehat{h}_{n,k}) + C_n(d_\theta + k) \right], \tag{6.2}$$

where $\widehat{\alpha}_{n,k} = (\widehat{\theta}_{n,k}, \widehat{h}_{n,k})$ denotes the sieve ML estimate given k basis functions, C_n is a sequence of positive non-decreasing values, $L_n(\cdot)$ is the log-likelihood function defined in (5.3) (also see (7.2)). It is clear that when $C_n = 2$ or $\log(n)$, the IC defined in (6.2) becomes Akaike information criterion (AIC) or Bayesian information criterion (BIC) respectively. See Konishi and Kitagawa (2008) for other choices of C_n (and ICs).

The IC is a computationally simple and useful way to select the number of basis functions. When the unknown function h_0 can be represented by finite many basis functions, it is well know that BIC can consistently select these basis functions though AIC will pick up a larger model. On the other hand, when infinite many basis functions are needed for representing the unknown true function h_0, the IC is useful to reduce the dimension of the model and improve the variance of the sieve estimator in finite samples. However, in this scenario, one at least has to show that the sieve ML estimator based on the selected basis functions is consistent. This result is proved under the following condition.

Assumption 6.1 *(i) For any finite fixed k, there is $\alpha_{p,k} = (\theta_{p,k}, h_{p,k}) \in \Theta \times \mathcal{H}_k$ such that $\alpha_{p,k} \neq \alpha_0$ and $\widehat{\alpha}_{p,k} - \alpha_{p,k} = o_P(1)$; (ii) there is a sequence $c(k)$ such that $\inf_{\alpha \in \Theta \times \mathcal{H}_k \setminus \{\alpha_0\}} E\left[\ell(Z, \alpha_0) - \ell(Z, \alpha)\right] \geq c(k) > 0$ for all k; (iii) $\sup_{\alpha \in \Theta \times \mathcal{H}_k} |\mu_n\left[\ell(Z, \alpha)\right]| = o_P(1)$; (iv) there is a sequence $k_n^* \to \infty$ as $n \to \infty$ such that $\left\|\widehat{\alpha}_{k_n^*} - \alpha_0\right\| = o_P(1)$; (v) $|\mu_n\left[\ell(Z, \alpha_0)\right]| = o_P(1)$, and for any sequence $\eta_n = o(1)$, there is*

$$\sup_{\{\alpha \in \mathcal{A}_{k_n} : \|\alpha - \alpha_0\| \leq \eta_n\}} \left| \frac{1}{n} \sum_{i=1}^n \left[\ell(Z_i, \alpha) - \ell(Z_i, \alpha_0)\right] \right| = o_P(1).$$

Assumption 6.1(i) defines the pseudo true parameter $\alpha_{p,k} = (\theta_{p,k}, h_{p,k})$ when the infinite dimensional parameter h_0 is parametrically specified/misspecified. Alternatively, the pseudo true parameter $\alpha_{p,k}$ can be defined as a solution to $\max_{\Theta \times \mathcal{H}_k} E\left[\ell(Z, \theta, h)\right]$. Assumption 6.1(ii) is the identification condition of the true parameter $\alpha_0 = (\theta_0, h_0)$. The uniform law of large numbers assumed in Assumption 6.1(iii) and the stochastic equicontinuity condition assumed in Assumption 6.1(v) are easy to verify. Assumption 6.1(iv) assumes the existence of a consistent estimator $\widehat{\alpha}_{k_n^*}$ of α_0.

Lemma 6.1 *Suppose that Assumption 6.1 hold. If $k_n^* C_n / n = o(1)$, then $\widehat{k}_n \to_P \infty$ as $n \to \infty$.*

According to Lemma 6.1, to ensure that asymptotically all basis functions are selected C_n could not diverge to infinity too fast. It is clear that both AIC and BIC satisfy the requirement that $k_n^* C_n / n = o(1)$. It will be interesting to investigate the specific order that $\widehat{k}_n$ diverges to infinity

and to study the properties of the sieve ML estimators based on the selected basis functions. Due to the length of this paper, we leave these for future research.

In this simulation study, we first investigate the finite sample performance of the sieve ML estimators. Different sample sizes with different numbers of basis functions are considered. 5000 simulated samples are generated for each combination of the sample size and number of basis functions. The performance of the sieve ML estimator $\widehat{\theta}_n$ is evaluated by finite sample bias, standard deviation and root of mean square error. Second, we investigate the finite sample performance of the sieve ML estimates when the number of basis functions is selected by AIC or BIC. Third, the finite sample performance of the inference procedures based on the Gaussian approximation and Chi-square approximation are evaluated by their respective coverage probabilities and lengths of confidence intervals at 95% confidence level. As the inference based on the Chi-square approximation is computationally intensive, we only consider 1000 simulated samples in this study. The simulation results are summarized in Tables B.1, B.2 and B.3.

Table B.1 indicates that the structural coefficients are well estimated by the sieve ML method. The finite sample bias of the sieve ML estimators $\widehat{\theta}_{1,n}$ and $\widehat{\theta}_{2,n}$ are small, even when the sample size is only 100. Given the sample size, the finite sample bias decreases as the number of the basis functions increases, though the standard error increases. When we increase the sample size n to 500, the standard error of $\widehat{\theta}_n$ decreases. The simulation results in Table B.1 make it clear that the finite sample properties of the sieve ML estimators are sensitive to the choice of the number of basis functions. From the perspective of minimizing the mean square error, the best choice of k_n would be 3 when the sample size $n = 100$ and the best choice of k_n would be 5 when $n = 500$. Table B.2 presents the results of the sieve ML estimators based on the number of basis functions selected by AIC and BIC. It is clear that AIC tends to select more basis functions. As a result, its estimator has smaller bias but larger standard error in finite samples. When the sample size is small, the performances of AIC and BIC (and their related sieve ML estimators) are similar, but when the sample size is large, the performance of AIC is much better. Table B.3 presents the finite sample performance of the confidence intervals constructed using the Gaussian approximation and the Chi-square approximation. It is clear that the confidence intervals based on the Gaussian approximation have coverage probability far below the nominal size 95%, particularly when the sample size and the number of the basis functions are small. On the other hand, the Chi-square confidence intervals have more accurate coverage probability, though they are slightly longer than the Gaussian confidence intervals. When the sample size is increased, both confidence intervals become narrow and their coverage probabilities are improved.

7 An Empirical Application

In this section, we apply the results of Sections 5 and 6 to a semiparametric duration analysis of the second birth in China. The data is from the China Health and Nutrition Survey (CHNS), which is an on-going survey conducted by researchers from the University of North Carolina, the National Institute of Nutrition and Food Safety, and the Chinese Center for Disease Control and Prevention. The CHNS is a panel data set and we use its 2006 survey in this empirical example. The spell T is computed as the number of years between the first and the second birth. The twins are treated as one birth. The control variables X include the gender of the first child, the years of schooling, a dummy variable indicating whether there was bonus to accepting the one child policy, and a dummy variable indicating the location of the household (rural or city). Let S_i be a binary random variable and $S_i = 1$ denote the event that woman i will eventually deliver a second birth. We use the Weibull density function to specify the conditional density of the spell T given the unobserved heterogeneity U, a set of control variables X and the event $S = 1$, i.e.

$$g(T|U, X, \theta_0, S = 1) = \theta_{1,0} T^{\theta_{1,0}-1} \exp\left[X'\theta_{2,0} + U - T^{\theta_{1,0}} \exp\left(X'\theta_{2,0} + U\right)\right]. \tag{7.1}$$

The variable S is not observed for individuals who will potentially (but currently do not) have a second child. To avoid this complication and make this example simple, we restrict our attention to the sample of women whose fertility history is established (i.e. we focus on Chinese women whose ages are larger than 40, and assume that women older than 40 have negligible probability of having more kids).

We use the Hermite polynomials $h_n^2(u)$ specified in (6.1) to approximate the unknown density $h^2(u)$ of the unobservable heterogeneity U. The structural coefficient θ_0 and the density function $h_0^2(u)$ are estimated by the sieve ML estimation (5.3), which we rewrite as (7.2):

$$\max_{\theta\in\Theta,\ h_n\in\mathcal{H}_n} n^{-1} \sum_{i=1}^{n} \log\left[A_1(\theta, Z_i) \int \exp\left[u - A_2(\theta, Z_i)\exp(u)\right] h_n^2(u) du\right], \tag{7.2}$$

where $A_1(\theta, Z) \equiv \theta_1 T^{\theta_1 - 1} \exp(X'\theta_2)$ and $A_2(\theta, Z) \equiv T^{\theta_1} \exp(X'\theta_2)$.

Tables B.4 contains the sieve ML estimates of the structural coefficients. AIC and BIC select two different sub-models in this example. The number of the basis functions selected by AIC is 4 while the selection of BIC is 3. The estimators of the structural parameters based on AIC are presented in the upper panel of Table B.4 and the estimators based on BIC are presented in the lower panel. The empirical results from sieve ML estimation indicate that the hazard rates of the women whose first kids are girls and who live in the rural areas are higher than those who live in city with their first kids being boys. The results confirm the boy preference culture in China, particularly in its rural area. The bonus to the households which accept the one child policy, on the other hand, decrease the hazard rate of having a second child. The estimate of the coefficient of

bonus, in some sense, measures the effect of the one child policy on the hazard rate. Our empirical results indicate that education has positive effect on the hazard rate. One possible explanation is that the education is a proxy variable for the household income, which increases the demand of kids.

8 Conclusion

In this paper we present sieve inference procedures for semi-nonparametric models with i.i.d. data. We first provide simple consistent variance estimators of the plug-in sieve M estimators of possibly irregular functionals. Our numerical equivalence result shows that one can use the standard errors of misspecified parametric M estimators as valid estimators of the semiparametric asymptotic standard errors of the sieve M estimators. We further establish that the sieve LR statistic is asymptotically chi-square distributed regardless of whether a functional is regular or irregular. These results are easily applicable to construct confidence sets for possibly irregular parameters in many semi-nonparametric models, such as the models studied in Elbers and Ridder (1982), Gallant and Nychika (1987), Chen, Heckman and Vytlacil (1998), Ridder and Woutersen (2003), Chamberlain (1986, 2010), Bierens (2013) and others. We apply our general theory to sieve ML estimation of and inference on a semiparametric duration model with unobserved nonparametric heterogeneity of Heckman and Singer (1984). In a contemporaneous work, Chen and Pouzo (2012) consider sieve minimum distance inference on irregular parameters of conditional moment restrictions containing unknown functions of endogenous regressors.

References

[1] Andrews, D., 1991. Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models. *Econometrica,* 59, 307-345.

[2] Andrews, D., M. Schafgans, 1998. Semiparametric Estimation of the Intercept of a Sample Selection Model. *Review of Economic Studies*, 65, 497-517.

[3] Bickel, P., C. Klaassen, Y. Ritov, J. Wellner, 1993. *Efficient and Adaptive Estimation for Semiparametric Models*. New York: Springer.

[4] Bierens, H., 2013. Consistency and Asymptotic Normality of Sieve ML Estimators Under Low-Level Conditions. *Econometric Theory,* forthcoming.

[5] Birge, L., P. Massart 1998. Minimum Contrast Estimators on Sieves: Exponential Bounds and Rates of Convergence. *Bernoulli*, 4, 329-375.

[6] Chamberlain, G., 1986. Asymptotic Efficiency in Semiparametric Models with Censoring. *Journal of Econometrics*, 32, 189-218.

[7] Chamberlain, G., 2010. Binary Response Models for Panel Data: Identification and Information. *Econometrica*, 78, 159-168.

[8] Chen, X., 2007. Large Sample Sieve Estimation of Semi-Nonparametric Models. In: Heckman, J.J., Leamer, E. (Eds.), *Handbook of Econometrics*, Vol. 5. North-Holland, Amsterdam.

[9] Chen, X., J. Heckman, E. Vytlacil, 1998. Non/Semiparametric Identification and Estimation of a Dynamic Discrete-Time Discrete-Choice Models with Unobserved Heterogeneity. University of Chicago, *unpublished manuscript*.

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

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

[12] Chen, X., D. Pouzo, 2012. Sieve Quasi Likelihood Ratio Inference on Semi/nonparametric Conditional Moment Models. *unpublished manuscript*.

[13] Chen, X., J. Racine, N. Swanson, 2001. Semiparametric ARX Neural Network Models with an Application to Forecasting Inflation. *IEEE Tran. Neural Networks,* 12, 674-683.

[14] Chen, X., H. White, 1999. Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators. *IEEE Tran. Information Theory,* 45, 682-691.

[15] Eastwood, B., A. Gallant, 1991. Adaptive Rules for Seminonparametric Estimators that Achieve Asymptotic Normality. *Econometric Theory*, 7, 307-340.

[16] Elbers, C., G. Ridder, 1982. True and Spurious Duration Dependence: the Identifiability of the Proportional Hazard Model. *Review of Economic Studies,* 49, 403-409.

[17] Fan, J., C. Zhang, J. Zhang, 2001. Generalized Likelihood Ratio Statistics and Wilks Phenomenon. *Annals of Statistics* 29, 640–652.

[18] Gallant, A.R., D. Nychka, 1987. Semi-non-parametric Maximum Likelihood Estimation. *Econometrica*, 55, 363-390.

[19] Gallant, A.R., G. Souza, 1991. On the Asymptotic Normality of Fourier Flexible Form Estimates. *Journal of Econometrics*, 50, 329-353.

[20] Gallant, A.R., H. White, 1992. On Learning the Derivatives of an Unknown Mapping with Multilayer Feedforward Networks. *Neural Networks* 5, 129-138.

[21] Graham, B., J. Powell, 2012. Identification and Estimation of Average Partial Effects in "Irregular" Correlated Random Coefficient Panel Data Models. *Econometrica*, 80, 2105-2152.

[22] Grenander, U., 1981. *Abstract Inference*. New York: Wiley Series.

[23] Hahn, J., 1994. The Efficiency Bound of the Mixed Proportional Hazard Model. *Review of Economic Studies*, 61, 607-629.

[24] Heckman, J., B. Singer, 1984. A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. *Econometrica*, 52, 271-320.

[25] Honoré, B., 1990. Simple Estimation of a Duration Model with Unobserved Heterogeneity. *Econometrica*, 58, 453–473.

[26] Honoré, B., 1994. A Note on the Rate of Convergence of Estimators of Mixtures of Weibulls. *Manuscript*, Northwestern University.

[27] Horowitz, J., 1992. A Smoothed Maximum Score Estimator for the Binary Response Model. *Econometrica*, 60, 505–531.

[28] Huang, J., 1998. Projection Estimation in Multiple Regression with Application to Functional ANOVA Models. *The Annals of Statistics*, 26, 242-272.

[29] Huang, J., 2003. Local Asymptotics for Polynomial Spline Regression. *Annals of Statistics*, 31(5), 1600-1635.

[30] Ishwaran, H., 1996. Identifiability and Rates of Estimation for Scale Parameters in Location Mixture Models. *Annals of Statistics*, 24, 1560-1571.

[31] Ishwaran, H., 1999. Information in Semiparametric Mixtures of Exponential Families. *Annals of Statistics*, 27, 159-177.

[32] Khan, S., 2012. Distribution Free Estimation of Heteroskedastic Binary Choice Models Using Probit Criterion Functions. *Journal of Econometrics,* forthcoming.

[33] Khan, S., D. Nekipelov, 2012. Information Structure and Statistical Information in Discrete Response Models. Duke and UC Berkeley, *unpublished manuscript*.

[34] Khan, S., E. Tamer, 2010. Irregular Identification, Support Conditions, and Inverse Weight Estimation. *Econometrica*, 78, 2021-2042.

[35] Koenker, R. 2005. *Quantile Regressions*. Cambridge University Press, Cambridge.

[36] Koenker, R., G. Bassett, 1978. Regression Quantiles. *Econometrica*, 46, 33-50.

[37] Konishi, S., G. Kitagawa, 2008. *Information Criteria and Statistical Modeling*, New York: Springer.

[38] Newey, W.K., 1990. Semiparametric Efficiency Bounds. *Journal of Applied Econometrics*, 5, 99-135.

[39] Newey, W.K., 1997. Convergence Rates and Asymptotic Normality for Series Estimators. *Journal of Econometrics,* 79, 147-168.

[40] Powell, J., 1994. Estimation of Semiparametric Models. In: Engle III, R.F., McFadden, D.F. (Eds.), *Handbook of Econometrics*, Vol. 4. North-Holland, Amsterdam.

[41] Ridder, G., Woutersen, T., 2003. The Singularity of the Information Matrix of the Mixed Proportional Hazard Model. *Econometrica*, 71, 1579-1589.

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

[43] Shen, X., J. Shi, 2005. Sieve Likelihood Ratio Inference on General Parameter Space. *Science in China Series A: Mathematics*, 48(1), 67-78.

[44] Shen, X., W. Wong 1994. Convergence Rate of Sieve Estimates. *Annals of Statistics,* 22, 580-615.

[45] Stone, C., 1990. Large-sample Inference for Log-spline Models. *Annals of Statistics*, 18(2), 717-741.

[46] Van de Geer, S. 2000. *Empirical Processes in M Estimation*. Cambridge University Press.

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

[48] Van der Vaart, A., J. Wellner, 1996. *Weak Convergence and Empirical Processes: with Applications to Statistics*. New York: Springer-Verlag.

[49] White, H., J. Wooldridge, 1991. Some Results on Sieve Estimation with Dependent Observations. In Barnett, W.A., J. Powell and G. Tauchen (eds.), *Non-parametric and Semi-parametric Methods in Econometrics and Statistics*, 459-493, Cambridge: Cambridge University Press.

[50] Wong, W., T. Severini, 1991. On Maximum Likelihood Estimation in Infinite Dimensional Parameter Spaces. *Annals of Statistics,* 19, 603-632.

[51] Zhou, S., Shen, X. Wolfe, D.A., 1998. Local Asymptotics for Regression Splines and Confidence Regions. Annals of Statistics, 26, 1760–1782.

Appendix

A Proof of the main results

Proof of Proposition 2.4. We establish the result by verifying that Assumptions 2.2 and 2.3 of Lemma 2.1 are satisfied. Define $\|v\|_{\sup} = ||v_\theta||_E + \sum_{j=1}^q \sup_{x_j \in \mathcal{X}_j} \left|v_{h_j}(x_j)\right|$ for any $v = (v_\theta, v_{h_1}, ..., v_{h_q}) \in \mathcal{V}$. Then under Condition 2.1(v), we deduce that for all $v = (v_\theta, v_{h_1}, ..., v_{h_q}) \in \mathcal{V}_n$ with $v_{h_j} = P_{k_n}^j(\cdot)'\beta_{k_n}^j$ and $\beta_{k_n}' = (\beta_{k_n}^{1\prime}, ..., \beta_{k_n}^{q\prime})$,

$$\begin{aligned}\sup_{v\in\mathcal{V}_n} \frac{\|v\|_{\sup}^2}{\|v\|_{sd}^2} &\leq \frac{q+1}{\tau(1-\tau)} \frac{||v_\theta||_E^2 + \sum_{j=1}^q \sup_{x\in\mathcal{X}_j} \left|v_{h_j}^2(x)\right|}{(v_\theta', \beta_{k_n}')E[\overline{P}_{k_n}(X)\overline{P}_{k_n}(X)'](v_\theta', \beta_{k_n}')'} \\ &\lesssim \frac{||v_\theta||_E^2 + \xi_n^2||\beta_{k_n}||^2}{||v_\theta||_E^2 + ||\beta_{k_n}||^2} \lesssim \xi_n^2. \end{aligned} \tag{A.1}$$

Thus for $u_n^* \equiv v_n^* / \|v_n^*\|_{sd}$ we have:

$$\|u_n^*\|_{\sup} = \sup_x \frac{|v_n^*(x)|}{\|v_n^*\|_{sd}} \lesssim \xi_n \quad \text{and} \tag{A.2}$$

$$E\left[|u_n^*(X)|^2\right] = \frac{E\left[|v_n^*(X)|^2\right]}{\tau(1-\tau)E\left[|v_n^*(X)|^2\right]} = \frac{1}{\tau(1-\tau)}. \tag{A.3}$$

In this example we have

$$\begin{aligned}\ell(Z,\alpha) &= [1\{Y \leq \alpha(X)\} - \tau][Y - \alpha(X)] \\ &= -|Y - \alpha(X)| \times [\tau 1\{Y > \alpha(X)\} + (1-\tau)1\{Y \leq \alpha(X)\}],\end{aligned}$$

and $\Delta(Z,\alpha_0)[\alpha - \alpha_0] = -(1\{U \leq 0\} - \tau)[\alpha(X) - \alpha_0(X)]$ with $U = Y - \alpha_0(X)$. By definition, $\Delta(Z,\alpha_0)[v] = -(1\{U \leq 0\} - \tau)v(X)$ is linear in v and hence Assumption 2.2(i) is satisfied. For any $\alpha \in \mathcal{N}_n$, by Knight's identity we have:

$$\ell(Z,\alpha) - \ell(Z,\alpha_0) = \Delta(Z,\alpha_0)[\alpha - \alpha_0] + r(Z,\alpha,\alpha_0)$$

where conditional on X,

$$r(Z,\alpha,\alpha_0) = -\int_0^{\alpha(X)-\alpha_0(X)} (1\{U \leq t\} - 1\{U \leq 0\})\, dt.$$

By definition, $\alpha^* - \alpha = \varepsilon_n u_n^*$ with $\varepsilon_n = o(n^{-1/2})$, hence, with changing of variable $t = \varepsilon_n s + (\alpha(X) - \alpha_0(X))$,

$$\begin{aligned}&\ell(Z,\alpha^*) - \ell(Z,\alpha) - \Delta(Z,\alpha_0)[\varepsilon_n u_n^*] = r(Z,\alpha^*,\alpha_0) - r(Z,\alpha,\alpha_0) \\ =& \int_{\alpha^*(X)-\alpha_0(X)}^{\alpha(X)-\alpha_0(X)} (1\{U \leq t\} - 1\{U \leq 0\})dt \\ =& \varepsilon_n \int_{u_n^*(X)}^0 (1\{U \leq \varepsilon_n s + (\alpha(X) - \alpha_0(X))\} - 1\{U \leq 0\})ds. \end{aligned} \tag{A.4}$$

Therefore Assumption 2.2(ii) is satisfied if the following stochastic equicontinuity result holds:

$$\sup_{\alpha\in\mathcal{N}_n}\mu_n\left\{\int_{u_n^*(X)}^0(1\{U\le\varepsilon_n s+(\alpha(X)-\alpha_0(X))\}-1\{U\le 0\})ds\right\}=o_P(n^{-1/2}). \tag{A.5}$$

Denote $m(Z,\alpha)=\int_{u_n^*(X)}^0(1\{Y\le\varepsilon_n s+\alpha(X)\}-1\{U\le 0\})ds$. Following the verification of Examples 1 and 2 in CLvK, we can show that for all $\delta\in(0,1]$,

$$\begin{aligned}
&E[\sup_{\alpha\in\mathcal{N}_n}\sup_{\alpha'\in\mathcal{N}_n:||\alpha'-\alpha||_\infty\le\delta}\left|m(Z,\alpha')-m(Z,\alpha)\right|^2]\\
\lesssim\ & E\left[\sup_{\alpha\in\mathcal{N}_n}1\{u_n^*(X)\le 0\}\int_{u_n^*(X)}^0 E\left([1\{Y\le\varepsilon_n s+\alpha(X)+\delta\}-1\{Y\le\varepsilon_n s+\alpha(X)-\delta\}]|X\right)ds\right]\\
&+E\left[\sup_{\alpha\in\mathcal{N}_n}1\{u_n^*(X)>0\}\int_0^{u_n^*(X)} E\left([1\{Y\le\varepsilon_n s+\alpha(X)+\delta\}-1\{Y\le\varepsilon_n s+\alpha(X)-\delta\}]|X\right)ds\right]\\
\equiv\ & S_{1,n}+S_{2,n}.
\end{aligned}\tag{A.6}$$

Let $F_{Y|X}(\cdot)$ denote the conditional cdf of Y given X. Under Conditions 2.1(i) and 2.3(i) and equation (A.3), we have:

$$\begin{aligned}
S_{1,n} &= E\left[\sup_{\alpha\in\mathcal{N}_n}1\{u_n^*(X)\le 0\}\int_{u_n^*(X)}^0\left[F_{Y|X}(\varepsilon_n s+\alpha(X)+\delta)-F_{Y|X}(\varepsilon_n s+\alpha(X)-\delta)\right]ds\right]\\
&\lesssim \delta.
\end{aligned}\tag{A.7}$$

Similarly we have $S_{2,n}\lesssim\delta$. Let $\mathcal{F}_n=\{m(Z,\alpha):\alpha\in\mathcal{N}_n\}$. Applying Theorem 3 in CLvK or Lemma 4.2 in Chen (2007), Condition 2.1(i)(ii) with $\gamma>1$ and equations (A.6)-(A.7) now imply that $\int_0^\infty H_{[]}^{1/2}(w,\mathcal{F}_n,\|\cdot\|_{L^2(P_Z)})dw<\infty$ and that (A.5) is satisfied. Hence Assumption 2.2(ii) is verified. Denote $K(\alpha_0,\alpha)\equiv E\left[\ell(Z,\alpha_0)-\ell(Z,\alpha)\right]\ge 0$. For any $\alpha\in\mathcal{N}_n$ we have

$$\begin{aligned}
0 &\le K(\alpha_0,\alpha)=E\left\{\int_0^{\alpha(X)-\alpha_0(X)}[F(s|X)-F(0|X)]\,ds\right\}\\
&= \frac{\|\alpha-\alpha_0\|^2}{2}+E\left[\int_0^{\alpha(X)-\alpha_0(X)}\int_0^s[f(u|X)-f(0|X)]duds\right]\\
&\equiv \frac{\|\alpha-\alpha_0\|^2}{2}+I(\alpha).
\end{aligned}$$

By Condition 2.3(i), we can deduce that for any $\alpha\in\mathcal{N}_n$,

$$\begin{aligned}
|I(\alpha)| &= \left|E\left[\int_0^{\alpha(X)-\alpha_0(X)}\int_0^s[f(u|X)-f(0|X)]duds\right]\right|\\
&\le cE\left[1\{\alpha(X)-\alpha_0(X)\ge 0\}\int_0^{\alpha(X)-\alpha_0(X)}\left(\int_0^s udu\right)ds\right]\\
&\quad +cE\left[1\{\alpha(X)-\alpha_0(X)\le 0\}\int_{\alpha(X)-\alpha_0(X)}^0\left(\int_s^0 udu\right)ds\right]\\
&\lesssim E\left[|\alpha(X)-\alpha_0(X)|^3\right]\le\|\alpha-\alpha_0\|_s^2\|\alpha-\alpha_0\|_{\sup}.
\end{aligned}$$

By Condition 2.3(ii), Remark 2.2 and the definition of $\mathcal{N}_n$, we obtain:

$$\sup_{\alpha \in \mathcal{N}_n} |I(\alpha)| = o(n^{-1}).$$

Hence Assumption 2.2(iii) is satisfied.

By equation (A.2) and $\left(\xi_n n^{-1/2}\right) = o(1)$ (which is implied by Condition 2.3(ii)), there is a $\kappa > 0$ such that

$$\lim_{n \to \infty} \left\{ n^{-\frac{\kappa}{2}} \left(E[|v_n^*(X)|^2] \right)^{-\frac{2+\kappa}{2}} E\left[|v_n^*(X)|^{2+\kappa} \right] \right\} = 0.$$

This and the Lyapounov CLT imply Assumption 2.3. The claimed result now follows from Lemma 2.1.

■

Proof of Lemma 3.1. Given Assumption 3.1, using virtually the same proof as that of Lemma 5.1 in Chen, Sun and Liao (2012) for time series data, we obtain:

$$\left| \frac{||\widehat{v}_n^*||^2}{||v_n^*||^2} - 1 \right| = o_P(1) \text{ and } \frac{||\widehat{v}_n^* - v_n^*||}{||v_n^*||} = o_P(1). \tag{A.8}$$

Given Assumption 3.2(i) and the second result in (A.8), and by the triangle inequality,

$$\begin{aligned} o_P(1) &= \frac{||\widehat{v}_n^* - v_n^*||_{sd}}{||v_n^*||_{sd}} \geq \left| \frac{||\widehat{v}_n^*||_{sd}}{||v_n^*||_{sd}} - 1 \right| = \left| \frac{||\widehat{v}_n^*||_{sd}}{||\widehat{v}_n^*||_{sd,n}} \frac{||\widehat{v}_n^*||_{sd,n}}{||v_n^*||_{sd}} - 1 \right| \\ &= \left| \frac{||\widehat{v}_n^*||_{sd}}{||\widehat{v}_n^*||_{sd,n}} \left(\frac{||\widehat{v}_n^*||_{sd,n}}{||v_n^*||_{sd}} - 1 \right) + \left(\frac{||\widehat{v}_n^*||_{sd}}{||\widehat{v}_n^*||_{sd,n}} - 1 \right) \right|. \end{aligned} \tag{A.9}$$

For any $v_1, v_2 \in \mathcal{V}_n$, we define

$$\langle v_1, v_2 \rangle_{sd} \equiv E\left\{ \Delta(Z, \alpha_0)[v_1] \Delta(Z, \alpha_0)[v_2] \right\}$$

and

$$\langle v_1, v_2 \rangle_{sd,n} \equiv \frac{1}{n} \sum_{i=1}^{n} \Delta(Z_i, \widehat{\alpha}_n)\left[v_1\right] \Delta(Z_i, \widehat{\alpha}_n)\left[v_2\right].$$

By Assumption 3.2 and the triangle inequality, we have

$$\begin{aligned} &\frac{\left| \|\widehat{v}_n^*\|_{sd,n}^2 - ||\widehat{v}_n^*||_{sd}^2 \right|}{||\widehat{v}_n^*||_{sd}^2} \asymp \frac{\left| \|\widehat{v}_n^*\|_{sd,n}^2 - ||\widehat{v}_n^*||_{sd}^2 \right|}{||\widehat{v}_n^*||^2} \leq \sup_{\alpha \in \mathcal{N}_n, v \in \mathcal{V}_n} \frac{|\langle v, v \rangle_{sd,n} - \langle v, v \rangle_{sd}|}{||v||^2} \\ \leq &\sup_{\alpha \in \mathcal{N}_n, v \in \mathcal{W}_n} \left| \mu_n \left[\left(\Delta(Z, \alpha)[v] \right)^2 \right] \right| + \sup_{\alpha \in \mathcal{N}_n, v \in \mathcal{W}_n} \left| E\left[\left(\Delta(Z, \alpha)[v] \right)^2 - \left(\Delta(Z, \alpha_0)[v] \right)^2 \right] \right| \\ = &\ o_P(1). \end{aligned} \tag{A.10}$$

Hence

$$\left| \frac{||\widehat{v}_n^*||_{sd,n}}{||\widehat{v}_n^*||_{sd}} - 1 \right| = o_P(1). \tag{A.11}$$

This and (A.9) imply that

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

■

Proof of Corollary 3.3. For Result (1), note that under Assumption 3.1, we have (A.8) holds. By $\|v\|_{sd} = \|v\|$ for all $v \in \mathcal{V}_n$, we have $||v_n^*||_{sd} = ||v_n^*||$. Hence the claimed result follows if we could show that

$$\left| \frac{\|\widehat{v}_n^*\|_n}{||v_n^*||} - 1 \right| \to_P 0. \tag{A.12}$$

Using the triangle inequality and the second result in (A.8), we have

$$\begin{aligned} o_P(1) &= \frac{||\widehat{v}_n^* - v_n^*||}{||v_n^*||} \geq \left| \frac{||\widehat{v}_n^*||}{||v_n^*||} - 1 \right| \\ &= \left| \frac{||\widehat{v}_n^*||}{||\widehat{v}_n^*||_n} \left(\frac{||\widehat{v}_n^*||_n}{||v_n^*||} - 1 \right) + \left(\frac{||\widehat{v}_n^*||}{||\widehat{v}_n^*||_n} - 1 \right) \right|. \end{aligned} \tag{A.13}$$

Let $E_Z[\cdot]$ denote the expectation taken w.r.t. Z. By Assumption 3.1(i)(ii) and the triangle inequality, we have

$$\begin{aligned} \frac{\left| \|\widehat{v}_n^*\|_n^2 - ||\widehat{v}_n^*||^2 \right|}{||\widehat{v}_n^*||^2} &= \frac{| - \frac{1}{n}\sum_{i=1}^n r(Z_i, \widehat{\alpha}_n)[\widehat{v}_n^*, \widehat{v}_n^*] + E_Z[r(Z, \alpha_0)[\widehat{v}_n^*, \widehat{v}_n^*]] |}{||\widehat{v}_n^*||^2} \\ \leq \sup_{\alpha \in \mathcal{N}_n, v \in \mathcal{W}_n} |\mu_n[r(Z, \alpha)[v, v]]| &+ \sup_{\alpha \in \mathcal{N}_n, v \in \mathcal{W}_n} |E[r(Z, \alpha)[v, v] - r(Z, \alpha_0)[v, v]]| = o_P(1). \end{aligned} \tag{A.14}$$

Hence $|\|\widehat{v}_n^*\|_n / ||\widehat{v}_n^*|| - 1| = o_P(1)$, which, together with (A.13), implies $|\|\widehat{v}_n^*\|_n / ||v_n^*|| - 1| = o_P(1)$.

For Result (2), since $\|v\|_{sd} = \|v\|$ for all $v \in \mathcal{V}_n$, we know that the Riesz representers of the functional $\frac{\partial f(\alpha_0)}{\partial \alpha}[\cdot]$ defined w.r.t. $\|\cdot\|_{sd}$ and $\|\cdot\|$ are the same. Using the similar arguments in showing (A.8) but replacing Assumption 3.1 with Assumption 3.3, we can show that

$$\left| \frac{||\widetilde{v}_n^*||_{sd}^2}{||v_n^*||_{sd}^2} - 1 \right| = o_P(1) \text{ and } \frac{||\widetilde{v}_n^* - v_n^*||_{sd}}{||v_n^*||_{sd}} = o_P(1). \tag{A.15}$$

Using the triangle inequality and the second result in (A.15), we have

$$\begin{aligned} o_P(1) &= \frac{||\widetilde{v}_n^* - v_n^*||_{sd}}{||v_n^*||_{sd}} \geq \left| \frac{||\widetilde{v}_n^*||_{sd}}{||v_n^*||_{sd}} - 1 \right| \\ &= \left| \frac{||\widetilde{v}_n^*||_{sd}}{||\widetilde{v}_n^*||_{sd,n}} \left(\frac{||\widetilde{v}_n^*||_{sd,n}}{||v_n^*||_{sd}} - 1 \right) + \left(\frac{||\widetilde{v}_n^*||_{sd}}{||\widetilde{v}_n^*||_{sd,n}} - 1 \right) \right|. \end{aligned} \tag{A.16}$$

Using Assumption 3.3(i)(ii) and the triangle inequality, we can use similar arguments in showing (A.14) to get

$$\frac{\left| \|\widetilde{v}_n^*\|_{sd,n}^2 - ||\widetilde{v}_n^*||_{sd}^2 \right|}{||\widetilde{v}_n^*||_{sd}^2} = o_P(1).$$

Hence $|\,\|\widetilde{v}_n^*\|_{sd,n}\,/||\widetilde{v}_n^*||_{sd}-1| = o_P(1)$, which, together with (A.16), implies $|\,||\widetilde{v}_n^*||_{sd,n}/\,||v_n^*||_{sd}-1| = o_P(1)$. ■

Proof of Lemma 3.2. Using the inverse formula of partitioned matrix, we have

$$\begin{aligned}
&(-E_n\{r(Z_i,\widehat{\alpha}_n)[\overline{P}_{k_n}(X_i),\overline{P}_{k_n}(X_i)]\})^{-1} \\
= &\begin{pmatrix} I_n^{-1} & -I_n^{-1}I_{12,n}I_{22,n}^{-1} \\ -J_n^{-1}I_{21,n}I_{11,n}^{-1} & J_n^{-1} \end{pmatrix} \equiv \begin{pmatrix} A_{11} & -A_{12} \\ -A_{21} & A_{22} \end{pmatrix}. \qquad \text{(A.17)}
\end{aligned}$$

We can decompose $E_n\{\Delta(Z_i,\widehat{\alpha}_n)[\overline{P}_{k_n}(X_i)]\Delta(Z_i,\widehat{\alpha}_n)[\overline{P}_{k_n}(X_i)]'\}$ as

$$E_n\{\Delta(Z_i,\widehat{\alpha}_n)[\overline{P}_{k_n}(X_i)]\Delta(Z_i,\widehat{\alpha}_n)[\overline{P}_{k_n}(X_i)]'\} \equiv \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}, \qquad \text{(A.18)}$$

where

$$\begin{aligned}
B_{11} &= E_n\left\{\Delta(Z_i,\widehat{\alpha}_n)[P_{k_n}(X_i)]\Delta(Z_i,\widehat{\alpha}_n)[P_{k_n}(X_i)]'\right\}, \\
B_{12} &= B_{21}' = E_n\left\{\Delta(Z_i,\widehat{\alpha}_n)[P_{k_n}(X_i)]\Delta(Z_i,\widehat{\alpha}_n)[P_{d_\theta}(e)]'\right\}, \\
B_{22} &= E_n\left\{\Delta(Z_i,\widehat{\alpha}_n)[P_{d_\theta}(e)]\Delta(Z_i,\widehat{\alpha}_n)[P_{d_\theta}(e)]'\right\}.
\end{aligned}$$

By the definition of $\widehat{V}_{k_n\times k_n}$, the decompositions in (A.17) and (A.18), we get

$$\widehat{V}_{k_n\times k_n} = A_{11}B_{11}A_{11} - A_{12}B_{21}A_{11} - A_{11}B_{12}A_{21} + A_{12}B_{22}A_{21}.$$

Next, note that

$$\begin{aligned}
||\widehat{v}_{h,n}^*||_{sd,n}^2 &= E_n\{\Delta(Z_i,\widehat{\alpha}_n)[\widehat{v}_{h,n}^*]\Delta(Z_i,\widehat{\alpha}_n)[\widehat{v}_{h,n}^*]'\} \\
&= \begin{pmatrix} I_n^{-1}P_{k_n}(\overline{x}) \\ -I_{22,n}^{-1}I_{21,n}I_n^{-1}P_{k_n}(\overline{x}) \end{pmatrix}' \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \begin{pmatrix} I_n^{-1}P_{k_n}(\overline{x}) \\ -I_{22,n}^{-1}I_{21,n}I_n^{-1}P_{k_n}(\overline{x}) \end{pmatrix} \\
&= P_{k_n}(\overline{x})'I_n^{-1}B_{11}I_n^{-1}P_{k_n}(\overline{x}) - P_{k_n}(\overline{x})'I_n^{-1}I_{12,n}I_{22,n}^{-1}B_{21}I_n^{-1}P_{k_n}(\overline{x}) \\
&\quad -P_{k_n}(\overline{x})'I_n^{-1}B_{12}I_{22,n}^{-1}I_{21,n}I_n^{-1}P_{k_n}(\overline{x}) + P_{k_n}(\overline{x})'I_n^{-1}I_{12,n}I_{22,n}^{-1}B_{22}I_{22,n}^{-1}I_{21,n}I_n^{-1}P_{k_n}(\overline{x}) \\
&= P_{k_n}(\overline{x})'[A_{11}B_{11}A_{11} - A_{12}B_{21}A_{11} - A_{11}B_{12}A_{21} + A_{12}B_{22}A_{21}]P_{k_n}(\overline{x}) \\
&= P_{k_n}(\overline{x})'\widehat{V}_{k_n\times k_n}P_{k_n}(\overline{x}) = \widehat{Var}[\widehat{h}(\overline{x})]
\end{aligned}$$

which gives Result (1).

We next show the second result. By the definition of $\widehat{V}_{d_\theta\times d_\theta}$, the decompositions in (A.17) and (A.18), we get

$$\widehat{V}_{d_\theta\times d_\theta} = A_{21}B_{11}A_{12} - A_{22}B_{21}A_{12} - A_{21}B_{12}A_{22} + A_{22}B_{22}A_{22}.$$

By definition, we have

$$\begin{aligned}
||\widehat{v}^*_{\theta,n}||^2_{sd,n} &= E_n\{\Delta(Z_i,\widehat{\alpha}_n)[\widehat{v}^*_{\theta,n}]\Delta(Z_i,\widehat{\alpha}_n)[\widehat{v}^*_{\theta,n}]'\} \\
&= \begin{pmatrix} -I^{-1}_{11.n}I_{12,n}J^{-1}_n\lambda \\ J^{-1}_n\lambda \end{pmatrix}' \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \begin{pmatrix} -I^{-1}_{11,n}I_{12,n}J^{-1}_n\lambda \\ J^{-1}_n\lambda \end{pmatrix} \\
&= \lambda' J^{-1}_n I_{21,n} I^{-1}_{11,n} B_{11} I^{-1}_{11,n} I_{12,n} J^{-1}_n\lambda - \lambda' J^{-1}_n I_{21,n} I^{-1}_{11,n} B_{12} J^{-1}_n\lambda \\
&\quad -\lambda' J^{-1}_n B_{21} I^{-1}_{11,n} I_{12,n} J^{-1}_n\lambda + \lambda' J^{-1}_n B_{22} J^{-1}_n\lambda \\
&= \lambda'[A_{21}B_{11}A_{12} - A_{22}B_{21}A_{12} - A_{21}B_{12}A_{22} + A_{22}B_{22}A_{22}]\lambda \\
&= \lambda'\widehat{V}_{d_\theta\times d_\theta}\lambda,
\end{aligned}$$

which gives Result (2). ■

Proof of Theorem 4.1. As $f_j(\cdot)$ satisfies Assumption 2.1(i)(ii), using similar arguments in the proof of Theorem 2.1, we can deduce that (for $j = 1, ..., m$)

$$\left|\left\langle \widehat{\alpha}_n - \alpha_0, u^*_{j,n}\right\rangle - \mu_n\left\{\Delta(Z,\alpha_0)[u^*_{j,n}]\right\}\right| = o_P(n^{-\frac{1}{2}}) \tag{A.19}$$

which together with Assumption 4.1(iii) implies that (for $j = 1, ..., m$)

$$\left|\langle u^*_{j,n}, \widehat{\alpha}_n - \alpha_0\rangle\right| \le \left|\mu_n\left\{\Delta(Z,\alpha_0)[u^*_{j,n}]\right\}\right| + o_P(n^{-\frac{1}{2}}) = O_P(n^{-\frac{1}{2}}). \tag{A.20}$$

We first consider that $\widehat{\alpha}^*_n = \widehat{\alpha}_n - \sum_{j=1}^m \langle u^*_{j,n}, \widehat{\alpha}_n - \alpha_0\rangle u^*_{j,n}$. Using (A.19), (A.20) and Assumption 4.1(i)-(ii), we deduce that

$$\begin{aligned}
& L_n(\widehat{\alpha}_n) - L_n(\widehat{\alpha}^*_n) \\
= \;& E[\ell(Z,\widehat{\alpha}_n) - \ell(Z,\widehat{\alpha}^*_n)] + \mu_n\left\{\Delta(Z,\alpha_0)[\widehat{\alpha}_n - \widehat{\alpha}^*_n]\right\} \\
& +\mu_n\left\{\ell(Z,\widehat{\alpha}_n) - \ell(Z,\widehat{\alpha}^*_n) - \Delta(Z,\alpha_0)[\widehat{\alpha}_n - \widehat{\alpha}^*_n]\right\} \\
= \;& \frac{1}{2}||\widehat{\alpha}^*_n - \widehat{\alpha}_n||^2 + \langle \widehat{\alpha}^*_n - \widehat{\alpha}_n, \widehat{\alpha}_n - \alpha_0\rangle + \mu_n\left\{\Delta(Z,\alpha_0)[\widehat{\alpha}_n - \widehat{\alpha}^*_n]\right\} + o_P(n^{-1}) \\
= \;& \frac{1}{2}||\widehat{\alpha}^*_n - \widehat{\alpha}_n||^2 - \sum_{j=1}^m \left\langle \widehat{\alpha}_n - \alpha_0, u^*_{j,n}\right\rangle\left[\left\langle \widehat{\alpha}_n - \alpha_0, u^*_{j,n}\right\rangle - \mu_n\left\{\Delta(Z,\alpha_0)[u^*_{n,j}]\right\}\right] + o_P(n^{-1}) \\
= \;& \frac{1}{2}||\widehat{\alpha}^*_n - \widehat{\alpha}_n||^2 + o_P(n^{-1}).
\end{aligned} \tag{A.21}$$

By definition $\mathbf{f}(\alpha_0) = 0$, which together with Assumption 2.1(ii) implies that $\langle \alpha - \alpha_0, u^*_{j,n}\rangle = o(n^{-\frac{1}{2}})$ for $\alpha \in \{\alpha \in \mathcal{N}_n : \mathbf{f}(\alpha) = 0\}$ and $j = 1, ..., m$. Hence for the constrained sieve M estimator $\widetilde{\alpha}_n$, we have $\langle \widetilde{\alpha}_n - \alpha_0, u^*_{j,n}\rangle = o_P(n^{-\frac{1}{2}})$, which together with (A.20) implies that

$$\langle \widehat{\alpha}_n - \widehat{\alpha}^*_n, \widetilde{\alpha}_n - \alpha_0\rangle = \sum_{j=1}^m \langle u^*_{n,j}, \widehat{\alpha}_n - \alpha_0\rangle\left\langle u^*_{n,j}, \widetilde{\alpha}_n - \alpha_0\right\rangle = o_P(n^{-1}). \tag{A.22}$$

Next denote $\widetilde{\alpha}_n^* = \widetilde{\alpha}_n + \sum_{j=1}^m \langle u_{n,j}^*, \widehat{\alpha}_n - \alpha_0 \rangle u_{n,j}^*$. As $\left(\frac{\partial f_1(\alpha_0)}{\partial \alpha}[\cdot], ..., \frac{\partial f_m(\alpha_0)}{\partial \alpha}[\cdot] \right)$ is linearly independent, we can assume that $\langle v_{n,j_1}^*, v_{n,j_2}^* \rangle = 0$ for any $j_1 \neq j_2$. Using the result in (A.22), we have

$$\begin{aligned}
\langle \widetilde{\alpha}_n - \widetilde{\alpha}_n^*, \widetilde{\alpha}_n^* - \alpha_0 \rangle &= -\langle \widehat{\alpha}_n - \widehat{\alpha}_n^*, \widetilde{\alpha}_n^* - \alpha_0 \rangle \\
&= -\langle \widehat{\alpha}_n - \widehat{\alpha}_n^*, \widetilde{\alpha}_n^* - \widetilde{\alpha}_n \rangle - \langle \widehat{\alpha}_n - \widehat{\alpha}_n^*, \widetilde{\alpha}_n - \alpha_0 \rangle \\
&= -\sum_{j=1}^m \langle u_{j,n}^*, \widehat{\alpha}_n - \alpha_0 \rangle \langle u_{j,n}^*, \widetilde{\alpha}_n^* - \widetilde{\alpha}_n \rangle + o_P(n^{-1}) \\
&= -\sum_{j=1}^m \left| \langle u_{j,n}^*, \widehat{\alpha}_n - \alpha_0 \rangle \right|^2 \left\| u_{j,n}^* \right\|^2 + o_P(n^{-1}) \\
&= -\sum_{j=1}^m \left| \langle u_{j,n}^*, \widehat{\alpha}_n - \alpha_0 \rangle \right|^2 + o_P(n^{-1}) \qquad \text{(A.23)}
\end{aligned}$$

where the third equality is by (A.22), the fourth inequality is by $\langle v_{n,j_1}^*, v_{n,j_2}^* \rangle = 0$ for any $j_1 \neq j_2$, the last equality is by $||v_{j,n}^*|| = ||v_{j,n}^*||_{sd}$. As $\widetilde{\alpha}_n \in \mathcal{N}_n$, using (A.23) and similar arguments in deriving (A.22), we get

$$\begin{aligned}
& L_n(\widetilde{\alpha}_n^*) - L_n(\widetilde{\alpha}_n) \\
= \; & E[\ell(Z, \widetilde{\alpha}_n^*) - \ell(Z, \widetilde{\alpha}_n)] + \mu_n \{\Delta(Z, \alpha_0)[\widetilde{\alpha}_n^* - \widetilde{\alpha}_n]\} \\
& + \mu_n \{\ell(Z, \widetilde{\alpha}_n^*) - \ell(Z, \widetilde{\alpha}_n) - \Delta(Z, \alpha_0)[\widetilde{\alpha}_n^* - \widetilde{\alpha}_n]\} \\
= \; & \frac{1}{2} ||\widetilde{\alpha}_n - \widetilde{\alpha}_n^*||^2 + \langle \widetilde{\alpha}_n - \widetilde{\alpha}_n^*, \widetilde{\alpha}_n^* - \alpha_0 \rangle + \mu_n \{\Delta(Z, \alpha_0)[\widetilde{\alpha}_n^* - \widetilde{\alpha}_n]\} + o_P(n^{-1}) \\
= \; & \frac{1}{2} ||\widetilde{\alpha}_n - \widetilde{\alpha}_n^*||^2 + \sum_{j=1}^m \langle u_{n,j}^*, \widehat{\alpha}_n - \alpha_0 \rangle \left[\mu_n \left\{ \Delta(Z, \alpha_0)[u_{n,j}^*] \right\} - \langle u_{j,n}^*, \widehat{\alpha}_n - \alpha_0 \rangle \right] + o_P(n^{-1}) \\
= \; & \frac{1}{2} ||\widetilde{\alpha}_n - \widetilde{\alpha}_n^*||^2 + o_P(n^{-1}) = \frac{1}{2} ||\widehat{\alpha}_n - \widehat{\alpha}_n^*||^2 + o_P(n^{-1}). \qquad \text{(A.24)}
\end{aligned}$$

By the definition of $\widehat{\alpha}_n$ and the result in (A.24), we deduce that

$$L_n(\widehat{\alpha}_n) - L_n(\widetilde{\alpha}_n) \geq L_n(\widetilde{\alpha}_n^*) - L_n(\widetilde{\alpha}_n) = \frac{1}{2} ||\widehat{\alpha}_n - \widehat{\alpha}_n^*||^2 + o_P(n^{-1}). \qquad \text{(A.25)}$$

Next, using similar trick in Shen and Shi (2005), we can find some $\alpha_{\mathbf{f},n}^* \in \{\alpha \in \mathcal{N}_n : \mathbf{f}(\alpha) = 0\}$ such that

$$L_n(\widehat{\alpha}_n^*) - L_n(\alpha_{\mathbf{f},n}^*) = o_P(n^{-1}). \qquad \text{(A.26)}$$

Using the definition of $\widetilde{\alpha}_n$ and the results in (A.21) and (A.26), we get

$$\begin{aligned}
L_n(\widehat{\alpha}_n) - L_n(\widetilde{\alpha}_n) &\leq L_n(\widehat{\alpha}_n) - L_n(\widehat{\alpha}_n^*) + L_n(\widehat{\alpha}_n^*) - L_n(\alpha_{\mathbf{f},n}^*) \\
&= \frac{1}{2} ||\widehat{\alpha}_n^* - \widehat{\alpha}_n||^2 + o_P(n^{-1}). \qquad \text{(A.27)}
\end{aligned}$$

Equations (A.19), (A.25) and (A.27) imply that

$$\begin{aligned} 2n\left[L_n(\widehat{\alpha}_n) - L_n(\widetilde{\alpha}_n)\right] &= n||\widehat{\alpha}_n^* - \widehat{\alpha}_n||^2 + o_P(1) = n\sum_{j=1}^m \left[\langle u_{n,j}^*, \widehat{\alpha}_n - \alpha_0\rangle\right]^2 + o_P(1) \\ &= \sum_{j=1}^m \left[\sqrt{n}\mu_n\left\{\Delta(Z,\alpha_0)[u_{n,j}^*]\right\} + o_P(1)\right]^2 + o_P(1) \end{aligned} \tag{A.28}$$

where the second equality uses $\langle v_{n,j_1}^*, v_{n,j_2}^*\rangle = 0$ for any $j_1 \neq j_2$ and $||v_{j,n}^*|| = ||v_{j,n}^*||_{sd}$. Now (4.2) follows by applying Assumption 4.1(iii) and the Continuous Mapping Theorem in (A.28). ■

Proof of Proposition 5.6. We establish the result by verifying that assumptions of Lemma 2.1 and Corollary 3.3(2) are satisfied. Let $\mathcal{F}_n = \left\{\frac{\partial \log[p(Z,\alpha)]}{\partial\alpha}[u_n^*] - \frac{\partial \log[p(Z,\alpha_0)]}{\partial\alpha}[u_n^*] : \alpha \in \mathcal{N}_n\right\}$. Under Condition 5.4(i) we have $\log N_{[]}(\varepsilon, \mathcal{F}_n, \|\cdot\|_{L^2(P)}) \lesssim \log N(\varepsilon, \mathcal{A}_n, \|\cdot\|_A)$ with $\mathcal{A}_n = \Theta \times \mathcal{H}_n$. Since $\int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{A}_n, \|\cdot\|_A)}d\varepsilon < \infty$, applying Lemma 4.2 in Chen (2007) we obtain:

$$\sup_{\alpha\in\mathcal{N}_n} \mu_n \left\{\frac{\partial \log[p(Z,\alpha)]}{\partial\alpha}[u_n^*] - \frac{\partial \log[p(Z,\alpha_0)]}{\partial\alpha}[u_n^*]\right\} = o_P(n^{-\frac{1}{2}})$$

which implies that Assumption 2.2(ii)' (and hence Assumption 2.2(ii)) is satisfied. Next note that

$$\begin{aligned} -K(\alpha_0,\alpha) &= E\left[\frac{\partial \log[p(Z,\alpha_0)]}{\partial\alpha}[\alpha-\alpha_0] + \frac{\partial^2 \log[p(Z,\alpha_0)]}{2\partial\alpha\partial\alpha'}[\alpha-\alpha_0,\alpha-\alpha_0]\right. \\ &\quad \left. + \frac{\partial^2 \log[p(Z,\widetilde{\alpha})]}{2\partial\alpha\partial\alpha'}[\alpha-\alpha_0,\alpha-\alpha_0] - \frac{\partial^2 \log[p(Z,\alpha_0)]}{2\partial\alpha\partial\alpha'}[\alpha-\alpha_0,\alpha-\alpha_0]\right] \end{aligned} \tag{A.29}$$

where $\widetilde{\alpha}$ lies between α and α_0 (pathwisely). By the restriction $\int p(z,\alpha_0)dz = 1$, we have for all $v \in \mathcal{V}$:

$$E\left[\frac{\partial \log[p(Z,\alpha_0)]}{\partial\alpha}[v]\right] = 0 \text{ and } E\left[p^{-1}(Z,\alpha_0)\frac{\partial^2 p(Z,\alpha_0)}{\partial\alpha\partial\alpha'}[v,v]\right] = 0 \tag{A.30}$$

which implies that

$$E\left[\frac{\partial^2 \log p(Z,\alpha_0)}{2\partial\alpha\partial\alpha'}[\alpha-\alpha_0,\alpha-\alpha_0]\right] = -\frac{1}{2}||\alpha-\alpha_0||_{sd}^2. \tag{A.31}$$

Using Condition 5.4(ii), uniformly over $\alpha \in \mathcal{N}_n, \widetilde{\alpha} \in \mathcal{N}_0$,

$$\begin{aligned} &E\left[\left|\frac{\partial^2 \log[p(Z,\widetilde{\alpha})]}{\partial\alpha\partial\alpha}[\alpha-\alpha_0,\alpha-\alpha_0] - \frac{\partial^2 \log[p(Z,\alpha_0)]}{\partial\alpha\partial\alpha}[\alpha-\alpha_0,\alpha-\alpha_0]\right|\right] \\ \leq\ &\|\alpha-\alpha_0\|_{sd}^2 \sup_{v\in\mathcal{N}_0:\|v\|_{sd}=1} E\left[\left|\frac{\partial^2 \log[p(Z,\widetilde{\alpha})]}{\partial\alpha\partial\alpha}[v,v] - \frac{\partial^2 \log[p(Z,\alpha_0)]}{\partial\alpha\partial\alpha}[v,v]\right|\right] = o(n^{-1}). \end{aligned}$$

Hence Assumption 2.2(iii)" is verified. Condition 5.4(iii) directly assumes the Lyapounov CLT Assumption 2.3'. Hence all assumptions of Lemma 2.1 are satisfied.

It remains to verify Assumption 3.3 of Corollary 3.3(2). First we note that by the definition of $\mathcal{W}_n$, we have: $\sup_{v\in\mathcal{W}_n}\left\|\frac{\partial \log[p(Z,\alpha_0)]}{\partial\alpha}[v]\right\|_{L^2(P)} \leq const.$ By Condition 5.5(i), Hölder inequality and the triangle inequality,

$$\begin{aligned}
& \sup_{\alpha\in\mathcal{N}_n, v_1,v_2\in\mathcal{W}_n}\left|E\left[\frac{\partial \log[p(Z,\alpha)]}{\partial\alpha}[v_1]\frac{\partial \log[p(Z,\alpha)]}{\partial\alpha}[v_2]\right]-\langle v_1,v_2\rangle\right| \\
\leq & \sup_{\alpha\in\mathcal{N}_n, v\in\mathcal{W}_n}\left\|\frac{\partial \log[p(Z,\alpha)]}{\partial\alpha}[v]-\frac{\partial \log[p(Z,\alpha_0)]}{\partial\alpha}[v]\right\|^2_{L^2(P)} \\
& +2\sup_{v\in\mathcal{W}_n}\left\|\frac{\partial \log[p(Z,\alpha_0)]}{\partial\alpha}[v]\right\|_{L^2(P)}\sup_{\alpha\in\mathcal{N}_n, v\in\mathcal{W}_n}\left\|\frac{\partial \log[p(Z,\alpha)]}{\partial\alpha}[v]-\frac{\partial \log[p(Z,\alpha_0)]}{\partial\alpha}[v]\right\|_{L^2(P)} \\
= & o(1),
\end{aligned}$$

hence Assumption 3.3(i) is satisfied. For any $\alpha,\alpha'\in\mathcal{N}_n$ and $v_1,v_2,v_3,v_4\in\mathcal{W}_n$, using the triangle inequality and Hölder inequality, we have

$$\begin{aligned}
& \left|\frac{\partial \log p(Z,\alpha)}{\partial\alpha}[v_1]\frac{\partial \log p(Z,\alpha)}{\partial\alpha}[v_2]-\frac{\partial \log p(Z,\alpha')}{\partial\alpha}[v_3]\frac{\partial \log p(Z,\alpha')}{\partial\alpha}[v_4]\right| \\
\leq & \left|\frac{\partial \log p(Z,\alpha)}{\partial\alpha}[v_1]-\frac{\partial \log p(Z,\alpha')}{\partial\alpha}[v_3]\right|\times\left|\frac{\partial \log p(Z,\alpha)}{\partial\alpha}[v_2]\right| \\
& +\left|\frac{\partial \log p(Z,\alpha)}{\partial\alpha}[v_2]-\frac{\partial \log p(Z,\alpha')}{\partial\alpha}[v_4]\right|\times\left|\frac{\partial \log p(Z,\alpha')}{\partial\alpha}[v_3]\right| \\
\leq & \left[I_1(Z,\alpha,\alpha';v_1,v_3)+I_2(Z,\alpha,\alpha';v_2,v_4)\right]\times\sup_{\alpha\in\mathcal{N}_n, v\in\mathcal{W}_n}\left|\frac{\partial \log p(Z,\alpha)}{\partial\alpha}[v]\right|, \qquad \text{(A.32)}
\end{aligned}$$

with

$$I_1(Z,\alpha,\alpha';v_1,v_3)\leq\left|\frac{\partial \log p(Z,\alpha)}{\partial\alpha}[v_1]-\frac{\partial \log p(Z,\alpha')}{\partial\alpha}[v_1]\right|+\left|\frac{\partial \log p(Z,\alpha')}{\partial\alpha}[v_1-v_3]\right|, \qquad \text{(A.33)}$$

$$I_2(Z,\alpha,\alpha';v_2,v_4)\leq\left|\frac{\partial \log p(Z,\alpha)}{\partial\alpha}[v_2]-\frac{\partial \log p(Z,\alpha')}{\partial\alpha}[v_2]\right|+\left|\frac{\partial \log p(Z,\alpha')}{\partial\alpha}[v_2-v_4]\right|. \qquad \text{(A.34)}$$

By Condition 5.5(iii) and the definition of $\mathcal{W}_n$ we have:

$$\begin{aligned}
& \sup_{\alpha'\in\mathcal{N}_n, v_1,v_3\in\mathcal{W}_n: v_1\neq v_3}\left|\frac{\partial \log p(Z,\alpha')}{\partial\alpha}[v_1-v_3]\right| \\
\leq & \|v_1-v_3\|_{sd}\times\sup_{\alpha'\in\mathcal{N}_n, v\in\mathcal{W}_n}\left|\frac{\partial \log p(Z,\alpha')}{\partial\alpha}[v]\right|\leq\|v_1-v_3\|_{sd}\times D_{3,n}(Z). \qquad \text{(A.35)}
\end{aligned}$$

By inequalities (A.32), (A.33), (A.34) and (A.35), Conditions 5.5(ii) and (iii), and the definitions of $\mathcal{N}_n$ and $\mathcal{W}_n$, we have that the class $\mathcal{F}_n^{sd}=\left\{\frac{\partial \log p(Z,\alpha)}{\partial\alpha}[v_1]\frac{\partial \log p(Z,\alpha)}{\partial\alpha}[v_2]:\alpha\in\mathcal{N}_n, v_1,v_2\in\mathcal{W}_n\right\}$ has finite $L^1(P)-$covering number with bracketing. This and Theorem 2.4.1 of van der Vaart and Wellner (1996) imply that

$$\sup_{\alpha\in\mathcal{N}_n, v_1,v_2\in\mathcal{W}_n}\mu_n\left\{\frac{\partial \log p(Z,\alpha)}{\partial\alpha}[v_1]\frac{\partial \log p(Z,\alpha)}{\partial\alpha}[v_2]\right\}=o_P(1),$$

which verifies Assumption 3.3(ii). Assumption 3.3(iii) is directly assumed. Hence Assumption 3.3 holds and by Corollary 3.3(2), we have $|\,||\widetilde{v}_n^*||_{sd,n}/\,||v_n^*||_{sd} - 1| = o_P(1)$. ■

Proof of Lemma 6.1. Let k^* be any fixed finite positive integer. The corresponding parametric M estimator with k^* many basis functions are defined as $\widehat{\alpha}_{n,k^*} = (\widehat{\theta}_{n,k^*}, \widehat{h}_{n,k^*})$. Using Assumption 6.1(iii) and the assumption that $C_n/n = o(1)$, we have

$$\begin{aligned} n^{-1}I_n(k^*) &= -\frac{2}{n}\sum_{i=1}^{n}\ell(Z_i, \widehat{\theta}_{n,k^*}, \widehat{h}_{n,k^*}) + \frac{C_n(d_\theta + k^*)}{n} \\ &= -2E_Z\left[\ell(Z, \widehat{\theta}_{n,k^*}, \widehat{h}_{n,k^*})\right] - 2\mu_n\left[\ell(Z, \widehat{\theta}_{n,k^*}, \widehat{h}_{n,k^*})\right] + o(1) \\ &= -2E_Z\left[\ell(Z, \widehat{\theta}_{n,k^*}, \widehat{h}_{n,k^*})\right] + o_P(1) \end{aligned} \tag{A.36}$$

where $E_Z[\cdot]$ denotes the expectation taken w.r.t. the distribution of Z. On the other hand, using Assumption 6.1(iv)(v) and the assumption that $k_n^* C_n/n = o(1)$, we have

$$\begin{aligned} n^{-1}I_n(k_n^*) &= -\frac{2}{n}\sum_{i=1}^{n}\ell(Z_i, \widehat{\theta}_{n,k_n^*}, \widehat{h}_{n,k_n^*}) + \frac{C_n(d_\theta + k_n^*)}{n} \\ &= -\frac{2}{n}\sum_{i=1}^{n}\ell(Z_i, \theta_0, h_0) - \frac{2}{n}\sum_{i=1}^{n}\left[\ell(Z_i, \widehat{\theta}_{n,k_n^*}, \widehat{h}_{n,k_n^*}) - \ell(Z_i, \theta_0, h_0)\right] + o(1) \\ &= -2E_Z\left[\ell(Z_i, \theta_0, h_0)\right] + o_P(1). \end{aligned} \tag{A.37}$$

Equations (A.36) and (A.37) imply that

$$\frac{I_n(k^*) - I_n(k_n^*)}{n} = 2E_Z\left[\ell(Z, \theta_0, h_0) - \ell(Z, \widehat{\theta}_{n,k^*}, \widehat{h}_{n,k^*})\right] + o_P(1)$$

which, together with Assumption 6.1(ii), implies that $\frac{I_n(k^*) - I_n(k_n^*)}{n} > 0$ w.p.a.1. This finishes the proof. ■

B Tables

Table B.1 Sieve ML Estimation of Semiparametric Mixture Model

	$n = 100$ and $k_n = 1$				$n = 100$ and $k_n = 2$			
	Mean	Bias	Std	RMSE	Mean	Bias	Std	RMSE
$\widehat{\theta}_{1,n}$	0.8766	0.1234	0.0803	0.1472	0.8788	0.1212	0.0792	0.1448
$\widehat{\theta}_{2,n}$	0.8764	0.1236	0.1646	0.2058	0.8785	0.1215	0.1653	0.2051
	$n = 100$ and $k_n = 3$				$n = 100$ and $k_n = 4$			
	Mean	Bias	Std	RMSE	Mean	Bias	Std	RMSE
$\widehat{\theta}_{1,n}$	0.9893	0.0107	0.1327	0.1332	1.0276	0.0276	0.1404	0.1431
$\widehat{\theta}_{2,n}$	0.9891	0.0109	0.2130	0.2133	1.0263	0.0263	0.2215	0.2231
	$n = 500$ and $k_n = 2$				$n = 500$ and $k_n = 3$			
	Mean	Bias	Std	RMSE	Mean	Bias	Std	RMSE
$\widehat{\theta}_{1,n}$	0.8647	0.1353	0.0359	0.1400	0.9584	0.0416	0.0577	0.0712
$\widehat{\theta}_{2,n}$	0.8636	0.1364	0.0722	0.1543	0.9572	0.0428	0.0902	0.0998
	$n = 500$ and $k_n = 4$				$n = 500$ and $k_n = 5$			
	Mean	Bias	Std	RMSE	Mean	Bias	Std	RMSE
$\widehat{\theta}_{1,n}$	0.9604	0.0396	0.0613	0.0729	1.0134	0.0134	0.0975	0.0984
$\widehat{\theta}_{2,n}$	0.9599	0.0401	0.0912	0.0996	1.0128	0.0128	0.1215	0.1222

Table B.1: 5000 simulated samples are used to calculate the finite sample mean(Mean), bias (Bias), standard deviation (Std) and root of mean square error (RMSE) of the sieve MLEs of the structural coefficients.

Table B.2 Sieve ML Estimation Based on AIC and BIC

	$n = 100$ and AIC ($\widehat{k}_n = 1.2$)				$n = 100$ and BIC ($\widehat{k}_n = 1.0$)			
	Mean	Bias	Std	RMSE	Mean	Bias	Std	RMSE
$\widehat{\theta}_{1,n}$	0.8903	0.1097	0.0997	0.1482	0.8768	0.1232	0.0806	0.1473
$\widehat{\theta}_{2,n}$	0.8905	0.1095	0.1786	0.2095	0.8766	0.1234	0.1652	0.2062
	$n = 500$ and AIC ($\widehat{k}_n = 3.8$)				$n = 500$ and BIC ($\widehat{k}_n = 2.2$)			
	Mean	Bias	Std	RMSE	Mean	Bias	Std	RMSE
$\widehat{\theta}_{1,n}$	0.9579	0.0421	0.0700	0.0817	0.8754	0.1246	0.0493	0.1340
$\widehat{\theta}_{2,n}$	0.9566	0.0434	0.0982	0.1074	0.8738	0.1262	0.0798	0.1493

Table B.2: 5000 simulated samples are used to calculate the finite sample mean(Mean), bias (Bias), standard deviation (Std) and root of mean square error (RMSE) of the sieve MLEs of the structural coefficients.

Table B.3 Inference Based on Gaussian and Chi-square Approximation

	$n = 50$ and $k_n = 3$				$n = 100$ and $k_n = 1$			
	CP-θ_{10}	LT-θ_{10}	CP-θ_{20}	LT-θ_{20}	CP-θ_{10}	LT-θ_{10}	CP-θ_{20}	LT-θ_{20}
G.N.	0.8510	0.5990	0.8730	0.9922	0.4740	0.2539	0.8090	0.5734
C.S.	0.9190	0.6409	0.9290	1.1335	0.4970	0.2583	0.8250	0.5873
	$n = 100$ and $k_n = 2$				$n = 100$ and $k_n = 3$			
	CP-θ_{10}	LT-θ_{10}	CP-θ_{20}	LT-θ_{20}	CP-θ_{10}	LT-θ_{10}	CP-θ_{20}	LT-θ_{20}
G.N.	0.4710	0.2479	0.7920	0.5596	0.8550	0.4317	0.8900	0.6866
C.S.	0.5280	0.2638	0.8270	0.5914	0.9070	0.4618	0.9300	0.7913
	$n = 100$ and $k_n = 4$				$n = 150$ and $k_n = 4$			
	CP-θ_{10}	LT-θ_{10}	CP-θ_{20}	LT-θ_{20}	CP-θ_{10}	LT-θ_{10}	CP-θ_{20}	LT-θ_{20}
G.N.	0.8580	0.4558	0.9010	0.7284	0.8600	0.3772	0.8970	0.5930
C.S.	0.9250	0.4796	0.9390	0.8266	0.9110	0.3881	0.9320	0.6777

Table B.3: 1000 simulated samples are used and inferences are conducted based on 0.95 confidence level. CP and LT denote the convergence probability and length of confidence interval for the related structural coefficients respectively. G.N. and C.S. denote the Gaussian approximation and Chi-square approximation respectively.

Table B.4 Duration Analysis of the Second Birth in China[1]

AIC	Sieve MLE	Std	Gaussian CIs[5]	Chi-square CIs[5]
log*(Duration)*	3.3621	0.0975	(3.1710 3.5532)	(3.2720 3.4496)
Constant	-5.3175	0.2624	(-5.8317 -4.8032)	(-5.4516 -5.1844)
Gender of 1st kid[2]	0.0584	0.1256	(-0.1877 0.3045)	(-0.1182 0.2322)
Years of schooling	0.0088	0.0181	(-0.0267 0.0442)	(-0.0102 0.0275)
Bonus[3]	-0.2270	0.1331	(-0.4880 0.0339)	(-0.4164 -0.0418)
Household type[4]	0.7900	0.1769	(0.4433 1.1366)	(0.6427 0.9365)
BIC	Sieve MLE	Std	Gaussian CIs	Chi-square CIs
log*(Duration)*	3.2404	0.0718	(3.0090 3.2905)	(3.0642 3.2321)
Constant	-4.9435	0.2284	(-5.3232 -4.4281)	(-5.0064 -4.7476)
Gender of 1st kid	0.0397	0.0840	(-0.0993 0.2298)	(-0.1100 0.2309)
Years of schooling	0.0075	0.0157	(-0.0253 0.0364)	(-0.0129 0.0235)
Bonus	-0.2836	0.1118	(-0.4816 -0.0434)	(-0.4459 -0.0818)
Household type	0.7326	0.1591	(0.4291 1.0526)	(0.5974 0.8824)

Table B.4: 1. Sample size n=694; 2. gender dummy variable equals 1 if the 1st kid is a girl and 0 otherwise; 3. bonus dummy variable equals 1 if there are subsidies awarded to the household accepting the one child policy; 4. household type is 1 if it is registered in rural area and 0 otherwise; 5. confidence intervals are constructed under 0.95