Economics Working Paper Series

2017/014

Inference in Nonparametric Series Estimation with Data-Dependent Undersmoothing

Byunghoon Kang

The Department of Economics
Lancaster University Management School
Lancaster LA1 4YX
UK

LUMS home page: http://www.lancaster.ac.uk/lums/

Inference in Nonparametric Series Estimation with Data-Dependent Undersmoothing

Byunghoon Kang*

Department of Economics, Lancaster University

First version December 9, 2014; Revised May 16, 2017

Abstract

Existing asymptotic theory for inference in nonparametric series estimation typically imposes an undersmoothing condition that the number of series terms is sufficiently large to make bias asymptotically negligible. However, there is no formally justified data-dependent method for this in practice. This paper constructs inference methods for nonparametric series regression models and introduces tests based on the infimum of t-statistics over different series terms. First, I provide an empirical process theory for the t-statistics indexed by the number of series terms. Using this result, I show that test based on the infimum of the t-statistics and its asymptotic critical value controls asymptotic size with undersmoothing condition. Using this test, we can construct a valid confidence interval (CI) by test statistic inversion that has correct asymptotic coverage probability. Allowing asymptotic bias without the undersmoothing condition, I show that CI based on the infimum of the t-statistics bounds coverage distortions. In an illustrative example, nonparametric estimation of wage elasticity of the expected labor supply from Blomquist and Newey (2002), proposed CI is close to or tighter than those based on the standard CI with the possible ad hoc choice of series terms.

Keywords: Nonparametric series regression, Pointwise confidence interval, Smoothing parameter choice, Specification search, Undersmoothing.

JEL classification: C12, C14.

*This paper is a revised version of my first chapter in the Ph.D. thesis at UW-Madison. I am deeply indebted to Bruce Hansen for his continuous guidance and suggestions. I am also grateful to Jack Porter, Xiaoxia Shi and Joachim Freyberger for thoughtful discussions. Thanks to Michal Kolesár, Denis Chetverikov, Yixiao Sun, Andres Santos, Patrik Guggenberger, Federico Bugni, Joris Pinkse, Liangjun Su, and Myung Hwan Seo for helpful conversations and criticism. I acknowledge support by Kwanjeong Educational Foundation Graduate Research Fellowship and Leon Mears Dissertation Fellowship from UW-Madison. All errors are my own. Email: b.kang1@lancaster.ac.uk, Homepage: https://sites.google.com/site/davidbhkang

1 Introduction

I consider the following nonparametric regression model;

$$\begin{aligned} y_i &= g_0(x_i) + \varepsilon_i, \\ E(\varepsilon_i | x_i) &= 0 \end{aligned} \tag{1.1}$$

where $\{y_i, x_i\}_{i=1}^n$ is i.i.d. with scalar response variable y_i, vector of covariates $x_i \in \mathbb{R}^{d_x}$, and $g_0(x) = E(y_i | x_i = x)$ is the conditional mean function. Examples falling into the model (1.1) include nonparametric estimation of the Mincer equation, gasoline demand, and labor supply function (see, among many others, Heckman, Lochner and Todd (2006), Hausman and Newey (1995), Blomquist and Newey (2002), Blundell and MaCurdy (1999), and references therein). Addressing potential misspecification of the parametric model, nonparametric series methods have several advantages, as they can easily impose shape restrictions such as additive separability or concavity, and implementation is easy because the estimation method is least squares. However, implementation in practice requires a choice of *the number of series terms*, K. Estimation and inference may largely depend on its choice in finite samples. Moreover, required K may vary with different data sets to accommodate the smoothness of unknown function and different sample sizes, as well as whether the goal is estimation or inference.

Existing theory for the asymptotic normality and valid inference imposes so-called *undersmoothing* (i.e., *overfitting*) condition that is a faster rate of K than the mean-squared error (MSE) optimal convergence rates to make bias asymptotically negligible relative to standard deviation. The undersmoothing condition has been imposed, particularly for valid inference, in many nonparametric series methods both in theory and in practice, as there is no theory for bias-corrections available to date. Ignoring asymptotic bias with the undersmoothing assumption, one can apply the conventional confidence interval (CI) using the standard normal critical value with estimates and standard errors based on some choice of "sufficiently large" K (larger than MSE optimal K). However, the asymptotic theory does not provide specific guidelines for choosing a "large" number of series terms to make bias small in practice. With given sample sizes n, some possibly ad hoc methods in practice select $\widehat{K} = \widetilde{K} \cdot n^{\gamma}$ with some pre-selected $\widetilde{K}$ and a specific rate of γ that satisfies the undersmoothing level, which is generally unknown. However, there is no formally justified data-dependent method to choose K that gives the desired level of undersmoothing in series regression literature.

Due to these unsatisfactory results for the inference procedure both in theory and practice, a specification search seems necessary, i.e., search over different series terms $K \in [\underline{K}, \bar{K}]$. For example, a researcher may use quadratic, cubic, or quartic terms in polynomial regres-

sion, or try a different number of knots in regression spline to see how the estimate and standard error change. Moreover, some data-dependent selection rules that are valid for estimation (such as cross-validation) and some rule-of-thumb methods that are suggested for inference, also require evaluating estimates with different Ks. If researchers evaluate different specifications with a different number of series terms and select one specification as a baseline model, it is not clear how this randomness affects the standard inference.

In this paper, I construct inference methods in nonparametric series regression given the range of different series terms. I consider the testing problem for a regression function at a point and introduce tests based on *infimum of the studentized t-statistics* over different series terms. To describe intuition heuristically, we may decompose infimum t-statistic as follows

$$\inf_K |T_n(K)| \approx \inf_K |N(0,1) + \frac{Bias(K)}{SE(K)}|$$

where $T_n(K), Bias(K), SE(K)$ denote t-statistic, bias and standard error of the series estimator using K terms, respectively. The test based on infimum t-statistics and searching for small t-statistics have a similar motivation to the one on which the undersmoothing condition is theoretically based: using faster rates of K than the optimal MSE rate (using "large" K that has a small bias and large variance) so that makes the second term, $\frac{Bias(K)}{SE(K)}$, small. Many papers in nonparametric series estimation literature typically suggested to increase the number of series terms and include additional terms than those cross-validation chooses for inference (for example, see Newey (2013), Newey, Powell, and Vella (1999)). Although I do not consider data-dependent methods that satisfy desired undersmoothing rates in this paper, I formally justify this conventional wisdom by introducing the infimum test statistic and provide an inference method based on its asymptotic distribution as an alternative data-dependent undersmoothing.

For this, I first provide an empirical process theory for the t-statistics, which I shall call *t-statistic process*, indexed by the number of series terms. The main contribution of this paper is to derive a uniform asymptotic distribution theory for the entire sequences of t-statistics over a range of K. Existing asymptotic normality of the t-statistic in the literature holds under a deterministic sequence of $K \to \infty$ as the sample size increases. I impose an assumption on the set of deterministic sequences $\mathcal{K}_n$ where the number of series terms $K \in \mathcal{K}_n$ can be indexed by continuous parameter π, a 'fraction' of the largest series terms $\bar{K}$, and this is important for our purpose to show the weak convergence of the empirical process.

Using this result, I show that test based on the infimum of the t-statistics and its asymptotic critical value control the asymptotic size (null rejection probability) with the under-

smoothing condition for all Ks in a set. Allowing asymptotic bias without the undersmoothing condition, I also analyze the effect of bias on the asymptotic size of the test. Even allowing the asymptotic bias, the test based on the infimum t-statistic bound the size distortions, in the sense that the asymptotic size is bounded above by the asymptotic size of a test with a t-statistic that has the smallest bias. The infimum t-statistic is less sensitive to the asymptotic bias: it naturally excludes small K with large bias and selects among some large Ks under the null.

I also construct a valid pointwise confidence interval for the conditional mean function that has nominal asymptotic coverage probability by test statistic inversion. The proposed CI based on infimum test statistic can be easily constructed using estimates and standard errors for the set of Ks. It is obtained as the union of all CIs by replacing the standard normal critical value with the critical value from the asymptotic distribution of the infimum t-statistic. We can approximate the asymptotic critical value using a simple Monte Carlo or weighted bootstrap method. Similar to the asymptotic size results, I show that proposed CI bounds the coverage distortions even when asymptotic bias exists. I also find that our proposed CI performs well in Monte Carlo experiments; coverage probability of the CI based on the infimum t-statistic is close to the nominal level in various simulation setups. As an illustrative example, I revisit nonparametric estimation of wage elasticity of the expected labor supply, as in Blomquist and Newey (2002). Given the table in Blomquist and Newey (2002), the proposed CI is tighter than the standard CI with the largest number of series terms as well as close to the standard CI with some "large" K.

This paper also provides a valid CI after selecting the number of series terms. By adjusting the conventional normal critical value to the critical value from supremum of the t-statistics over all series terms, we can adjust uncertainty due to the choice of series terms. This paper gives a valid post-selection CI that has a correct coverage with any choice of $\widehat{K}$ among some ranges. By enlarging the CI with critical values larger than the normal critical value, this post-selection CI can accommodate bias, although it does not explicitly deal with bias problems. We expect this lead to a tighter CI than those based on the Bonferroni-type critical value, as I incorporate the dependence structure of the t-statistics from our asymptotic distribution theory.

I also investigate inference methods in partially linear model setup. Focusing on the common parametric part, choice problems also occur for the number of approximating terms or the number of covariates in estimating the nonparametric part. Unlike the nonparametric object of interest that has a slower convergence than $n^{1/2}$ rate (e.g., regression function or regression derivative), t-statistics for the parametric object of interest are asymptotically equivalent for all sequences of K under standard rate conditions $K/n \to 0$ as $n \to \infty$. To

fully account for the dependency of the t-statistics with the different sequences of Ks in the partially linear model setup, this requires a different approximation theory than standard first order approximation results. Using the recent results of Cattaneo, Jansson, and Newey (2015a), I develop a joint asymptotic distribution of the studentized t-statistics over a different number of series terms. By focusing on the faster rate of K that grows as fast as the sample size n and using larger variance than the standard variance formula, we can account for the dependency of t-statistics with different Ks. In this setup, I also propose methods to construct CIs that are similar to the nonparametric regression setup and provide their asymptotic coverage properties.

1.1 Related literature

The literature on nonparametric series estimation is vast, but data-dependent series term selection and its impact on estimation or inference are comparatively less developed. Perhaps the most widely used data-dependent rule in practice is cross-validation. Asymptotic optimality results have been developed (see, for example, Li (1987), Andrews (1991b), Hansen (2015)) in terms of asymptotic equivalence between integrated mean squared error (IMSE) of the nonparametric estimator with $\widehat{K}_{\texttt{cv}}$ selected by minimizing the cross-validation criterion and IMSE of the infeasible optimal estimator. However, there are two problems with cross-validation selected $\widehat{K}_{\texttt{cv}}$ for the valid inference. First, it is asymptotically equivalent to selecting K to minimize IMSE, and thus it does not satisfy the undersmoothing condition needed for asymptotic normality without bias terms. Therefore, a t-statistic based on $\widehat{K}_{\texttt{cv}}$ will be asymptotically invalid for inference. Second, $\widehat{K}_{\texttt{cv}}$ selected by cross-validation will itself be random and not deterministic. Thus, it is not clear whether the t-statistic based on $\widehat{K}_{\texttt{cv}}$ has a standard asymptotic normal distribution which is derived from a deterministic sequence of K.

Novel recent papers by Horowitz (2014), Chen and Christensen (2015a) develop the state-of-the-art data-dependent methods in the nonparametric instrumental variables (NPIV) estimation (see also other references therein). They develop data-driven methods for choosing sieve dimension in that resulting NPIV estimators attain the optimal sup-norm or L^2 norm rates adaptive to the unknown smoothness of $g_0(x)$. In this paper, we focus on the inference rather than estimation with the similar issues arise from using cross-validation.

Moreover, this paper is also closely related to the previous methods that conceptually require increasing K until t-statistic is "small enough". For example, among many others, Newey (2013) suggested increasing K until standard errors are large relative to small changes in objects of interest, Newey, Powell, and Vella (1999) suggested using more terms than

those cross-validation chooses, and Horowitz and Lee (2012) suggested increasing K until the integrated variance suddenly increases and then adding an additional term. They discuss these methods work well in practice and simulation. Using similar ideas, I provide formal inference methods based on asymptotic distribution results of the infimum test statistic with appropriate critical values smaller than the standard normal critical values.

Several important papers have investigated the asymptotic properties of series (and sieves) estimators, including papers by Andrews (1991a), Eastwood and Gallant (1991), Newey (1997), Chen and Shen (1998), Huang (2003a), Chen (2007), Chen and Liao (2014), Chen, Liao, and Sun (2014), Belloni, Chernozhukov, Chetverikov, and Kato (2015), and Chen and Christensen (2015b), among many others. Under i.i.d. or weakly dependent data, they focused on Sup/L^2-norm convergence rates, asymptotic normality of series estimators, and pointwise/uniform inference on linear/nonlinear functionals under a deterministic sequence of K. This paper extends the asymptotic normality of the t-statistic under a single sequence of K to the uniform central limit theorem of the t-statistic for the sequences of K over a set, and focuses on a pointwise inference on $g_0(x)$, which is an irregular (i.e., slower than $n^{1/2}$ rate) and linear functional, under i.i.d. data.

For the kernel-based density or regression estimation, the data-dependent bandwidth selection problem is well known. Several rule-of-thumb methods and plug-in optimal bandwidths have been proposed (see Härdle and Linton (1994), Li and Racine (2007) for references). A recent paper by Calonico, Cattaneo and Farrell (2015) compared higher-order coverage properties of undersmoothing and explicit bias-corrections and derived coverage optimal bandwidth choices in kernel estimation. See also Hall and Horowitz (2013), Schennach (2015) and references therein for various recent work on related bias issues and nonparametric inference for the kernel estimator. Unlike the kernel-based methods, little is known about the statistical properties of data-dependent selection rules (e.g., rates of $\widehat{K}_{\mathtt{cv}}$) and asymptotic distribution with data-dependent methods in series estimation. In general, the main technical difficulty arises from the lack of an explicit asymptotic bias formula for the series estimator (see Zhou, Shen, and Wolfe (1998) and Huang (2003b) for exceptions with some specific sieves). Thus, it is difficult to derive an asymptotic theory for the bias-correction, or some plug-in formulas compare with kernel estimation.

An important recent paper that is concurrent with this paper, Armstrong and Kolesár (2015) considered inference methods in kernel estimation with bandwidth snooping. Focusing on the supremum of the t-statistics over the bandwidths, they developed confidence intervals that are uniform in bandwidths. Considering supremum statistic is motivated by the sensitivity analysis as a correction for the multiple testing problems. Moreover, considering different bandwidths and the test based on the supremum of the studentized t-statistics

has been used to achieve adaptive inference procedures when smoothness of the function is unknown (See Horowitz and Spokoiny (2001), and also Armstrong (2015)). See Appendix C for the similar coverage results (uniform in series terms) as in Armstrong and Kolesár (2015). The main focus of this paper is to analyze the undersmoothing assumption with their effect on the size of the tests and to develop tests which can control size distortions even allowing large asymptotic bias, which can be crucial in series estimation context.[1]

The outline of the paper is as follows. I first introduce basic nonparametric series regression setup in Section 2. In Section 3, I provide an empirical process theory for the t-statistic sequences over a set. Section 4 introduces infimum of the t-statistic and describes the asymptotic null distributions of the test statistic. Then, I provide the asymptotic size results of the test and implementation procedure for the critical value. Section 5 introduces CIs based on the infimum test statistic and provides their coverage properties. Section 6 analyzes valid post model selection inference in this setup. Section 7 extends our inference methods to the partially linear model setup. Section 8 includes Monte Carlo experiments in various setups. Section 9 illustrates proposed inference methods using the nonparametric estimation of wage elasticity of the expected labor supply, as in Blomquist and Newey (2002), then Section 10 concludes. Appendix A and B include all proofs, figures, and tables. Appendix C discuss inference procedures based on the supremum of the t-statistics.

1.2 Notation

I introduce some notation will be used in the following sections. I use $||A|| = \sqrt{tr(A'A)}$ for the Euclidean norm. Let $\lambda_{min}(A), \lambda_{max}(A)$ denote the minimum and maximum eigenvalues of a symmetric matrix A, respectively. $o_p(\cdot)$ and $O_p(\cdot)$ denote the usual stochastic order symbols, convergence in probability and bounded in probability. $\xrightarrow{d}$ denotes convergence in distribution and $\Rightarrow$ denotes weak convergence. I use the notation $a \wedge b = \min\{a, b\}$, $a \vee b = \max\{a, b\}$, and denote $\lfloor a \rfloor$ as the largest integer less than the real number a. For two sequences of positive real numbers a_n and b_n, $a_n \lesssim b_n$ denotes $a_n \leq cb_n$ for all n sufficiently large with some constant $c > 0$ that is independent of n. $a_n \asymp b_n$ denotes $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For a given random variable $\{X_i\}$ and $1 \leq p < \infty$, $L^p(X)$ is the space of all L^p norm bounded functions with $||f||_{L^p} = [E||f(X_i)||^p]^{1/p}$ and $\ell^\infty(X)$ denotes the space of all bounded functions under sup-norm, $||f||_\infty = \sup_{x \in \mathcal{X}} |f(x)|$ for the bounded real-valued functions f on the support $\mathcal{X}$. Let also $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\}$, $\mathbb{R}_{+,\infty} = \mathbb{R}_+ \cup \{+\infty\}$, $\mathbb{R}_{[+\infty]} = \mathbb{R} \cup \{+\infty\}$ and $\mathbb{R}_{[\pm\infty]} = \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$.

[1] We may also consider other types of test statistics, for example, "median" or "average" of the t-statistics. Any types of test statistics that are continuous transformation of joint t-statistics with its appropriate critical value leads to the tests that control the asymptotic size with undersmoothing.

2 Model framework and estimation

I first introduce the nonparametric series regression setup in the model (1.1). Given a random sample $\{y_i, x_i\}_{i=1}^n$, we are interested in the conditional mean $g_0(x) = E(y_i|x_i = x)$ at a point $x \in \mathcal{X} \subset \mathbb{R}^{d_x}$. All the results derived in this paper are the pointwise inference in x and I will omit the dependence on x if there is no confusion.

We consider the sequence of approximating model indexed by the number of series terms $K \equiv K(n)$. Let $\widehat{g}_K(x)$ be an estimator of $g_0(x)$ using the first K vectors of approximating functions $P_K(x) = (p_1(x), \cdots, p_K(x))'$ from basis functions $p(x) = (p_1(x), p_2(x), \cdots)'$. Standard examples for the basis functions are power series, Fourier series, orthogonal polynomials (e.g., Hermite polynomials), or splines with evenly sequentially spaced knots.

Series estimator $\widehat{g}_K(x)$ is then obtained by standard least square (LS) estimation of y_i on regressors P_{Ki}

$$\widehat{g}_K(x) = P_K(x)'\widehat{\beta}_K, \qquad \widehat{\beta}_K = (P^{K\prime}P^K)^{-1}P^{K\prime}Y \tag{2.1}$$

where $P_{Ki} \equiv P_K(x_i) = (p_1(x_i), p_2(x_i), \cdots, p_K(x_i))', P^K = [P_{K1}, \cdots, P_{Kn}]', Y = (y_1, \cdots y_n)'$. $\widehat{g}_K(x)$ is an estimator of the best linear L^2 approximation for $g_0(x)$, i.e., $P_K(x)'\beta_K$ where β_K is defined as the best linear projection coefficients $\beta_K \equiv (E[P_{Ki}P_{Ki}'])^{-1}E[P_{Ki}y_i]$. Define the approximation error using K series terms as $r_K(x) = g_0(x) - P_K(x)'\beta_K$ for $x \in \mathcal{X}$. Also define $r_{Ki} \equiv r_K(x_i)$, $p_i \equiv p(x_i) = (p_{1i}, p_{2i}, \cdots,)'$. We can write the model using K approximating terms as the following projection model

$$y_i = P_{Ki}'\beta_K + \varepsilon_{Ki}, \quad E[P_{Ki}\varepsilon_{Ki}] = 0 \tag{2.2}$$

where $\varepsilon_{Ki} \equiv r_{Ki} + \varepsilon_i$.

For simplicity of notation, I define the true regression function at a point as $\theta_0 \equiv g_0(x)$. Let $\widehat{\theta}_K \equiv \widehat{g}_K(x)$ and $\theta_K \equiv P_K(x)'\beta_K$. Define the series variance

$$\begin{aligned} &V_K \equiv V_K(x) = P_K(x)'Q_K^{-1}\Omega_K Q_K^{-1}P_K(x), \\ &Q_K = E(P_{Ki}P_{Ki}'), \quad \Omega_K = E(P_{Ki}P_{Ki}'\varepsilon_i^2) \end{aligned} \tag{2.3}$$

where $Q_K^{-1}\Omega_K Q_K^{-1}$ is the conventional asymptotic variance formula for the LS estimator $\widehat{\beta}_K$.

We use notion of testing setup and consider two-sided testing for θ

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0. \tag{2.4}$$

The studentized t-statistic for H_0 is

$$T_n(K, \theta_0) \equiv \frac{\sqrt{n}(\widehat{g}_K(x) - g_0(x))}{V_K^{1/2}} = \frac{\sqrt{n}(\widehat{\theta}_K - \theta_0)}{V_K^{1/2}}. \tag{2.5}$$

Under standard regularity conditions (will be discussed in Section 3) including an undersmoothing rate for deterministic sequence $K \to \infty$ as $n \to \infty$, the asymptotic distribution of the t-statistic is well known

$$T_n(K, \theta_0) \xrightarrow{d} N(0, 1).$$

See, for example, Andrews (1991a), Newey (1997), Belloni et al. (2015), Chen and Christensen (2015b) among many others. In the next section, I formally develop an asymptotic distribution theory of $T_n(K, \theta_0)$ over a set $\mathcal{K}_n$.

3 Asymptotic distribution

3.1 Weak convergence of t-statistic process

In this section, I provide an asymptotic theory of the joint t-statistics over a set. First, I introduce following set $\mathcal{K}_n$ to construct empirical process theory of the t-statistics over $K \in \mathcal{K}_n$ that can be indexed by the continuous and fixed parameter π, which is a 'fraction' of the largest series term $\bar{K}$.

Assumption 3.1. *(Set of number of series terms) Let $\mathcal{K}_n$ as*

$$\mathcal{K}_n = \{K : K \in [\underline{\pi}\bar{K}, \bar{K}]\}$$

where $K = \pi\bar{K}$ with a fixed constant $\pi \in \Pi = [\underline{\pi}, 1]$, $\underline{\pi} > 0$, and $\bar{K} \equiv \bar{K}(n) \to \infty$ as $n \to \infty$.

Assumption 3.1 considers a range of the number of series terms and considers an infinite sequence of approximations indexed by $\pi \in \Pi$. Note that $\mathcal{K}_n$ is indexed by sample size n as I will impose rate conditions for the largest series term $\bar{K}$ in the next Assumption 3.2. Together with the Assumption 3.2 below, set $\mathcal{K}_n$ in Assumption 3.1 considers the sequence of models that has the same rate of K, i.e., $K \asymp K'$ for any $K, K' \in \mathcal{K}_n$. Note that standard inference methods in nonparametric regression setup typically consider a singleton set $\mathcal{K}_n = \{K\}$ with $K \to \infty$ as $n \to \infty$.

Next, define the following empirical process, $T_n^*(\pi, \theta)$, as

$$T_n^*(\pi, \theta) \equiv T_n(\lfloor \pi \bar{K} \rfloor, \theta), \quad \pi \in \Pi, \tag{3.1}$$

where $T_n(K, \theta)$ is defined in (2.5), i.e., $T_n^*(\pi, \theta)$ is a t-statistic evaluated at the parameter θ using $\lfloor \pi \bar{K} \rfloor$ number of series terms. Note that $T_n^*(\pi, \theta)$ is indexed by $\pi \in \Pi$ and is a step function of π.

Also, I impose mild regularity conditions that are standard in nonparametric series regression literature and are satisfied by well-known basis functions. For each $K \in \mathcal{K}_n$, define $\zeta_K \equiv \sup_{x \in \mathcal{X}} ||P_K(x)||$ as the largest normalized length of the regressor vector and $\lambda_K \equiv (\lambda_{min}(Q_K))^{-1/2}$ for $K \times K$ design matrix $Q_K = E(P_{Ki} P_{Ki}')$.

Assumption 3.2. *(Regularity conditions)*

(i) $\{y_i, x_i\}_{i=1}^n$ *are i.i.d random variables satisfying the model* (1.1).

(ii) $\sup_{x \in \mathcal{X}} E(\varepsilon_i^2 | x_i = x) < \infty$, $\inf_{x \in \mathcal{X}} E(\varepsilon_i^2 | x_i = x) > 0$, *and* $\sup_{x \in \mathcal{X}} E(\varepsilon_i^2 \{|\varepsilon_i| > c(n)\} | x_i = x) \to 0$ *for any sequence* $c(n) \to \infty$ *as* $n \to \infty$.

(iii) For each $K \in \mathcal{K}_n$, *as* $K \to \infty$, *there exists* η *and* c_K, ℓ_K *such that*

$$\sup_{x \in \mathcal{X}} |g_0(x) - P_K(x)'\eta| \leq \ell_K c_K, \quad E[(g_0(x_i) - P_K(x_i)'\eta)^2]^{1/2} \leq c_K.$$

(iv) $\sup_{K \in \mathcal{K}_n} \lambda_K \lesssim 1$.

(v) $\sup_{K \in \mathcal{K}_n} \zeta_K \sqrt{(\log K)/n}(1 + \sqrt{K} \ell_K c_K) + \ell_K c_K \to 0$ *as* $n \to \infty$.

I closely follow assumptions in recent papers by Belloni et al. (2015), Chen and Christensen (2015b) and impose rate conditions of K uniformly over $\mathcal{K}_n$. Other standard regularity conditions in the literature (e.g., Newey (1997)) can also be used here. Assumption 3.2(ii) imposes moment conditions and standard uniform integrability conditions. ζ_K, c_K, ℓ_K in Assumption 3.2(iii)-(v) are satisfied with various basis functions. For example, if the support $\mathcal{X}$ is a cartesian product of compact connected intervals (e.g. $\mathcal{X} = [0,1]^{d_x}$) and the probability density of x_i is bounded below zero, then $\zeta_K \lesssim K$ for power series and other orthogonal polynomial series, and $\zeta_K \lesssim \sqrt{K}$ for regression splines, Fourier series, and wavelet series. c_K and ℓ_K in Assumption 3.2(iii) vary with the different basis and can be replaced by series specific bounds. For example, if $g_0(x)$ belongs to the Hölder space of smoothness p, then $c_K \lesssim K^{-p/d_x}, \ell_K \lesssim K$ for power series, $c_K \lesssim K^{-(p \wedge s_0)/d_x}, \ell_K \lesssim 1$ for spline and wavelet series

of order s_0 (see Newey (1997), Chen (2007), Belloni et al. (2015), and Chen and Christensen (2015b) for more discussions on c_K, ℓ_K, ζ_K with various sieve bases).

When the probability density function of x_i is uniformly bounded above and bounded away from zero over the compact support $\mathcal{X}$ and orthonormal basis is used, then $\lambda_K \lesssim 1$ (see, for example, Proposition 2.1 in Belloni et al. (2015) and Remark 2.2 in Chen and Christensen (2015b)). The rate conditions in Assumption 3.2(v) can be replaced by the specific bounds of ζ_K, c_K, ℓ_K. For example, for the power series, Assumption 3.2(v) reduced to $\sup_{K\in\mathcal{K}_n} \sqrt{K^2(\log K)/n}(1 + K^{3/2-p/d_x}) + K^{1-p/d_x} = \sqrt{\bar{K}^2(\log \bar{K})/n}(1 + \bar{K}^{3/2-p/d_x}) + \bar{K}^{1-p/d_x} \to 0$ under Assumption 3.1.

For notational simplicity, it is convenient to define $P_\pi(x) \equiv P_{\lfloor \bar{K}\pi \rfloor}(x)$, $P_{\pi i} \equiv P_\pi(x_i) = P_{\lfloor \bar{K}\pi \rfloor i}$ and $r_\pi \equiv r_\pi(x) = r_{\lfloor \bar{K}\pi \rfloor}(x)$ under Assumption 3.1. The series variance can be defined as $V_\pi \equiv V_\pi(x) = ||\Omega_\pi^{1/2} Q_\pi^{-1} P_\pi(x)||^2$, where $\Omega_\pi = E(P_{\pi i} P_{\pi i}' \varepsilon_i^2), Q_\pi = E(P_{\pi i} P_{\pi i}')$. Under Assumptions 3.1 and 3.2, the t-statistic process under H_0 can be decomposed as follows

$$T_n^*(\pi, \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{P_\pi(x)' P_{\pi i} \varepsilon_i}{V_\pi^{1/2}} - \sqrt{n} V_\pi^{-1/2} r_\pi + o_p(1), \quad \pi \in \Pi \tag{3.2}$$

where the first term converges to a standard normal distribution for any $\pi \in \Pi$, and the second term does not necessarily converge to 0 even with large sample sizes due to approximation errors. We want to show weak convergence of the empirical process $\{T_n^*(\pi, \theta_0) : \pi \in \Pi\}$. This empirical process has a covariance kernel

$$\Sigma_n(\pi_1, \pi_2) \equiv \frac{P_{\pi_1}(x)' E(P_{\pi_1 i} P_{\pi_2 i}' \varepsilon_i^2) P_{\pi_2}(x)}{V_{\pi_1}^{1/2} V_{\pi_2}^{1/2}}, \qquad \pi_1, \pi_2 \in \Pi. \tag{3.3}$$

We expect that the limiting Gaussian process has a covariance function as a limit of the sequence of covariance functions $\Sigma_n(\pi_1, \pi_2)$, which is assumed to exist by the following assumption. I also impose rate restrictions on series variances.

Assumption 3.3.

(i) $\Sigma(\pi_1, \pi_2) = \lim_{n\to\infty} \Sigma_n(\pi_1, \pi_2)$ *exists and* $\Sigma(\pi_1, \pi_2) < 1$ *for any* $\pi_1, \pi_2 \in \Pi$.

(ii) $\lim_{n\to\infty} V_\pi^{1/2} (\bar{K}\pi)^{-\eta} = c$ *uniformly in* $\pi \in \Pi$ *for some constants* $c, \eta > 0$.

Assumption 3.3(i) guarantees well-defined covariance function for the tight Gaussian process in $\ell^\infty(\Pi)$ and a positive definite variance-covariance matrix for its finite dimensional limit distributions. Assumption 3.3(ii) may be stronger than necessary, but required to prove weak convergence of the t-statistic process. Assumption 3.3(ii) requires that series

variance is increasing in K at some rates uniformly in $K \in \mathcal{K}_n$ under Assumption 3.1, i.e., $\lim_{n\to\infty} V_K^{1/2} K^{-\eta} = c$. This assumption holds with $\eta = 1/2$ when we consider a point $x \in \mathcal{X}$ where $V_K^{1/2} \propto K^{1/2}$. Moreover, Assumption 3.3(ii) is a sufficient condition for Assumption 3.3(i) under homoskedasticity. Under conditional homoskedasticity, $E(\varepsilon_i^2 | x_i = x) = \sigma^2$, the limit of the covariance kernel $\Sigma(\pi_1, \pi_2)$ reduces to the simple form

$$\Sigma(\pi_1, \pi_2) = \lim_{n\to\infty} \frac{V_{\pi_1 \wedge \pi_2}^{1/2}}{V_{\pi_1 \vee \pi_2}^{1/2}} \tag{3.4}$$

for any $\pi_1, \pi_2 \in \Pi$, i.e., the covariance kernel is the limit of the ratio of standard deviations. If we further assume Assumption 3.3(ii), then $\Sigma(\pi_1, \pi_2) = (\frac{\pi_1 \wedge \pi_2}{\pi_1 \vee \pi_2})^\eta$. With $\eta = 1$, this coincides with the covariance kernel of a scaled Brownian motion process $\mathbb{Z}(\pi)/\sqrt{\pi}, \pi \in \Pi$.

Next, I define the asymptotic bias for the sequence of models indexed by π as the limit of the second term in (3.2)

$$\nu(\pi) \equiv \lim_{n\to\infty} -\sqrt{n} V_\pi^{-1/2} r_\pi. \tag{3.5}$$

Under the following undersmoothing condition, $\nu(\pi) = 0$ for all $\pi \in \Pi$. To assess the effect of bias on inference, we will consider a distinction between results imposing the undersmoothing condition or not.

Assumption 3.4. *(Undersmoothing)* $\sup_{K \in \mathcal{K}_n} |\sqrt{n} V_K^{-1/2} \ell_K c_K| \to 0$ *as* $n \to \infty$.

When we use explicit bounds $\ell_K c_K \lesssim K^{-p/d_x}$ for spline or wavelet series, Assumption 3.4 reduces to $\sup_{K \in \mathcal{K}_n} \sqrt{n} V_K^{-1/2} K^{-p/d_x} = o(1)$. When we further consider $V_K^{1/2} \propto K^{1/2}$, Assumption 3.1 and 3.4 together imply that Assumption 3.4 is provided by $\sqrt{n} \bar{K}^{1/2 - p/d_x} \to 0$ for power series.

Next theorem is our first main result which provides uniform central limit theorem of the t-statistic process for nonparametric LS series estimation.

Theorem 3.1. *Under Assumptions 3.1, 3.2, 3.3 and* $\sup_\pi |\nu(\pi)| < \infty$,

$$T_n^*(\pi, \theta_0) \Rightarrow \mathbb{T}(\pi) + \nu(\pi), \quad \pi \in \Pi, \tag{3.6}$$

where $\mathbb{T}(\pi)$ *is a mean zero Gaussian process on* $\ell^\infty(\Pi)$ *with covariance kernel* $E(\mathbb{T}(\pi_1)\mathbb{T}(\pi_2)) = \Sigma(\pi_1, \pi_2)$ *for any* $\pi_1, \pi_2 \in \Pi$, *and* $\nu(\pi)$ *is defined in* (3.5). *In addition, if Assumption 3.4 is*

satisfied, then

$$T_n^*(\pi, \theta_0) \Rightarrow \mathbb{T}(\pi), \quad \pi \in \Pi. \tag{3.7}$$

Theorem 3.1 provides weak convergence of the t-statistic process $T_n^*(\pi, \theta_0), \pi \in \Pi$. This is an asymptotic theory for the entire sequence of t-statistics $T_n(K, \theta_0), K \in \mathcal{K}_n$. The t-statistic process converges weakly to a mean zero Gaussian process $\mathbb{T}(\pi)$ plus the asymptotic bias $\nu(\pi)$. Proof of Theorem 3.1 needs to verify a uniform-entropy condition and apply empirical process theory in van der Vaart and Wellner (1996, Theorem 2.11.22).

Remark 3.1 (Rate conditions)**.** Note that the asymptotic bias $|\nu(\pi)| = 0$ if $\bar{K}$ increases faster than the optimal MSE rate (undersmoothing). $0 < |\nu(\pi)| < \infty$ if $\bar{K}$ increases at the optimal MSE rate, and $|\nu(\pi)| = +\infty$ if $\bar{K}$ increases slower than the optimal MSE rate (oversmoothing). Theorem 3.1 does not allow oversmoothing rates as we require $\sup_\pi |\nu(\pi)| < \infty$. Assumption 3.1 does not consider different rates of K satisfying asymptotic normality of series estimators, however, these are the class of sequences to be able to provide uniform central limit theorem of the t-statistic process. As studentized t-statistic is normalized by the standard deviation $V_K^{1/2}$ which may increase differently with different rates of K, two t-statistics with different rates of K can be asymptotically independent, thus hard to incorporate dependency (see discussions in Section 3.2 with alternative $\mathcal{K}_n$ allowing different rates of K).

Remark 3.2 (Other functionals)**.** Here, I focus on the leading example, where $\theta_0 = g_0(x)$ for some fixed point $x \in \mathcal{X}$, but I may consider other linear functionals $\theta_0 = a(g_0(\cdot))$, such as the regression derivates $a(g_0(x)) = \frac{d}{dx} g_0(x)$. All the results in this paper can be applied to irregular (slower than $n^{1/2}$ rate) linear functionals with estimators $\widehat{\theta} = a(\widehat{g}_K(x)) = a_K(x)'\widehat{\beta}_K$ and appropriate transformation of basis $a_K(x) = (a(p_1(x), \cdots, a(p_K(x)))'$. While verification of previous results for regular ($n^{1/2}$ rate) functionals, such as integrals and weighted average derivative, is beyond the scope of this paper, I examine similar results for the partially linear model setup in Section 7.

3.2 Asymptotic normality with different rates

Next, we provide different asymptotic distribution of the sequence of t-statistics with an alternative set $\mathcal{K}_n$ constructed to allow different rates of Ks. A following alternative set assumption allows for optimal mean squared error rates of K as well as oversmoothing rates which increase slower than the optimal MSE rates.

Assumption 3.5. *(Alternative set with different rates) Let $\mathcal{K}_n$ as*

$\mathcal{K}_n = \{\underline{K} = K_1, \cdots, K_m, \cdots, \bar{K} = K_M\}$ *where* $K_m \equiv \lfloor \tau n^{\phi_m} \rfloor$ *for constant* $\tau > 0$, $0 < \phi_1 < \phi_2 < \cdots < \phi_M$, *and fixed* M. *Define asymptotic bias for the sequence of models as* $\nu(m) \equiv -\lim_{n\to\infty} \sqrt{n} V_{K_m}^{-1/2} r_{K_m}$. *Assume that the largest model* $\bar{K}$ *satisfies* $\sqrt{n} V_{\bar{K}}^{-1/2} \ell_{\bar{K}} c_{\bar{K}} \to 0$ *as* $n \to \infty$.

Assumption 3.5 can consider rates of K from oversmoothing ($|\nu(m)| = +\infty$) to undersmoothing ($|\nu(m)| = 0$) with different ϕ_m. Here, $\underline{K}$ can increase slower than the optimal MSE rates and $\bar{K}$ satisfies undersmoothing rates. Assumption 3.5 considers a broader range of Ks than the Assumption 3.1 as alternative set allows slower than optimal MSE rates. Together with Assumption 3.2, there exist explicit rate restrictions on ϕ_m uniformly over m to guarantee an asymptotic normality of each single t-statistic. Undersmoothing assumption for the $\bar{K}$, i.e. $\nu(M) = 0$, is a modeling device considering a broad range of K and large enough $\bar{K}$ so that satisfy undersmoothing.

Under Assumption 3.5, joint t-statistics do not converge in distribution to a bounded random vector if any of the elements $|\nu(m)| = +\infty$ with oversmoothing sequences. If $\nu(m) = \pm\infty$ for some m, then it can be shown that corresponding t-statistic $T_n(K_m, \theta_0)$ diverges in probability to $\pm\infty$. This matters when we obtain the asymptotic distribution of the test statistic that is some continuous transformation of the joint t-statistics because continuous mapping theorem cannot be directly applied. This technical difficulty does not arise when we consider t-statistics centered at the true conditional mean function (θ_0) with relative bias (r_{K_m}) and standard error ($\sqrt{V_{K_m}/n}$): joint t-statistics converge in distribution to a mean zero normal random vector. However, allowing $|\nu(m)| = +\infty$ is important for our analysis to consider asymptotic bias and undersmoothing (and oversmoothing) condition with their effects on inference.

To obtain the asymptotic distribution under $\sup_m |\nu(m)| = +\infty$, we provide formal proofs which combine arguments in inference on CIs for the parameters in moment inequality literature as in Andrews and Guggenberger (2009). For this, we define the continuous function on the extended real space as follows; $S : A \to B$ is continuous at $t \in A$ if $t' \to t$ for $t \in A$ implies $S(t') \to S(t)$ for any set A.

Theorem 3.2. *Under Assumptions 3.2 and 3.5, following holds for any continuous function* $S(t)$ *at all* $t \in \mathbb{R}^{M-1}_{[\pm\infty]} \times \mathbb{R}$,

$$S(T_n(\theta_0)) \xrightarrow{d} S(Z + \nu)$$

where $T_n(\theta) = (T_n(K_1, \theta), \cdots, T_n(K_M, \theta))', Z = (Z_1, \cdots, Z_M)' \sim N(0, \Sigma)$, $\nu = (\nu(1), \cdots, \nu(M))'$ *are* $M \times 1$ *vectors provided that* Σ *exists and is a finite positive definite matrix*

with $\Sigma_{jl} = \lim_{n\to\infty} \Sigma_{jl,n}$, *and* $\Sigma_{jl,n} = \frac{P_{K_j}(x)'E(P_{K_j i}P'_{K_l i}\varepsilon_i^2)P_{K_l}(x)}{V_{K_j}^{1/2}V_{K_l}^{1/2}}$. *If* $\nu(m) = \pm\infty$, *then the corresponding element of* $Z + \nu$ *equals* $\pm\infty$.

Note that we do not require either Assumption 3.4 (undersmoothing) or $\sup_m |\nu(m)| < \infty$ in Theorem 3.2. Variance-covariance matrix Σ is similarly defined as in Theorem 3.1. Moreover, if $V_K^{1/2} \asymp K^\eta$ at some point x with $\eta > 0$, then for any $j < l$, $\Sigma_{jl,n} \leq C\frac{V_{K_j}^{1/2}}{V_{K_l}^{1/2}}$ for some constant $C > 0$ by Assumption 3.2(ii) and the upper bound converges to 0 as $n \to 0$ by Assumption 3.5, thus $\Sigma_{jl,n} \to 0$ and $\Sigma = I_M$.

Remark 3.3 (Rate conditions (continued))**.** Note that Assumption 3.5 only considers finite sequences, i.e., $|\mathcal{K}_n| = M$. Assumption 3.5 is useful to consider the effect of bias on inference problems allowing a broader range of K including oversmoothing rates (see Section 4 for formal results). On the other hand, Assumption 3.1 considers sequences of K with the same rates which only differ in constant π. Thus, Theorem 3.1 gives the joint asymptotic distribution of t-statistics that has either zero bias for all $K \in \mathcal{K}_n$ or non-zero bounded bias for all $K \in \mathcal{K}_n$.

4 Test statistic

In this section, I introduce an infimum test statistic and analyze its asymptotic null distribution based on Theorem 3.1 and 3.2. Then, I provide an asymptotic size result of the tests, and methods to obtain critical values for inference procedures.

I consider following test statistic

$$\operatorname{Inf} T_n(\theta) \equiv \inf_{K\in\mathcal{K}_n} |T_n(\lfloor K \rfloor, \theta)|, \tag{4.1}$$

where either Assumption 3.1 or Assumption 3.5 can be used for $\mathcal{K}_n$. Note that $\operatorname{Inf} T_n(\theta) = \inf_{\pi\in\Pi} |T_n^*(\pi, \theta)|$ under Assumption 3.1 and $\operatorname{Inf} T_n(\theta) = \inf_{m=1,\cdots,M} |T_n(K_m, \theta)|$ under Assumption 3.5.

As I denoted in the introduction, there are several reasons to consider $\operatorname{Inf} T_n(\theta)$ in the series regression context. First of all, small t-statistic centered at the true θ_0 corresponds to the approximation with a certain choice of series terms that has a small bias and large variance, which is good for inference similar to what undersmoothing assumption does by eliminating asymptotic bias, theoretically. Moreover, this is also closely related to some rule-of-thumb methods suggested by several papers to choose undersmoothed K (see, for example, Newey (2013), Newey, Powell and Vella (1999)).

4.1 Asymptotic distribution of the test statistic

Asymptotic null limiting distribution of the infimum test statistic follows immediately from Theorem 3.1 and 3.2.

Corollary 4.1. *1. Under Assumptions 3.1, 3.2, 3.3 and* $\sup_\pi |\nu(\pi)| < \infty$, $\operatorname{Inf} T_n(\theta_0) \xrightarrow{d} \inf_{\pi\in[\underline{\pi},1]} |\mathbb{T}(\pi) + \nu(\pi)|$, *where* $\mathbb{T}(\pi)$ *is the mean zero Gaussian process defined in Theorem 3.1. In addition, if Assumption 3.4 holds, then* $\operatorname{Inf} T_n(\theta_0) \xrightarrow{d} \xi_{\inf} \equiv \inf_{\pi\in[\underline{\pi},1]} |\mathbb{T}(\pi)|$.

2. Under Assumptions 3.2 and 3.5, $\operatorname{Inf} T_n(\theta_0) \xrightarrow{d} \inf_{m=1,\cdots,M} |Z_m + \nu(m)|$, *where* Z_m *is an element of* $M \times 1$ *normal vector* $Z \sim N(0, \Sigma)$ *and* $\nu = (\nu(1), \cdots, \nu(M))'$ *is defined in Theorem 3.2.*

Corollary 4.1.1 derives the asymptotic null limiting distribution of $\operatorname{Inf} T_n(\theta)$ under $\mathcal{K}_n$ with same rates of K (Assumption 3.1) and Corollary 4.1.2 provides the asymptotic distribution under alternative $\mathcal{K}_n$ with different rates of K (Assumption 3.5).

Whether some asymptotic bias $|\nu(m)|$ are unbounded or not, Corollary 4.1.2 shows that $\operatorname{Inf} T_n(\theta_0)$ converge in distribution to the bounded random variable. Under the null H_0, $\operatorname{Inf} T_n(\theta)$ exclude all small Ks corresponding to oversmoothing rates (where the bias is of larger order than the standard error) and select among large Ks with optimal MSE rates and undersmoothing rates (where the bias is of smaller order), asymptotically. Using this Corollary, I discuss the effect of asymptotic bias on the inference in Section 4.2 (for the size results) and Section 5 (for the coverage results).

4.2 Asymptotic size

I start by defining critical value $c^{\inf}_{1-\alpha}$ as $(1-\alpha)$ quantile of the asymptotic null distribution $\xi_{\inf} = \inf_{\pi\in[\underline{\pi},1]} |\mathbb{T}(\pi)|$ in Corollary 4.1.1, i.e., solves

$$P(\inf_{\pi\in[\underline{\pi},1]} |\mathbb{T}(\pi)| > c^{\inf}_{1-\alpha}) = \alpha \tag{4.2}$$

for $0 < \alpha < 1/2$.[2]

The asymptotic null distribution, $\xi_{\inf}$, can be completely defined by covariance kernel of the limiting Gaussian process $\mathbb{T}(\pi)$ in Theorem 3.1. In the special case where Assumption 3.3(ii) holds with $\eta = 1$ under homoskedasticity (as discussed below the equation (3.4)),

[2]Without imposing the undersmoothing assumption, asymptotic distribution of $\operatorname{Inf} T_n(\theta_0)$ in Corollary 4.1.1 also depend on asymptotic bias $\nu(\pi)$ as well. If $\nu(\pi)$ can be replaced by some estimates $\widehat{\nu}(\pi)$, then the critical value from $\inf_{\pi\in\Pi} |\mathbb{T}(\pi) + \widehat{\nu}(\pi)|$ can be used. We do not pursue this approach as it is a difficult problem beyond the scope of this paper.

$\inf_{\pi\in[\underline{\pi},1]}|\mathbb{T}(\pi)|$ can be approximated by $\inf_{\pi\in[\underline{\pi},1]}|\mathbb{Z}(\pi)/\sqrt{\pi}|$ with a Brownian motion $\mathbb{Z}(\pi)$. Then the critical value can be tabulated as a function of $\underline{\pi}=\underline{K}/\bar{K}$ with the smallest $\underline{K}$ and the largest $\bar{K}$, which can be viewed as an analogous result in Armstrong and Kolesár (2015) with the uniform Kernel (See Section 2 of Armstrong and Kolesár (2015)). In general, the limiting Gaussian process can not be written as some transformation of Brownian motion, so that the asymptotic critical value cannot be tabulated. However, critical values can be obtained by standard Monte Carlo simulation method or by the weighted bootstrap method and will be discussed in Section 4.3.

With abuse of notation, I also use $c^{\inf}_{1-\alpha}$ as $(1-\alpha)$ quantile of the $\inf_{m=1,\cdots,M}|Z_m|$ if Corollary 4.1.2 is used under Assumption 3.5. Next, I define $z_{1-\alpha/2}$ as $(1-\alpha/2)$ quantile of the standard normal distribution function, which solves $P(|Z|>z_{1-\alpha/2})=\alpha$, where $Z\sim N(0,1)$. Next Corollary provides the asymptotic size of the tests based on $\operatorname{Inf}T_n(\theta)$ follows from the Corollary 4.1.

Corollary 4.2. *1. Under Assumptions 3.1, 3.2, 3.3 and 3.4, following holds with critical value $c^{\inf}_{1-\alpha}$ defined in* (4.2) *and the normal critical value $z_{1-\alpha/2}$,*

$$\limsup_{n\to\infty}P(\operatorname{Inf}T_n(\theta_0)>c^{\inf}_{1-\alpha})=\alpha,\qquad \limsup_{n\to\infty}P(\operatorname{Inf}T_n(\theta_0)>z_{1-\alpha/2})\leq\alpha. \tag{4.3}$$

2. Suppose Assumptions 3.1, 3.2, and 3.3 hold. If $\sup_\pi|\nu_\pi|<\infty$, then following inequality holds

$$\limsup_{n\to\infty}P(\operatorname{Inf}T_n(\theta_0)>c^{\inf}_{1-\alpha})\leq F(c^{\inf}_{1-\alpha},\inf_\pi|\nu(\pi)|), \tag{4.4}$$

$$\limsup_{n\to\infty}P(\operatorname{Inf}T_n(\theta_0)>z_{1-\alpha/2})\leq F(z_{1-\alpha/2},\inf_\pi|\nu(\pi)|), \tag{4.5}$$

where $F(c,|\nu|)=1-\Phi(c-|\nu|)+\Phi(-c-|\nu|)$ with the standard normal cumulative distribution function $\Phi(\cdot)$.

3. Under Assumptions 3.2 and 3.5, following holds

$$\limsup_{n\to\infty}P(\operatorname{Inf}T_n(\theta_0)>c^{\inf}_{1-\alpha})\leq F(c^{\inf}_{1-\alpha},0), \tag{4.6}$$

$$\limsup_{n\to\infty}P(\operatorname{Inf}T_n(\theta_0)>z_{1-\alpha/2})\leq\alpha. \tag{4.7}$$

Under Assumption 3.1 (same rates of K), Corollary 4.2.1 shows that the tests based on $\operatorname{Inf}T_n(\theta)$ asymptotically control size assuming all $K\in\mathcal{K}_n$ satisfy the undersmoothing condition. As $\operatorname{Inf}T_n(\theta_0)\leq|T_n(K,\theta_0)|$ and $|T_n(K,\theta_0)|\xrightarrow{d}|N(0,1)|$ for any single $K\in$

$\mathcal{K}_n$, the test based on $\operatorname{Inf} T_n(\theta)$ using normal critical value also controls the asymptotic size, but conservative. Without undersmoothing assumption, Corollary 4.2.2 shows that the asymptotic size is bounded above by the asymptotic size of a single t-statistic with the smallest asymptotic bias $\inf_\pi |\nu(\pi)|$. Note that $F(c, |\nu|)$ is a monotone decreasing in c and increasing in $|\nu|$. See also Hall and Horowitz (2013), Hansen (2014) for the similar function and Figure 1 for the plots of $F(\cdot, \cdot)$ as a function of $|\nu|$ with some different c.

Note that $c^{\inf}_{1-\alpha} \leq z_{1-\alpha/2}$, so that $F(z_{1-\alpha/2}, 0) = \alpha \leq F(c^{\inf}_{1-\alpha}, 0)$. Moreover, the upper bounds of the asymptotic size can be small if the smallest bias $\inf_\pi |\nu(\pi)|$ is small. For example, when $c^{\inf}_{1-\alpha} = 1.5$, $F(c^{\inf}_{1-\alpha}, \inf_\pi |\nu(\pi)|) = 0.13$ for $\inf_\pi |\nu(\pi)| = 0$. (4.5) also shows that the test based on $\operatorname{Inf} T_n(\theta_0)$ with normal critical value controls size asymptotically if the smallest bias is 0.

Under Assumption 3.5 (different rates of K), Corollary 4.2.3 shows that asymptotic size of the tests based on $\operatorname{Inf} T_n(\theta)$ is bounded above even when we allowing 'large' asymptotic bias ($|\nu(m)| = \infty$) for several Ks in $\mathcal{K}_n$.[3]

Remark 4.1 (Largest K)**.** Although, there exist rate restrictions for $\bar{K}$ by Assumption 3.2 to be used for the asymptotic normal approximation, formal guidance or data-dependent results for the range of $\mathcal{K}_n = [\underline{K}, \bar{K}]$ are beyond the scope of this paper. Nevertheless, the test based on $\operatorname{Inf} T_n(\theta_0)$ and its asymptotic critical value $c^{\inf}_{1-\alpha}$ may have better power compare with the test based on $T_n(K, \theta_0)$ with the normal critical value for some large K. Suppose that $\operatorname{Inf} T_n(\theta_0) = |T_n(K, \theta_0)|$ for some large K (say $\bar{K}$) under the null, then the test based on $\operatorname{Inf} T_n(\theta_0)$ may have better power as this test compares with the smaller critical value than the normal critical value.

Also, note that the asymptotic size result in (4.7) relies on the inequality $\operatorname{Inf} T_n(\theta_0) \leq |T_n(\bar{K}, \theta_0)|$ and the fact that $T_n(\bar{K}, \theta_0) \xrightarrow{d} N(0, 1)$ under Assumption 3.5. But, theory can provide the bound of the asymptotic size in (4.7) as $F(z_{1-\alpha/2}, \inf_m |\nu(m)|)$ without any undersmoothing conditions on $K \in \mathcal{K}_n$. Asymptotic distribution result in Corollary 4.1.2 is still valid as long as $\inf_m |\nu(m)|$ is bounded. If we know (a priori) that $\bar{K}$ satisfies the undersmoothing condition and others not, then there's no point of searching over different

[3]If we further assume $V_K^{1/2} \asymp K^\eta, \eta > 0$ for all $K \in \mathcal{K}_n$ then $\Sigma = I_M$ and t-statistics are asymptotically independent (see discussions below Theorem 3.2). Then, we can get asymptotic size of the test as $\prod_{m=1}^M F(c^{\inf}_{1-\alpha}, |\nu(m)|)$. The asymptotic size is not affected by K_m such that $|\nu(m)| = \infty$ since $F(c, \infty) = 1$ for any constant $c > 0$. Further, suppose that the last M_1 number of Ks satisfy undersmoothing conditions and the others satisfy oversmoothing rates, i.e., $|\nu(m)| = \infty$ for $m = 1, \cdots, M - M_1$ and $|\nu(m)| = 0$ for the others. Then, the asymptotic size is equal to $\alpha^{M_1/M}$, as $c^{\inf}_{1-\alpha} = z_{1-\alpha^{1/M}/2}$ follows from Theorem 3.2 and $\Sigma = I_M$. In this special case, the asymptotic size is a decreasing function of the fraction of number of undersmoothing sequences M_1/M, and is equal to α when $|\nu(m)| = 0$ for all m, similar to Corollary 4.2.1. Using infimum t-statistic and larger critical value $z_{1-\alpha^{1/M_1}/2}$ (which is equal to the standard normal critical value when $M_1 = 1$) controls asymptotic size in this particular case, but this is not practically useful.

K; we may just use $\bar{K}$ for the inference. This may work well if $\bar{K}$ coincides with some size-optimal sequence $K^*(n) = \arg\min_K |P(T_n(K, \theta_0) > z_{1-\alpha/2}) - \alpha|$. However $K^*(n)$ is unknown, so the choice of $\bar{K}$ can be ad hoc in practice.

Remark 4.2 (Power of the test)**.** Although $\text{Inf}\, T_n(\theta)$ leads to the tests that control the asymptotic size or bound the size distortions, one reasonable concern is that possible low power property of the test compare with the other statistics (e.g., the supremum of the t-statistics). Investigating local asymptotic power comparisons of the level α test based on several different statistics, or some optimal property of the tests in this nonparametric regression context is important, but these are beyond the scope of this paper. I discuss the length of CIs based on inverting an infimum test statistic in Section 5. I also report the length of proposed CIs (Figures 5-6) and power of the tests (Figure 7) in Section 8 with various simulation setups.

The goal of this paper is to develop tests which can control size distortions even allowing large asymptotic bias for several different series approximations. I want to emphasize that bias issues can severely affect commonly used inference procedures (i.e., coverage of standard CI) in series estimation. For example, high-order polynomials can be highly sensitive to the choice of series terms. Using low-order polynomials or regression splines can help to reduce bias issues, but does not solve bias problem completely. Moreover, a test based on the other transformation of the t-statistics can be sensitive to the bias problems, thus may lead to size distortions of the tests. For example, see Appendix C for the inference based on the supremum test statistic under Assumption 3.5.

4.3 Critical values

In this section, I discuss detail descriptions to approximate critical values defined in (4.2). Here, I suggest using simple simulation methods to obtain critical values. To make implementation procedures simple, I impose following set assumption and conditional homoskedasticity.

Assumption 4.1. *(Set of finite number of series terms)*

$\mathcal{K}_n = \{\underline{K} \equiv K_1, \cdots, K_m, \cdots, \bar{K} \equiv K_M\}$ *where* $K_m = \pi_m \bar{K}$ *for constant* π_m*,* $0 < \underline{\pi} = \pi_1 < \pi_2 < \cdots < \pi_M = 1$*, fixed* M*, and* $\bar{K} = \bar{K}(n) \to \infty$ *as* $n \to \infty$*.*

Assumption 4.2. *(Conditional homoskedasticity)* $E(\varepsilon_i^2 | x_i = x) = \sigma^2$*.*

Assumption 4.1 is a finite dimensional version of Assumption 3.1 and is different with an alternative set (Assumption 3.5) that considers a different rate of Ks. Hereafter, without loss

of generality, we assume $K_1 < K_2 < \cdots < K_M$ and they are all integers. In finite samples, we only consider finite set $\mathcal{K}_n$, so the difference between Assumption 4.1 and 3.5 only matters in large samples. Conditional homoskedasticity assumption is only for a simpler implementation. Based on the general covariance structure defined in Theorem 3.1 and 3.2, we can construct a variance-covariance matrix using its sample analogs under the heteroskedastic error.

By Theorem 3.1, following finite dimensional convergence of the t-statistics holds under the Assumptions 3.2, 3.4, 4.1, and 4.2

$$(T_n(K_1, \theta_0), \cdots, T_n(K_M, \theta_0))' \xrightarrow{d} Z = (Z_1, \cdots, Z_M)', \quad Z \sim N(0, \Sigma), \tag{4.8}$$

where Σ is a variance-covariance matrix, $\Sigma_{jl} = \lim_{n\to\infty} V_{K_j}^{1/2}/V_{K_l}^{1/2}$ for any j and l, provided that Σ exists and is a finite positive definite matrix. (4.8) also holds under same assumptions as in Theorem 3.2. Note that the limiting distribution does not depend on θ_0 and variance-covariance matrix Σ can be consistently estimated by its sample counterparts. This requires estimators of the variance V_K that are consistent uniformly over $K \in \mathcal{K}_n$. Define least square residuals as $\widehat{\varepsilon}_{Ki} = y_i - P_{Ki}'\widehat{\beta}_K$, and let $\widehat{V}_K$ as the simple pluig-in estimator for V_K

$$\begin{aligned} &\widehat{V}_K = P_K(x)'\widehat{Q}_K^{-1}\widehat{\Omega}_K\widehat{Q}_K^{-1}P_K(x), \\ &\widehat{Q}_K = \frac{1}{n}\sum_{i=1}^n P_{Ki}P_{Ki}', \quad \widehat{\Omega}_K = \frac{1}{n}\sum_{i=1}^n P_{Ki}P_{Ki}'\widehat{\varepsilon}_{Ki}^2. \end{aligned} \tag{4.9}$$

Then, I define $\widehat{c}_{1-\alpha}^{\inf}$ based on the asymptotic null distribution of $\operatorname{Inf} T_n(\theta_0)$ as follows,

$$\begin{aligned} &\widehat{c}_{1-\alpha}^{\inf} \equiv (1-\alpha) \text{ quantile of } \inf_{m=1,\cdots,M} |Z_{m,\widehat{\Sigma}}|, \\ &\text{where } Z_{\widehat{\Sigma}} = (Z_{1,\widehat{\Sigma}}, \cdots, Z_{M,\widehat{\Sigma}})' \sim N(0, \widehat{\Sigma}), \quad \widehat{\Sigma}_{jj} = 1, \widehat{\Sigma}_{jl} = \widehat{V}_{K_j}^{1/2}/\widehat{V}_{K_l}^{1/2}. \end{aligned} \tag{4.10}$$

One can compute $\widehat{c}_{1-\alpha}^{\inf}$ by simulating B (typically $B = 1000$ or 5000) i.i.d. random vectors $Z_{\widehat{\Sigma}}^b \sim N(0, \widehat{\Sigma})$ and by taking $(1-\alpha)$ sample quantile of $\{\operatorname{Inf} T_n^b = \inf_m |Z_{m,\widehat{\Sigma}}^b| : b = 1, \cdots, B\}$.[4]

I impose following assumption on the consistency of variance estimator $\widehat{V}_K$ uniformly in $K \in \mathcal{K}_n$.

Assumption 4.3. $\sup_{K\in\mathcal{K}_n} |\frac{\widehat{V}_K}{V_K} - 1| = o_p(1)$ *as* $n, K \to \infty$.

[4]Under heteroskedastic error terms, we can construct $\widehat{\Sigma}_{j,l} = \frac{\widehat{V}_{K_{jl}}}{\widehat{V}_{K_j}^{1/2}\widehat{V}_{K_l}^{1/2}}$ for any $j < l$, where $\widehat{V}_{K_{jl}}$ is an sample analog estimator of $P_{K_j}(x)'E(P_{K_ji}P_{K_li}'\varepsilon_i^2)P_{K_l}(x)$ and $\widehat{V}_{K_j}, \widehat{V}_{K_l}$ are estimator of the variance V_{K_j}, V_{K_l}, respectively.

Assumption 4.3 is satisfied under same regularity conditions (Assumption 3.1 and 3.2) with an additional assumption. For example, if we further assume $\sup_{K\in\mathcal{K}_n} || \sum_{i=1}^n \tilde{P}_{Ki}\tilde{P}_{Ki}'\varepsilon_i^2 - E[\tilde{P}_{Ki}\tilde{P}_{Ki}'\varepsilon_i^2]|| = o_p(1)$ with an orthonormalized vector of basis functions $\tilde{P}_K(x) \equiv Q_K^{-1/2}P_K(x)$, then Assumption 4.3 holds. See Lemma 5.1 of Belloni et al. (2015), and also Lemma 3.1 and 3.2 of Chen and Christensen (2015b) for different sufficient conditions under mild rate restrictions and unconditional moment of the error terms.

Next, we consider a t-statistic $T_{n,\widehat{V}}(K,\theta) = \sqrt{\frac{n}{\widehat{V}_K}}(\widehat{\theta}_K - \theta_0)$ replacing variance of the series estimator V_K with $\widehat{V}_K$. Following Corollary provides the joint asymptotic distribution of $T_{n,\widehat{V}}(K,\theta)$ for $K \in \mathcal{K}_n$ and the validity of Monte Carlo critical values $\widehat{c}_{1-\alpha}^{\inf}$ defined in (4.10).

Corollary 4.3. *Under Assumptions 3.2, 3.4, 4.1, 4.2 and 4.3,*

$$(T_{n,\widehat{V}}(K_1,\theta_0),\cdots,T_{n,\widehat{V}}(K_M,\theta_0))' \xrightarrow{d} Z$$

where $Z = (Z_1,\cdots,Z_M)' \sim N(0,\Sigma)$ with a positive definite matrix Σ defined in (4.8)*. This also holds under the Assumptions 3.2, 3.4, 3.5, 4.2, and 4.3. Furthermore, $\widehat{c}_{1-\alpha}^{\inf} \xrightarrow{p} c_{1-\alpha}^{\inf}$ holds where $\widehat{c}_{1-\alpha}^{\inf}$ is defined in* (4.10) *and $c_{1-\alpha}^{\inf}$ is the $(1-\alpha)$ quantile of $\inf_{m=1,\cdots,M}|Z_m|$.*

Remark 4.3 (Weighted bootstrap)**.** Alternatively, we can use the weighted bootstrap method to approximate asymptotic critical values. Implementation of the weighted bootstrap method is as follows. First, generate i.i.d draws from exponential random variables $\{\omega_i\}_{i=1}^n$, independent of the data. Then, for each draw, calculate LS estimator weighted by $\omega_1,\cdots,\omega_n$ for each $K \in \mathcal{K}_n$ and construct weighted bootstrap t-statistic as follows

$$\begin{aligned}
&\hat{\beta}_K^b = \arg\min_b \frac{1}{n}\sum_{i=1}^n \omega_i(y_i - P_{Ki}'b)^2, \quad \hat{g}_K^b(x) = P_K(x)'\hat{\beta}_K^b, \\
&T_n^b(K) = \frac{\sqrt{n}(\hat{g}_K^b(x) - \hat{g}_K(x))}{\widehat{V}_K^{1/2}}.
\end{aligned} \tag{4.11}$$

Then, construct $\text{Inf}\, T_n^b = \inf_K |T_n^b(K)|$. Repeat this B times (1000 or 5000) and define $\widehat{c}_{1-\alpha}^{\inf,WB}$ as conditional $1-\alpha$ quantile of $\{\text{Inf}\, T_n^b : b = 1,\cdots,B\}$ given the data. Similar to Belloni et al. (2015), the idea behind the weighted bootstrap methods is as follows: if the limiting distribution of weighted bootstrap process is equal to the original process conditional on the data, then the weighted bootstrap process $\text{Inf}\, T_n^b$ also approximate the original limiting distribution $\inf_{\pi\in[\underline{\pi},1]}\mathbb{T}(\pi)$. However, the validity of the weighted bootstrap is beyond the scope of this paper and will be pursued for the future work.

5 Confidence interval

Now, I introduce confidence intervals for $\theta_0 = g_0(x)$ and provide their coverage properties. We consider a confidence interval based on inverting a test statistic for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Define $CI^{\texttt{Robust}}_{\texttt{inf}}$ as the nominal level $1-\alpha$ CI for θ based on infimum test statistics,

$$\begin{aligned} CI^{\texttt{Robust}}_{\texttt{inf}} &\equiv \{\theta : \inf_{K \in \mathcal{K}_n} |T_{n,\widehat{V}}(K,\theta)| \leq \widehat{c}^{\inf}_{1-\alpha}\} \\ &= \{\theta : |T_{n,\widehat{V}}(K,\theta)| > \widehat{c}^{\inf}_{1-\alpha}, \forall K\}^C = \bigcup_{K \in \mathcal{K}_n} \{\theta : |T_{n,\widehat{V}}(K,\theta)| \leq \widehat{c}^{\inf}_{1-\alpha}\} \\ &= [\inf_K(\widehat{\theta}_K - \widehat{c}^{\inf}_{1-\alpha} s(\widehat{\theta}_K)), \sup_K(\widehat{\theta}_K + \widehat{c}^{\inf}_{1-\alpha} s(\widehat{\theta}_K))] \end{aligned} \tag{5.1}$$

where $\widehat{c}^{\inf}_{1-\alpha}$ is the Monte Carlo critical value defined in Section 4.3, $s(\widehat{\theta}_K) \equiv \sqrt{\widehat{V}_K/n}$ is a standard error of series estimator $\widehat{\theta}_K$ using K series terms, and A^C denotes the complement of a set A. Note that $CI^{\texttt{Robust}}_{\texttt{inf}}$ can be easily obtained by using estimates $\widehat{\theta}_K$, standard errors $s(\widehat{\theta}_K)$, and a critical value $\widehat{c}^{\inf}_{1-\alpha}$. $CI^{\texttt{Robust}}_{\texttt{inf}}$ is the lower and the upper-end point of confidence intervals for all $K \in \mathcal{K}_n$ using critical value $\widehat{c}^{\inf}_{1-\alpha}$.

Similarly, I define $CI_{\texttt{inf}}$ based on $\operatorname{Inf} T_n(\theta)$ and the normal critical value $z_{1-\alpha/2}$,

$$\begin{aligned} CI_{\texttt{inf}} &\equiv \{\theta : \inf_{K \in \mathcal{K}_n} |T_{n,\widehat{V}}(K,\theta)| \leq z_{1-\alpha/2}\} \\ &= [\inf_K(\widehat{\theta}_K - z_{1-\alpha/2} s(\widehat{\theta}_K)), \sup_K(\widehat{\theta}_K + z_{1-\alpha/2} s(\widehat{\theta}_K))] \end{aligned} \tag{5.2}$$

Note that $CI_{\texttt{inf}}$ is the union of all confidence intervals for $K \in \mathcal{K}_n$ using conventional normal critical value $z_{1-\alpha/2}$.

Next Corollary shows valid coverage property of the above CIs, and it follows from Corollary 4.2 and 4.3.

Corollary 5.1. *1. Under Assumptions 3.2, 3.4, 4.1, 4.2, and 4.3,*

$$\liminf_{n\to\infty} P(\theta_0 \in CI^{Robust}_{inf}) = 1-\alpha, \qquad \liminf_{n\to\infty} P(\theta_0 \in CI_{inf}) \geq 1-\alpha. \tag{5.3}$$

2. Under Assumptions 3.2, 4.1, 4.2, 4.3, and $\sup_m |\nu(m)| < \infty$*,*

$$\liminf_{n\to\infty} P(\theta_0 \in CI^{Robust}_{inf}) \geq 1 - F(c^{\inf}_{1-\alpha}, \inf_m |\nu(m)|), \tag{5.4}$$

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{inf}) \geq 1 - F(z_{1-\alpha/2}, \inf_m |\nu(m)|). \tag{5.5}$$

3. Under Assumptions 3.2, 3.5, 4.2, and 4.3,

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{inf}^{Robust}) \geq 1 - F(c_{1-\alpha}^{\inf}, 0), \tag{5.6}$$

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{inf}) \geq 1 - \alpha. \tag{5.7}$$

Corollary 5.1.1 shows the validity of $CI_{\texttt{inf}}^{\texttt{Robust}}$, i.e., asymptotic coverage is equal to $1-\alpha$. The asymptotic coverage of $CI_{\texttt{inf}}$ is greater or equal than $1-\alpha$. Note that the Corollary 5.1.1 requires undersmoothing condition, i.e., no asymptotic bias for all Ks in $\mathcal{K}_n$.

Without undersmoothing assumption, Corollary 5.1.2 and 5.1.3 show that the coverage probability of $CI_{\texttt{inf}}^{\texttt{Robust}}$ and $CI_{\texttt{inf}}$ are bounded below by the coverage of single K with the smallest asymptotic bias, similar to the asymptotic size results in Corollary 4.2. For example, the lower bound in (5.6) is 0.87 when $c_{1-\alpha}^{\inf} = 1.5$. Furthermore, (5.7) shows that $CI_{\texttt{inf}}$ using standard normal critical value achieve nominal coverage probability $1-\alpha$. $CI_{\texttt{inf}}$ and $CI_{\texttt{inf}}^{\texttt{Robust}}$ bound coverage distortions even when we allow large asymptotic bias terms ($|\nu(m)| = \infty$) for several Ks in $\mathcal{K}_n$. In this sense, $CI_{\texttt{inf}}$ and $CI_{\texttt{inf}}^{\texttt{Robust}}$ are robust to the bias problems.

Although $CI_{\texttt{inf}}$ gives formally valid coverage allowing asymptotic bias, coverage property of the $CI_{\texttt{inf}}$ in (5.3) and (5.7) holds with inequality; thus it can be conservative. As the variance of series estimator increases with K, we expect $CI_{\texttt{inf}}$ can be comparable to the standard CI using normal critical values with some large K around $\bar{K}$. In contrast, $CI_{\texttt{inf}}^{\texttt{Robust}}$ has shorter length by using smaller critical value than the normal critical value.

Remark 5.1 (Length of the interval and the ranges of K)**.** Note that potential large length of the $CI_{\texttt{inf}}^{\texttt{Robust}}$ is also related to the possible low power property of the test. Also, note that the last equality from the definition of $CI_{\texttt{inf}}^{\texttt{Robust}}$ in (5.1) holds only when there is no dislocated CI, i.e., the intersection is nonempty at least for some two CIs using $\widehat{c}_{1-\alpha}^{\inf}$. Otherwise, using the superset widens the length of CI. As the variance of series estimator increases with K, we expect that the union of all confidence intervals may only be determined by some large Ks so that there is no dislocated CI. In general, dislocated confidence interval may show some evidence of significant bias for some specific models, but there is no guarantee that the union of the confidence interval is connected in practice.

Although this paper does not consider the data-dependent choice of $\mathcal{K}_n$, a possible large length of CI can be avoidable if $\underline{K}$ is reasonably large and this is exactly the condition needed in Corollary 5.1 to have a correct coverage. Furthermore, the net effect of increasing largest $\bar{K}$ on the length of $CI_{\texttt{inf}}^{\texttt{Robust}}$ is not clear as it may decrease critical values $\widehat{c}_{1-\alpha}^{\inf}$ as well.

6 Post model selection inference

In this section, I provide methods to construct a valid CI that gives correct coverage after selecting the number of series terms considering supremum of the t-statistics.

I first consider the 'post model selection' t-statistic

$$|T_n(\widehat{K},\theta)|, \quad \widehat{K} \in \mathcal{K}_n$$

where $\widehat{K}$ is a possibly data-dependent rule chosen from $\mathcal{K}_n$. Then, we define following 'naive' post-selection CI with $\widehat{K}$ using the normal critical value $z_{1-\alpha/2}$,

$$CI^{\texttt{Naive}}_{\texttt{pms}} \equiv \{\theta : |T_n(\widehat{K},\theta)| \leq z_{1-\alpha/2}\} = [\widehat{\theta}_{\widehat{K}} - z_{1-\alpha/2}s(\widehat{\theta}_{\widehat{K}}), \widehat{\theta}_{\widehat{K}} + z_{1-\alpha/2}s(\widehat{\theta}_{\widehat{K}})]. \tag{6.1}$$

Conventional method of using normal critical values in (6.1) comes from the asymptotic normality of the t-statistic under deterministic sequence, i.e., when $\mathcal{K}_n = \{K\}$. However, it is not clear whether the asymptotic normality of the t-statistic $T_n(\widehat{K},\theta_0) \xrightarrow{d} N(0,1)$ holds with some random sequence of $\widehat{K}$. Even if we assume the asymptotic bias is negligible, the variability of $\widehat{K}$ introduced by some selection rules can affect the variance of the asymptotic distribution. Thus, it is not clear whether naive inference using standard normal critical value is valid. If the post model selection t-statistic, $T_n(\widehat{K},\theta_0)$ with some $\widehat{K}$, has non-normal asymptotic distribution, then the naive confidence interval $CI^{\texttt{Naive}}_{\texttt{pms}}$ may have coverage probability less than the nominal level $1-\alpha$.

Furthermore, $\widehat{K}$ with some data-dependent rules may not satisfy the undersmoothing rate conditions which ensure the asymptotic normality without bias terms. For example, suppose a researcher uses $\widehat{K} = \widehat{K}_{\texttt{cv}}$ selected by cross-validation. It is well known that the $\widehat{K}_{\texttt{cv}}$ is typically too "small" so that lead to a large bias by violating undersmoothing assumption needed to ensure asymptotic normality and the valid inference. If $\widehat{K}$ increases not sufficiently fast as the undersmoothing condition does, then the asymptotic distribution may have bias terms and resulting naive CI may have large coverage distortions.

Here, I construct a valid post-selection CI with $\widehat{K} \in \mathcal{K}_n$ by adjusting standard normal critical value to the critical value from a 'supremum' test statistic,

$$\text{Sup}\, T_n(\theta) \equiv \sup_{K \in \mathcal{K}_n} |T_n(K,\theta)|. \tag{6.2}$$

Note that $|T_n(\widehat{K},\theta_0)| \leq \text{Sup}\, T_n(\theta_0)$ for any choice of $\widehat{K} \in \mathcal{K}_n$, and $\text{Sup}\, T_n(\theta_0) \xrightarrow{} \xi_{\text{sup}} \equiv \sup_{\pi \in [\underline{\pi},1]} |\mathbb{T}(\pi)|$ under the same assumptions as in Corollary 4.1. Therefore, inference based on $|T_n(\widehat{K},\theta_0)|$ using asymptotic critical values from the limiting distribution of $\text{Sup}\, T_n(\theta_0)$

will be valid, but conservative. Similar to $c_{1-\alpha}^{\inf}$ defined in (4.2), I define asymptotic critical value $c_{1-\alpha}^{\sup}$ as $1-\alpha$ quantile of $\xi_{\sup}$. We can approximate this critical value by using Monte Carlo simulation based method similarly as in Section 4.3. To be specific, I define

$$\widehat{c}_{1-\alpha}^{\sup} \equiv (1-\alpha) \text{ quantile of } \sup_{m=1,\cdots,M} |Z_{m,\widehat{\Sigma}}|, \tag{6.3}$$

where $Z_{\widehat{\Sigma}} = (Z_{1,\widehat{\Sigma}}, \cdots, Z_{M,\widehat{\Sigma}})' \sim N(0, \widehat{\Sigma})$ and $\widehat{\Sigma}$ are defined in (4.10). We can verify $\widehat{c}_{1-\alpha}^{\sup} \xrightarrow{p} c_{1-\alpha}^{\sup}$ similar to Corollary 4.3.

Next, I define the following robust post-selection CI using the critical value $\widehat{c}_{1-\alpha}^{\sup}$ rather than the normal critical value $z_{1-\alpha/2}$ compare to $CI_{\texttt{pms}}^{\texttt{Naive}}$,

$$CI_{\texttt{pms}}^{\texttt{Robust}} \equiv [\widehat{\theta}_{\widehat{K}} - \widehat{c}_{1-\alpha}^{\sup} s(\widehat{\theta}_{\widehat{K}}), \widehat{\theta}_{\widehat{K}} + \widehat{c}_{1-\alpha}^{\sup} s(\widehat{\theta}_{\widehat{K}})], \quad \widehat{K} \in \mathcal{K}_n. \tag{6.4}$$

Next Corollary shows that the robust post-selection $CI_{\texttt{pms}}^{\texttt{Robust}}$ guarantees the asymptotic coverage as $1-\alpha$. Even though Corollary 6.1 does not implicitly use randomness of the specific data-dependent selection rules of $\widehat{K}$, $CI_{\texttt{pms}}^{\texttt{Robust}}$ can be useful as it can be applied to any selection rules among $\mathcal{K}_n$.

Corollary 6.1. *Under Assumptions 3.2, 3.4, 4.1, 4.2, and 4.3,*

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{\textit{pms}}^{\textit{Robust}}) \geq 1-\alpha. \tag{6.5}$$

(6.5) *also holds under Assumptions 3.2, 3.4, 3.5, 4.2 and 4.3.*

Corollary 6.1 imposes an undersmoothing (Assumption 3.4) and does not allow optimal MSE rates (e.g., $\widehat{K}_{\texttt{cv}}$). Thus $CI_{\texttt{pms}}^{\texttt{Robust}}$ does not deal with the bias problem explicitly. However, it accommodates bias by enlarging confidence interval using larger critical values $\widehat{c}_{1-\alpha}^{\sup}$ than the normal critical value. Moreover, we also expect $\widehat{c}_{1-\alpha}^{\sup}$ is smaller than the usual Bonferroni-type critical value. Bonferroni corrections use normal critical value $z_{1-\frac{\alpha}{2M}}$ replacing α with α/M. However, Bonferroni critical value can be too large especially when $|\mathcal{K}_n| = M$ is large, as it ignores dependence structure of the t-statistics.

7 Extension: partially linear model setup

In this section, I provide inference methods for the partially linear model (PLM) similar to the nonparametric regression setup.

Suppose we observe random samples $\{y_i, w_i, x_i\}_{i=1}^n$, where y_i is scalar response variable, $w_i \in \mathcal{W} \subset \mathbb{R}$ is treatment/policy variable of interest, and $x_i \in \mathcal{X} \subset \mathbb{R}^{d_x}$ is a set of explanatory

variables. For simplicity, we shall assume w_i is a scalar. I consider following partially linear model

$$y_i = \theta_0 w_i + g_0(x_i) + \varepsilon_i, \qquad E(\varepsilon_i | w_i, x_i) = 0. \tag{7.1}$$

We are interested in inference on treatment/policy effect θ_0 after approximating unknown function $g_0(x)$ by series terms/regressors $p(x_i)$ among a set of potential control variables. A number of regressors can be large if there are many available control variables, i.e., $p(x_i) = x_i$ or if there are large number of transformations of $p(x_i)$ are available such as polynomials and interactions of x_i. For notational simplicity, I use the similar notation as defined in nonparametric regression setup. Suppose we use K regressors $P_{Ki} = P_K(x_i)$, where $P_K(x) = (p_1(x), \cdots, p_K(x))'$ from the basis functions $p(x)$. The approximating model can be written as

$$y_i = \theta_0 w_i + P_{Ki}' \beta_K + \varepsilon_{Ki}, \tag{7.2}$$

where the error term $\varepsilon_{Ki} = r_{Ki} + \varepsilon_i$ and approximation error r_{Ki} are defined similarly as in Section 2. Then, series estimator $\widehat{\theta}_K$ for θ_0 using the first K approximating functions is obtained by standard LS estimation of y_i on w_i and P_{Ki}, and has the usual "partialling out" formula

$$\widehat{\theta}_K = (W' M_K W)^{-1} W' M_K Y \tag{7.3}$$

where $W = (w_1, \cdots, w_n)', M_K = I_K - P^K (P^{K'} P^K)^{-1} P^{K'}, P^K = [P_{K1}, \cdots, P_{Kn}]', Y = (y_1, \cdots, y_n)'$. Similar to the nonparametric regression model, we are interested in testing for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

The asymptotic normality and valid inference for the partially linear model have been developed in the literature. Donald and Newey (1994) derived the asymptotic normality of $\widehat{\theta}_K$ under standard rate conditions where $K/n \to 0$. See also Robinson (1988), Linton (1995) and references therein for the related results of the kernel estimators. Belloni, Chernozukhov, and Hansen (2014) analyzed asymptotic normality and uniformly valid inference for the post-double-selection estimator even when K is much larger than n under some form of sparsity condition. A recent paper by Cattaneo, Jansson, and Newey (2015a) provided a valid approximation theory for $\widehat{\theta}_K$ even when K grows at the same rate of n.

Different approximation theory using faster rate of K ($K/n \to c > 0$) than the standard rate conditions ($K/n \to 0$) is particularly useful for our purpose to better reflect the choice/search of smoothing parameters in finite samples. Unlike the nonparametric object

of interest in fully nonparametric model where variance term increases with K, $\widehat{\theta}_K$ has parametric ($n^{1/2}$) convergence rate and asymptotic variances are same as the semiparametric efficiency bound for all sequences under $K/n \to 0$, i.e., all estimators $\widehat{\theta}_K$ with different rate of Ks satisfying $K/n \to 0$ are asymptotically equivalent. This is also related to the well-known results of the two-step semiparametric estimation; asymptotic variance of two-step semiparametric estimators does not depend on the type of the first-step estimator and smoothing parameter sequences under certain conditions (see Newey (1994b)).

By using the higher order approximation theory that allows the number of series can grow as fast as sample size n, we can construct a joint distribution of the t-statistics with the different sequence of models. Under $K/n \to c$ for $c > 0$, the limiting normal distribution has a larger variance than the standard first-order asymptotic variance derived under $K/n \to 0$. Adjusted variances depend on the number of terms K so that I can provide an approximation theory that accounts the dependency of the t-statistics with different Ks.

Next, I impose regularity conditions that are used in Cattaneo, Jansson, and Newey (2015a, Assumption PLM) uniformly over $K \in \mathcal{K}_n$ where $\mathcal{K}_n$ is same as in the Assumption 4.1. Let $v_i \equiv w_i - g_{w0}(x_i)$ where $g_{w0}(x_i) \equiv E[w_i|x_i]$.

Assumption 7.1. *(Regularity conditions for Partially Linear Model)*

(i) $\{y_i, w_i, x_i\}$ *are i.i.d random variables satisfying the model* (7.1).

(ii) There exists constant $0 < c \leq C < \infty$ *such that* $E[\varepsilon_i^2|w_i, x_i] \geq c$ *and* $E[v_i^2|x_i] \geq c$, $E[\varepsilon_i^4|w_i, x_i] \leq C$ *and* $E[v_i^4|x_i] \leq C$.

(iii) $rank(P_K) = K$ *(a.s.) and* $M_{ii,K} \geq C$ *for* $C > 0$ *for all* $K \in \mathcal{K}_n$.

(iv) For all $K \in \mathcal{K}_n$, *there exists* γ_g, γ_{g_w},

$$\min_{\eta_g} E[(g_0(x_i) - \eta_g' P_{Ki})^2] = O(K^{-2\gamma_g}), \quad \min_{\eta_{gw}} E[(g_{w0}(x_i) - \eta_{g_w}' P_{Ki})^2] = O(K^{-2\gamma_{g_w}}).$$

Assumption 7.1 does not require $K/n \to 0$ which is required to get asymptotic normality in the literature (e.g., Donald and Newey (1994)). Similar to the Assumption 3.2(iii) in nonparametric setup, Assumption 7.1(iv) holds for the polynomials and splines basis. For example, 7.1(iv) holds with $\gamma_g = p_g/d_x, \gamma_{g_w} = p_w/d_x$ when $\mathcal{X}$ is compact and unknown functions $g_0(x), g_{w0}(x)$ has p_g, p_w continuous derivates, respectively.

From the results in Cattaneo, Jansson, and Newey (2015a), we have following decomposition for any $K \in \mathcal{K}_n$ under Assumptions 4.1, 7.1 and H_0,

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_K - \theta_0) &= (\frac{1}{n}W'M_KW)^{-1}\frac{1}{\sqrt{n}}W'M_KY \\ &= \widehat{\Gamma}_K^{-1}(\frac{1}{\sqrt{n}}\sum_i v_i M_{ii}^K \varepsilon_i + \frac{1}{\sqrt{n}}\sum_{i=1}^n \sum_{j=1,j\neq i}^n v_i M_{ij}^K \varepsilon_j) + o_p(1)\end{aligned} \tag{7.4}$$

where $\widehat{\Gamma}_K = W'M_KW/n$. Under homoskedasticity ($E[\varepsilon_i^2|w_i, x_i] = \sigma_\varepsilon^2$), $\sqrt{n}(\widehat{\theta}_K - \theta_0)$ is asymptotically normal with variance $V = \sigma_\varepsilon^2 E[v_i^2]^{-1}$ under any sequences $K \to \infty$ satisfying the standard rate conditions $K/n \to 0$. However, under the faster rate conditions on K imposed here, the second term in (7.4) is not negligible and converges to bounded random variables. Cattaneo, Jansson, and Newey (2015a) apply the central limit theorem of degenerate U-statistics for the second term, similar to the many instrument asymptotics analyzed in Chao, Swanson, Hausman, Newey and Woutersen (2012).

Now, consider the sequence of t-statistics $T_n(K, \theta), K \in \mathcal{K}_n$ for testing H_0. Under Assumptions 4.1, 7.1 and undersmoothing condition $nK^{-2(\gamma_g+\gamma_{g_w})} \to 0$, we get following asymptotic null limiting distributions for all deterministic sequence of $K \in \mathcal{K}_n$ assuming conditional homoskedasticity;

$$\begin{aligned}&T_n(K, \theta_0) = \sqrt{n}V_K^{-1/2}(\widehat{\theta}_K - \theta_0) \xrightarrow{d} N(0, 1), \\ &V_K = (1 - K/n)^{-1}V, \quad V = \sigma_\varepsilon^2 E[v_i^2]^{-1},\end{aligned}$$

where V_K coincides with V under $K/n \to 0$. Allowing K/n need not converge to zero requires "correction" term, $(1 - K/n)^{-1}$ taking into account for the remainder terms that are assumed "small" with the classical condition $K/n \to 0$. Note that the adjusted variance V_K is always greater than V when $K/n \nrightarrow 0$ and is an increasing function of K.

Next theorem is the main result of the partially linear model setup, analogous to nonparametric setup. Theorem 7.1 provides joint asymptotic distribution of the t-statistics $T_n(K, \theta_0)$ over $K \in \mathcal{K}_n$. It also provides the asymptotic coverage results of the CIs that are similarly defined as in Section 5 and 6.[5]

Theorem 7.1. *Suppose Assumptions 4.1 and 7.1 hold. Also, $n\bar{K}^{-2(\gamma_g+\gamma_{g_w})} \to 0$ as $\bar{K} \to \infty$. Assume $\bar{K}/n \to c$ $(0 < c < 1)$ and $E[\varepsilon_i^2|w_i, x_i] = \sigma_\varepsilon^2, E[v_i^2|x_i] = E[v_i^2]$. Then the joint null*

[5]Similar to the nonparametric setup, the asymptotic size results and the lower bounds of the asymptotic coverage for $CI_{\texttt{inf}}^{\texttt{Robust}}, CI_{\texttt{inf}}$ can be derived without undersmoothing assumption ($n\bar{K}^{-2(\gamma_g+\gamma_{g_w})} \to 0$),but omitted here for simplicity.

limiting distribution is given by

$$(T_n(K_1,\theta_0),\cdots,T_n(K_M,\theta_0))' \xrightarrow{d} Z=(Z_1,\cdots Z_M)' \sim N(0,\Sigma)$$

with variance-covariance matrix Σ *where* $\Sigma_{jl} \equiv \lim_{n\to\infty} V_{K_{j\wedge l}}^{1/2}/V_{K_{j\vee l}}^{1/2}$ *for* $j \neq l$*, and* $\Sigma_{jl}=1$ *for* $j=l$*. Moreover, under Assumptions 4.1, 4.3 and 7.1,*

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{inf}^{Robust}) = 1-\alpha, \quad \liminf_{n\to\infty} P(\theta_0 \in CI_{inf}) \geq 1-\alpha \tag{7.5}$$

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{pms}^{Robust}) \geq 1-\alpha \tag{7.6}$$

where $CI_{inf}^{Robust}, CI_{inf}$*, and* CI_{pms}^{Robust} *are similarly defined as in Section 5 and 6 with PLM estimator* $\widehat{\theta}_K$ *and variance estimator* $\widehat{V}_K$*, and the critical values* $\widehat{c}_{1-\alpha}^{\inf}, \widehat{c}_{1-\alpha}^{\sup}$*.*

Theorem 7.1 derives the joint asymptotic distribution of the $T_n(K,\theta_0)$ over $K \in \mathcal{K}_n$ for the parametric part in the partially linear model. Note that the variance-covariance matrix Σ is same as in nonparametric model setup (see equation (3.4) or (4.8)) under homoskedasticity. Variance-covariance matrix Σ_{jl} for any $j \neq l$ can be reduced under the condition $\bar{K}/n \to c$,

$$\Sigma_{jl} = \lim_{n\to\infty} \frac{V_{K_{j\wedge l}}^{1/2}}{V_{K_{j\vee l}^{1/2}}} = \lim_{n\to\infty} \frac{(1-K_{j\wedge l}/n)^{-1/2}V^{1/2}}{(1-K_{j\vee l}/n)^{-1/2}V^{1/2}} = \lim_{n\to\infty} \frac{(1-\pi_{j\wedge l}\bar{K}/n)^{-1/2}}{(1-\pi_{j\vee l}\bar{K}/n)^{-1/2}} = (\frac{1-c\pi_{j\vee l}}{1-c\pi_{j\wedge l}})^{1/2}.$$

Note that construction of CIs also requires consistent variance estimators $\widehat{V}_K$,

$$\widehat{V}_K = s^2\widehat{\Gamma}_K^{-1}, \quad s^2 = \frac{1}{n-1-K}(Y-W\widehat{\theta}_K)'M_K(Y-W\widehat{\theta}_K).$$

For consistent variance estimation results when $K/n \to c > 0$ and more discussions, see section 3.2 (Theorem 2) of Cattaneo, Jansson, and Newey (2015a) and also Cattaneo, Jansson, and Newey (2015b) even under conditional heteroskedastic error terms.

8 Monte Carlo simulations

This section investigates the small sample performance of the proposed methods in Sections 4-6. We are mainly interested in empirical coverage of CIs for the true functions $g(x)$ over the support of x with various specifications and different basis.

I consider the following data generating process similar to Newey and Powell (2003), Chen and Christensen (2015a),

$$
\begin{aligned}
y_i &= g(x_i) + \varepsilon_i, \\
x_i = \Phi(x_i^*), \begin{pmatrix} x_i^* \\ \varepsilon_i \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & \sigma^2 \end{pmatrix} \right)
\end{aligned} \tag{8.1}
$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function need to ensure compact support. I investigate following four functions for $g(x)$: $g_1(x) = 4x - 1$, $g_2(x) = \ln(|6x - 3| + 1)sgn(x - 1/2)$, $g_3(x) = \frac{\sin(7\pi x/2)}{1+2x^2(sgn(x)+1)}$, $g_4(x) = x - 1/2 + 5\phi(10(x - 1/2))$, where $\phi(\cdot)$ is the standard normal probability density function, and $sgn(\cdot)$ is the sign function. $g_1(x)$ and $g_2(x)$ are used in Newey and Powell (2003), Chen and Christensen (2015) and we label these functions as "linear" and "nonlinear" designs. $g_3(x)$ and $g_4(x)$ are rescaled versions used in Hall and Horowitz (2013), and we denote these as "highly nonlinear" designs. See Figure 2 for shapes of all four functions on the support $\mathcal{X} = [0, 1]$. In addition, I set $\sigma^2 = 1$ for all simulations results below. Results for $\sigma^2 = 0.5, 0.1$ show similar patterns, thus omitted.

I generate 5000 simulation replications for each different design with sample size $n = 100$. Then, I implement nonparametric series estimators using both power series bases and quadratic splines with evenly placed knots. In either case, K denotes the total number of estimated coefficients. I also set $\mathcal{K}_n = [2, 10]$ for the polynomials and $\mathcal{K}_n = [3, 13]$ for the splines. Then, I calculate a pointwise coverage of various CIs for all 40 grid points of x on [0,1]. To calculate critical values, 1000 additional Monte Carlo replications are also performed on each simulation iteration. Results for different sample sizes $n = 200, 400$ and results for the cubic spline regressions show similar patterns, thus omitted for brevity.

As a benchmark, I first consider post-selection CI with $\widehat{K}_{\texttt{cv}} \in \mathcal{K}_n$ selected to minimize leave-one-out cross-validation and using (naive) normal critical value, $CI^{\texttt{Naive}}_{\texttt{pms}} = [\widehat{\theta}_{\widehat{K}_{\texttt{cv}}} - z_{1-\alpha/2}s(\widehat{\theta}_{\widehat{K}_{\texttt{cv}}}), \widehat{\theta}_{\widehat{K}_{\texttt{cv}}} + z_{1-\alpha/2}s(\widehat{\theta}_{\widehat{K}_{\texttt{cv}}})]$. I also report coverage of $CI_{\texttt{maxK}} = [\widehat{\theta}_{\bar{K}} - z_{1-\alpha/2}s(\widehat{\theta}_{\bar{K}}), \widehat{\theta}_{\bar{K}} + z_{1-\alpha/2}s(\widehat{\theta}_{\bar{K}})]$ using the largest number of series terms $\bar{K}$. Next, I consider new CIs proposed in this paper, $CI^{\texttt{Robust}}_{\texttt{inf}}$ and $CI_{\texttt{inf}}$, based on the test statistics $\inf T_n(\theta)$ defined in Section 5. Finally, I examine robust post-selection CI, $CI^{\texttt{Robust}}_{\texttt{pms}}$ with $\widehat{K}_{\texttt{cv}}$, defined in Section 6. The critical values, $\widehat{c}^{\,\text{inf}}_{1-\alpha}$ and $\widehat{c}^{\,\text{sup}}_{1-\alpha}$ are constructed using the Monte Carlo methods.

Figure 3 reports nominal 95% coverage probability of all five CIs. Overall, $CI^{\texttt{Robust}}_{\texttt{inf}}$ performs well across the different simulation designs. Its empirical coverage is close to the nominal 95% level at many points over the support, even at the boundary. Coverage of $CI_{\texttt{inf}}$ is no less than the nominal level at almost all points, but $CI_{\texttt{inf}}$ using normal critical value seems quite conservative. $CI^{\texttt{Naive}}_{\texttt{pms}}$ using cross-validation selected series terms undercovers most of the cases: $\widehat{K}_{\texttt{cv}}$ is small and $CI^{\texttt{Naive}}_{\texttt{pms}}$ is somewhat narrow to cover the true function.

$CI_{\texttt{maxK}}$ slightly undercovers at many points, and works quite poorly at the boundary. $CI^{\texttt{Robust}}_{\texttt{pms}}$ using larger critical value $\widehat{c}^{\,\text{sup}}_{1-\alpha}$ seems to work well, but undercovers when there exists large bias for all K (for example, see coverage probability of $g_2(x = 0.4)$).

For the linear function $g_1(x)$, polynomials should approximate unknown function very well for all K, i.e., finite sample bias is expected to be small over $K \in \mathcal{K}_n$. In this setup, coverage of $CI^{\texttt{Robust}}_{\texttt{inf}}, CI_{\texttt{maxK}}$ are expected to be close to 95% and $CI_{\texttt{inf}}, CI^{\texttt{Robust}}_{\texttt{pms}}$ are expected to be conservative. Slightly undercover results in Figure 3-(a) for $CI_{\texttt{maxK}}$ are mostly due to the small sample sizes. However, given the small sample size, coverage of $CI^{\texttt{Robust}}_{\texttt{inf}}$ is fairly close to 95% and performs well even at the boundary compare with $CI_{\texttt{maxK}}$.

For the nonlinear function $g_2(x)$, coverage of all confidence intervals except $CI_{\texttt{inf}}$ is less than 0.95 at some particular points. For example, at $x = 0.4$ and 0.6, the coverage of $CI^{\texttt{Naive}}_{\texttt{pms}}, CI^{\texttt{Robust}}_{\texttt{pms}}$ are 0.77, 0.87, respectively. Although it is slightly below than 0.95, coverage of $CI^{\texttt{Robust}}_{\texttt{inf}}$ is 0.93, and this coincides with our theory that $CI^{\texttt{Robust}}_{\texttt{inf}}$ bounds the coverage distortions even when there exist biases for some $K \in \mathcal{K}_n$. With highly nonlinear function $g_4(x)$, $CI^{\texttt{Robust}}_{\texttt{inf}}$ does not achieve nominal coverage at point $x = 0.5$. At this single peak at $x = 0.5$, every polynomial approximation has a large bias. Possible poor coverage property at this point was also described in Hall and Horowitz (2013, Figure 3). Regression spline seems better for approximating $g_4(x)$ at this local point. Figure 4 shows the coverage probability of CIs using quadratic splines with a different number of knots. As we can see from Figure 4, $CI^{\texttt{Robust}}_{\texttt{inf}}$ with splines works better to achieve nominal coverage for $g_2(x = 0.4)$, $g_4(x = 0.5)$, and for other different functions as well.

In Figure 5, I compare the length of five CIs for the polynomial series. In the linear and nonlinear designs (for $g_1(x)$ and $g_2(x)$), the rank of the length in a narrower order is (roughly) as follows; $CI^{\texttt{Naive}}_{\texttt{pms}} < CI^{\texttt{Robust}}_{\texttt{pms}} \leq CI^{\texttt{Robust}}_{\texttt{inf}} < CI_{\texttt{maxK}} < CI_{\texttt{inf}}$. For the highly nonlinear design, the length of $CI^{\texttt{Robust}}_{\texttt{inf}}$ is similar or wider than $CI_{\texttt{maxK}}$ at some points where estimates are relatively sensitive across K. Figure 6 compares the length of CIs when splines are used, and it shows similar patterns with a polynomial approximation. Given the observations that $CI^{\texttt{Robust}}_{\texttt{inf}}$ has similar or slightly wider lengths than $CI_{\texttt{maxK}}, CI^{\texttt{Naive}}_{\texttt{pms}}$ and $CI^{\texttt{Robust}}_{\texttt{pms}}$, I want to highlight that it has better or similar coverage probability at most points than the other CIs as in Figure 4.

Note that the coverage probability of $CI_{\texttt{maxK}}$ can be better when $\bar{K}$ coincides with coverage optimal K^* that minimizes the distance $|P(\theta_0 \in CI(K)) - (1-\alpha)|$, where $CI(K)$ is a standard CI using K series terms and the normal critical value. However, as I already emphasized, there is no formal data-dependent method to choose such large enough K^*. Coverage optimal K^* also depends on the sample sizes and unknown smoothness of the underlying function. If $\bar{K}$ is smaller than the K^*, then $CI_{\texttt{maxK}}$ may undercover because of bias problems. If $\bar{K}$

is larger than K^*, then $CI_{\texttt{maxK}}$ can be too wide because of large standard errors, and it can have poor coverage because the normal distribution can be a poor approximation with large $\bar{K}$. In contrast, $CI^{\texttt{Robust}}_{\texttt{inf}}$ seems to perform well even in small sample sizes and to be less affected with the ranges of K, especially those with small K that has a large bias.

In addition to length comparisons, I also provide the power of the different test statistics. In Figure 7, I report power functions of the three different test statistics to test $H_0 : \theta = \theta_0$ against fixed alternatives $H_1 : \theta = \theta_0 + \delta$ where $\theta_0 = g_2(x)$ evaluated at some point x. As the power depends on the different point of interest x, I consider two cases where bias of series estimator for $g_2(x)$ is small ($x = 0.5$) and relatively large ($x = 0.4$). I plot the following rejection probability based on $\text{Inf}\, T_n(\theta), \text{Sup}\, T_n(\theta)$, and $|T_n(\widehat{K}, \theta)|$ with appropriate critical values as a functions of δ: (1) $P(|T_n(\widehat{K}_{\texttt{cv}}, \theta_0 + \delta)| > z_{1-\alpha/2})$ with $\widehat{K}_{\texttt{cv}}$; (2) $P(\text{Inf}\, T_n(\theta_0 + \delta) > \widehat{c}^{\text{inf}}_{1-\alpha})$; (3) $P(\text{Inf}\, T_n(\theta_0 + \delta) > z_{1-\alpha/2})$; (4) $P(\text{Sup}\, T_n(\theta_0 + \delta) > \widehat{c}^{\text{sup}}_{1-\alpha})$; (5) $P(|T_n(\widehat{K}_{\texttt{cv}}, \theta_0 + \delta)| > \widehat{c}^{\text{sup}}_{1-\alpha})$. Figure 7-(a) and (b) show that the tests based on $\text{Inf}\, T_n(\theta)$ control size or bound the size distortions when there exists bias for some Ks. The tests based on $\text{Sup}\, T_n(\theta)$ seems to have better power, but the size is not controlled even with moderate bias (Figure 7-(b)) and the size distortions can be huge with large bias for some K (Figure 7-(a)).

In sum, $CI^{\texttt{Robust}}_{\texttt{inf}}$ seems to work well in various simulation experiments: empirical coverage is close to the nominal level and less affected by bias issues. Regarding coverage, $CI_{\texttt{inf}}$ also performs well, but it can be quite conservative. In some simulation setups, coverage of $CI^{\texttt{Robust}}_{\texttt{pms}}$ is close to the nominal level, thus it is also advisable to report rather than the naive $CI^{\texttt{Naive}}_{\texttt{pms}}$.

9 Illustrative empirical application: Nonparametric estimation of labor supply function and wage elasticity with nonlinear budget sets

In this section, I illustrate our inference procedures by revisiting a paper by Blomquist and Newey (2002). Understanding how tax policy affects individual labor supply has been central issues in labor economics (see Hausman (1985) and Blundell and MaCurdy (1999), among many others). Blomquist and Newey (2002) estimate conditional mean of hours of work given the individual nonlinear budget sets using nonparametric series estimation. They also estimate other functionals such as wage elasticity of the expected labor supply and find some evidence of possible misspecification of the usual parametric model such as maximum likelihood estimation (MLE).

Specifically, Blomquist and Newey (2002) consider the following model by exploiting

additive structure from the utility maximization with piecewise linear budget sets. I use the similar notations with their paper,

$$h_i = g(x_i) + \varepsilon_i, \quad E(\varepsilon_i|x_i) = 0, \tag{9.1}$$

$$g(x_i) = g_1(y_J, w_J) + \sum_{j=1}^{J-1}[g_2(y_j, w_j, \ell_j) - g_2(y_{j+1}, w_{j+1}, \ell_j)], \tag{9.2}$$

where h_i is the hours of ith individual worked and $x_i = (y_1, \cdots, y_J, w_1, \cdots, w_J, \ell_1, \cdots, \ell_J, J)$ is the budget set that can be represented by the intercept y_j (non-labor income), slope w_j (marginal wage rates) and the end point ℓ_j of the jth segment in a piecewise linear budget with J segments. Here, Equation (9.2) for the conditional mean function follows from Theorem 2.1 of Blomquist and Newey (2002), and this additive structure substantially reduces the dimensionality issues. They consider following power series for $g(x)$,

$$p_k(x) = (y_J^{p_1(k)} w_J^{q_1(k)}, \sum_{j=1}^{J-1} \ell_j^{m(k)} (y_j^{p_2(k)} w_j^{q_2(k)} - y_{j+1}^{p_2(k)} w_{j+1}^{q_2(k)})), \quad p_2(k) + q_2(k) \geq 1. \tag{9.3}$$

From the Swedish "Level of Living" survey in 1973, 1980 and 1990, they pool the data from three waves and use the data for married or cohabiting men of ages 20-60. Changes in tax system over three different time periods gives a large variation in the budget sets. Sample size is $n = 2321$. See Section 5 of Blomquist and Newey (2002) for more detail descriptions. They estimate wage elasticity of the expected labor supply

$$E_w = \bar{w}/\bar{h}[\frac{\partial g(w, \cdots, w, \bar{y}, \cdots, \bar{y})}{\partial w}]|_{w=\bar{w}}, \tag{9.4}$$

which is the regression derivative of $g(x)$ evaluated at the mean of the net wage rates $\bar{w}$, virtual income $\bar{y}$ and level of hours $\bar{h}$.

Table 1 is the same table used in Blomquist and Newey (2002, Table 1). They report estimates $\widehat{E}_w$ and standard errors $SE_{\widehat{E}_w}$ with a different number of series terms by adding additional series terms. For example, estimates in the second raw use the term in the first row $(1, y_J, w_J)$ with the additional terms $(\Delta y, \Delta w)$. Here, $\ell^m \Delta y^p w^q$ denotes approximating term $\sum_j \ell_j^m (y_j^p w_j^q - y_{j+1}^p w_{j+1}^q)$. Blomquist and Newey (2002) also report cross-validation criteria, CV, for each specification. In their formula, series terms are chosen to maximize CV, which minimizes asymptotic MSE. In addition to their original table, I also report the standard 95% CI for each specification, i.e., $\widehat{E}_w \pm 1.96 SE_{\widehat{E}_w}$. From the table, it is ambiguous which large model (K) can be used for the inference and we do not have compelling data-dependent methods to select one of the large K for the confidence interval to be reported.

I report proposed robust confidence interval $CI^{\texttt{Robust}}_{\texttt{inf}}$ as well as $CI_{\texttt{inf}}$ and $CI^{\texttt{Robust}}_{\texttt{pms}}$ defined in Sections 5 and 6. For this, I exploit the covariance structure in the joint asymptotic distribution of the t-statistics under homoskedastic error: the variance-covariance matrix is only a function of the variance of series estimators. Therefore, construction of the critical value using the Monte Carlo method defined in (4.10) only requires estimated variance for different specifications that are already reported in the table of Blomquist and Newey (2002). We can implement critical value based on general variance forms under heteroskedasticity or bootstrap critical value using full dataset, but we do not pursue here to exploit computational advantages of our procedure. It is straightforward to construct the proposed CI without any replication of the data sets in this case. Using Monte-Carlo methods, estimated critical values are $\widehat{c}^{\text{inf}}_{1-\alpha} = 0.9668, \widehat{c}^{\text{sup}}_{1-\alpha} = 2.4764$, respectively (based on simulations using 10000 repetitions).

Robust CI based on the infimum of the t-statistics, $CI^{\texttt{Robust}}_{\texttt{inf}}$ is $[0.0271, 0.1111]$ and this is quite comparable to the CI with some large K, for example, $[0.0273, 0.1045]$ using all the additional terms up to the 6th row. Moreover, $CI^{\texttt{Robust}}_{\texttt{inf}}$ is substantially tighter than $CI_{\texttt{maxK}} = [0.0148, 0.1280]$ that uses the largest $\bar{K}$ as well as those based on the second largest series terms, $[0.0214, 0.1336]$.

$CI_{\texttt{inf}}$ using normal critical value is $[0.0148, 0.1384]$, and this turns out to be the union of CI with the largest and the third largest number of series terms. Naive post-selection CI with $\widehat{K}_{\texttt{cv}}$ is $CI^{\texttt{Naive}}_{\texttt{pms}} = [0.0247, 0.0839]$, and this seems somewhat narrow in this case. $CI^{\texttt{Robust}}_{\texttt{pms}}$ widens this naive confidence interval to $[0.0169, 0.0916]$.

Given the Table 1 reported in Blomquist and Newey (2002), $CI^{\texttt{Robust}}_{\texttt{inf}}$ seems robust to the choice of $[\underline{K}, \bar{K}]$. By sequentially excluding the largest model from the last column in Table 1 (decreasing $\bar{K}$), we get $CI^{\texttt{Robust}}_{\texttt{inf}}$ as $[0.0272, 0.1110], [0.0268, 0.1121], [0.0254, 0.0909], [0.0253, 0.0910], [0.0243, 0.0922], [0.0221, 0.0951]$ and corresponding critical values $\widehat{c}^{\text{inf}}_{1-\alpha}$ as $0.9643, 1.0043, 1.1316, 1.1433, 1.2360, 1.4541$, respectively. Moreover, by sequentially excluding the smallest model from the first column (increasing $\underline{K}$), we get $CI^{\texttt{Robust}}_{\texttt{inf}}$ as $[0.0379, 0.1144]$, $[0.0381, 0.1140], [0.0376, 0.1150], [0.0357, 0.1179], [0.0305, 0.1235]$ and the corresponding $\widehat{c}^{\text{inf}}_{1-\alpha}$ as $1.0879, 1.0732, 1.1092, 1.2136, 1.4165$, respectively. In all cases, $CI^{\texttt{Robust}}_{\texttt{inf}}$ is tighter than $CI_{\texttt{maxK}}$ with the new ranges $[\underline{K}, \bar{K}]$. Also, note that increasing $\bar{K}$ does not always increase the width of CI as it can decrease the critical value.

10 Conclusion

This paper considers the construction of inference methods given the range of different number of series terms in the nonparametric series regression model. New inference methods

proposed in this paper are based on two innovations. First, I provide an empirical process theory for the t-statistics indexed by the number of series terms over a set. Second, I introduce tests based on the infimum of the t-statistics over different series terms and show that the tests control the asymptotic size with undersmoothing condition or bound the size distortions without undersmoothing condition. Pointwise confidence interval for the true regression function is obtained by test statistic inversion. To construct the critical value and a valid CI, I suggest using a simple Monte Carlo simulation based method. In various simulation experiments, CI based on the infimum t-statistics performs well: coverage is close to the nominal level and less affected by finite sample bias. I illustrate proposed CI by revisiting empirical example of Blomquist and Newey (2002). I also provide methods of constructing a valid CI after selecting the number of series terms by adjusting the conventional normal critical value to the critical value based on the supremum of the t-statistics. Furthermore, I provide an extension of the proposed CIs in the partially linear model setup.

References

Andrews, D. W. K. (1991a): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models," *Econometrica*, 59, 307-345.

Andrews, D. W. K. (1991b): "Asymptotic Optimality of Generalized C_L, Cross-Validation, and Generalized Cross-Validation in Regression with Heteroskedastic Errors," *Journal of Econometrics*, 47, 359-377.

Andrews, D. W. K. and P. Guggenberger (2009): "Validity of Subsampling and "Plug-in Asymptotic" Inference for Parameters Defined by Moment Inequalities," *Econometric Theory*, 25, 669-709.

Armstrong, T. B. (2015): "Adaptive Testing on a Regression Function at a Point," *The Annals of Statistics*, 43, 2086-2101.

Armstrong, T. B. and M. Kolesár (2015): "A Simple Adjustment for Bandwidth Snooping," Working Paper.

Athey, S. and G.W. Imbens (2015): "A Measure of Robustness to Misspecification," *American Economic Review: Papers and Proceedings*, 105, 476-480.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015): "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Journal of Econometrics*, 186, 345-366.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*, 81, 608-650.

Blomquist, S. and W. K. Newey (2002): "Nonparametric Estimation with Nonlinear Budget Sets," *Econometrica*, 70, 2455-2480.

Blundell, R. and T. E. MaCurdy (1999): "Labor Supply: A Review of Alternative Approaches," *Handbook of Labor Economics*, In: O. Ashenfelter, D. Card (Eds.), vol. 3., Elsevier, Chapter 27.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2015): "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," Working paper.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2015a): "Alternative Asymptotics and the Partially Linear Model with Many Regressors," Working paper.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2015b): "Treatment Effects With Many Covariates and Heteroskedasticity," Working paper.

Chao, J. C., N. R. Swanson, J. A. Hausman, W. K. Newey, and T. Woutersen (2012): "Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments," *Econometric Theory*, 28, 42-86.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-nonparametric Models," *Handbook of Econometrics*, In: J.J. Heckman, E. Leamer (Eds.), vol. 6B., Elsevier, Chapter 76.

Chen, X. and T. Christensen (2015a): "Optimal Sup-Norm Rates, Adaptivity and Inference in Nonparametric Instrumental Variables Estimation," Cowles Foundation Discussion Paper 1923.

Chen, X. and T. Christensen (2015b): "Optimal Uniform Convergence Rates and Asymptotic Normality for Series Estimators Under Weak Dependence and Weak Conditions," *Journal of Econometrics*, 188, 447-465.

Chen, X. and Z. Liao (2014): "Sieve M inference on irregular parameters," *Journal of Econometrics*,182, 70-86.

Chen, X., Z. Liao, and Y. Sun (2014): "Sieve inference on possibly misspecified semi-nonparametric time series models," *Journal of Econometrics*,178, 639-658.

Chen, X. and X. Shen (1998): "Sieve extremum estimates for weakly dependent data," *Econometrica*, 66 (2), 289-314.

Donald, S. G. and W. K. Newey (1994): "Series Estimation of Semilinear Models," *Journal of Multivariate Analysis*, 50, 30-40.

Eastwood, B. J. and A.R. Gallant, (1991): "Adaptive Rules for Seminonparametric Estimators That Achieve Asymptotic Normality," *Econometric Theory*, 7, 307-340.

Gallant, A.R. and G. Souza (1991): "On the Asymptotic Normality of Fourier Flexible Form Estimates," *Journal of Econometrics*, 50, 329-353.

Hall, P. and J. Horowitz (2013): "A Simple Bootstrap Method for Constructing Nonparametric Confidence Bands for Functions," *The Annals of Statistics*, 41, 1892-1921.

Hansen B. E. (2014): "Robust Inference," Working paper.

Hansen B. E. (2015): "The Integrated Mean Squared Error of Series Regression and a Rosenthal Hilbert-Space Inequality," *Econometric Theory*, 31, 337-361.

Hansen, P.R. (2005): "A Test for Superior Predictive Ability," *Journal of Business and Economic Statistics*, 23, 365-380.

Härdle, W. and O. Linton (1994): "Applied Nonparametric Methods," *Handbook of Econometrics*, In: R. F. Engle, D. F. McFadden (Eds.), vol. 4., Elsevier, Chapter 38.

Hausman, J. A. (1985): "The Econometrics of Nonlinear Budget Sets", *Econometrica*, 53, 1255-1282.

Hausman, J. A. and W. K. Newey (1995): "Nonparametric Estimation of Exact Consumers Surplus and Deadweight Loss", *Econometrica*, 63, 1445-1476.

Heckman, J. J., L. J. Lochner, and P. E. Todd (2006): "Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond," *Handbook of the Economics of Education*, In: E. A. Hanushek, and F. Welch (Eds.), Vol. 1, Elsevier, Chapter 7.

Horowitz, J. L. (2014): "Adaptive Nonparametric Instrumental Variables Estimation: Empirical Choice of the Regularization Parameter," *Journal of Econometrics*, 180, 158-173.

Horowitz, J. L. and S. Lee (2012): "Uniform Confidence Bands for Functions Estimated Nonparametrically with Instrumental Variables," *Journal of Econometrics*, 168, 175-188.

Horowitz, J. L. and V. G. Spokoiny (2001): "An Adaptive, Rate-Optimal Test of a Parametric Mean-Regression Model Against a Nonparametric Alternative" *Econometrica*, 69, 599-631.

Huang, J. Z. (2003a): "Asymptotics for Polynomial Spline Regression Under Weak Conditions," *Statistics & Probability Letters*, 65, 207-216.

Huang, J. Z. (2003b): "Local Asymptotics for Polynomial Spline Regression," *The Annals of Statistics*, 31, 1600-1635.

Ichimura H. and P. E. Todd (2007): "Implementing Nonparametric and Semiparametric Esstimators," *Handbook of Econometrics*, In: J.J. Heckman, E. Leamer (Eds.), vol. 6B., Elsevier, Chapter 74.

Leamer, E. E. (1983): "Let's Take the Con Out of Econometrics," *The American Economic Review*, 73, 31-43.

Lepski, O. V. (1990): "On a problem of adaptive estimation in Gaussian white noise," *Theory of Probability and its Applications*, 35, 454-466.

Li, K. C. (1987): "Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15, 958-975.

Li, Qi, and J. S. Racine (2007): *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Linton, O. (1995): "Second order approximation in the partialy linear regression model," *Econometrica*, 63(5), 1079-1112.

Newey, W. K. (1994a): "Series Estimation of Regression Functionals," *Econometric Theory*, 10, 1-28.

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

Newey, W. K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147-168.

Newey, W. K. (2013): "Nonparametric Instrumental Variables Estimation," *American Economic Review: Papers & Proceedings*, 103, 550-556.

Newey, W. K. and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565-1578.

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

Robinson, P. M. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56(4), 931-954.

Romano, J. P. and M. Wolf (2005): "Stepwise Multiple Testing as Formalized Data Snooping," *Econometrica*, 73, 1237-1282.

Schennach, S. M. (2015): "A bias bound approach to nonparametric inference," *CEMMAP working paper CWP71/15*.

Tropp, J. A. (2015): *An Introduction to Matrix Concentration Inequalities*, Foundations and Trends in Machine Learning, Vol. 8: No.1-2, 1-230.

Van Der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

Varian, H. R. (2014): "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, 28, 3-28.

White, H. (2000): "A Reality Check for Data Snooping," *Econometrica*, 68, 1097-1126.

Zhou, S., X. Shen, and D.A. Wolfe (1998): "Local Asymptotics for Regression Splines and Confidence Regions," *The Annals of Statistics*, 26, 1760-1782.

A Proofs

In the Appendix, we define additional notations for the empirical process theory used in the proof of Theorem 3.1. Given measurable space $(S, \mathcal{S})$, let $\mathcal{F}$ as a class of measurable functions $f : \mathcal{S} \to \mathbb{R}$. For any probability measure Q on $(S, \mathcal{S})$, we define $N(\epsilon, \mathcal{F}, L_2(Q))$ as covering numbers relative to the $L_2(Q)$ norms, which is the minimal number of the $L_2(Q)$ balls of radius ϵ to cover $\mathcal{F}$ with $L_2(Q)$ norms $||f||_{Q,2} = (\int |f|^2 dQ)^{1/2}$. Uniform entropy numbers relative to L_2 are defined as $\sup_Q \log N(\epsilon ||F||_{Q,2}, \mathcal{F}, L_2(Q))$ where the supremum is over all discrete probability measures with an envelope function F.

Let the data $z_i = (\varepsilon_i, x_i)$ be i.i.d. random vectors defined on probability space $(\mathcal{Z} = \mathcal{E} \times \mathcal{X}, \mathcal{A}, P)$ with common probability distribution $P \equiv P_{\varepsilon,x}$. We think of $(\varepsilon_1, x_1), \cdots (\varepsilon_n, x_n)$ as the coordinates of the infinite product probability space. For notational convenience, we avoid discussing nonmeasurability issues and outer expectations (for the related issues, see van der Vaart and Wellner (1996)). Throughout the proofs, we denote $c, C > 0$ as a universal constant that does not depend on n.

A.1 Proof of Theorem 3.1

For any sequence $\{K(n) = \pi \bar{K}(n) : n \geq 1\} \in \prod_{n=1}^{\infty} \mathcal{K}_n$ under Assumptions 3.1 and 3.2, we first define orthonormalized vector of basis functions

$$\tilde{P}_K(x) \equiv Q_K^{-1/2} P_K(x) = E[P_{Ki} P_{Ki}']^{-1/2} P_K(x),$$
$$\tilde{P}_{Ki} = \tilde{P}_K(x_i), \tilde{P}^K = [\tilde{P}_{K1}, \cdots, P_{Kn}]'$$

We observe that

$$\widehat{g}_K(x) = P_K(x)'(P^{K\prime} P^K)^{-1} P^{K\prime} y = \tilde{P}_K(x)'(\tilde{P}^{K\prime} \tilde{P}^K)^{-1} \tilde{P}^{K\prime} y,$$
$$V_K(x) = P_K(x)' Q_K^{-1} \Omega_K Q_K^{-1} P_K(x) = \tilde{P}_K(x)' \tilde{\Omega}_K \tilde{P}_K(x),$$
$$\tilde{\Omega}_K = E(\tilde{P}_{Ki} \tilde{P}_{Ki}' \varepsilon_i^2).$$

Without loss of generality, we may impose normalization of $Q_{\bar{K}} = I$ or $Q_K = E(P_{Ki} P_{Ki}') = I_K$ uniformly over $K \in \mathcal{K}_n$, since $\widehat{g}_K(x)$ is invariant to nonsingular linear transformations of $P_K(x)$. However, we shall treat Q_K as unknown and deal with non-orthonormalized series terms here.

Next, we re-define pseudo-true value β_K in (2.2), with abuse of notation, using orthornormalized series terms $\tilde{P}_{Ki}$. That is, $y_i = \tilde{P}_{Ki}' \beta_K + \varepsilon_{Ki}, E[\tilde{P}_{Ki} \varepsilon_{Ki}] = 0$ where $\varepsilon_{Ki} = r_{Ki} + \varepsilon_i$, $r_K(x) = g_0(x) - \tilde{P}_K(x)' \beta_K, r_{Ki} = r_K(x_i)$, and $r_K \equiv (r_{K1}, \cdots r_{Kn})'$. We also define $\widehat{Q}_K \equiv$

$\frac{1}{n}\tilde{P}^{K\prime}\tilde{P}^{K}$, $\underline{\sigma}^2 \equiv \inf_x E[\varepsilon_i^2|x_i = x], \bar{\sigma}^2 \equiv \sup_x E[\varepsilon_i^2|x_i = x]$. We first provide useful lemmas which will be used in the proof of Theorem 3.1. Versions of proof of Lemma 1 are available in the literature, such as Newey (1997), Belloni et al. (2015) and Chen and Christensen (2015b), among others. For completeness, we provide the results of Lemma 1. Note that different rate conditions can be used in Assumption 3.2, but lead to different bounds in (A.1)-(A.3) in the following Lemma 1.

Lemma 1. *Under Assumptions 3.1 and 3.2, for any $K \in \mathcal{K}_n$, following holds*

$$||\widehat{Q}_K - I_K|| = O_p(\sqrt{\frac{\zeta_K^2 \lambda_K^2 \log K}{n}}), \tag{A.1}$$

$$R_1(K) \equiv \sqrt{\frac{1}{nV_K}}\tilde{P}_K(x)'\left(\widehat{Q}_K^{-1} - I_K\right)\tilde{P}^{K\prime}(\varepsilon + r_K) = O_p(\sqrt{\frac{\lambda_K^2\zeta_K^2 \log K}{n}}(1 + \ell_K c_K\sqrt{K})), \tag{A.2}$$

$$R_2(K) \equiv \sqrt{\frac{1}{nV_K}}\tilde{P}_K(x)'\tilde{P}^{K\prime}r_K = O_p(\ell_K c_K). \tag{A.3}$$

To provide (A.1) in Lemma 1, we first introduce matrix Bernstein inequality in Tropp (2015).

Lemma 2 (Theorem 6.1.1 of Tropp (2015))**.** *Consider a finite sequence $\{S_i\}$ of independent, random matrices with common dimension $d_1 \times d_2$. Assume that $ES_i = 0$, $||S_i|| \leq L$ for each i. Let $Z = \sum_i S_i$, and define*

$$v(Z) = \max\{||E(ZZ')||, ||E(Z'Z)||\}.$$

Then,

$$P(||Z|| \geq t) \leq (d_1 + d_2)\exp(\frac{-t^2/2}{v(Z)Lt/3}), \qquad \forall t \geq 0,$$
$$E||Z|| \leq \sqrt{2v(Z)\log(d_1 + d_2)} + \frac{1}{3}L\log(d_1 + d_2).$$

Proof of Lemma 1.

To provide bound in (A.1), we apply Lemma 2 by setting $S_i = \frac{1}{n}(\tilde{P}_{Ki}\tilde{P}_{Ki}' - E(\tilde{P}_{Ki}\tilde{P}_{Ki}'))$. Note that $\mathbb{E}\,S_i = 0$, $||S_i|| \leq L = \frac{1}{n}(\lambda_K^2\zeta_K^2 + 1)$, and $v(Z) = \frac{1}{n}||E(\tilde{P}_{Ki}\tilde{P}_{Ki}'\tilde{P}_{Ki}\tilde{P}_{Ki}') - E(\tilde{P}_{Ki}\tilde{P}_{Ki}')E(\tilde{P}_{Ki}\tilde{P}_{Ki}')|| \leq \frac{1}{n}(\lambda_K^2\zeta_K^2 + 1)$ by definition of λ_K, ζ_K and $E(\tilde{P}_{Ki}\tilde{P}_{Ki}') = I_K$. By

Lemma 2, we have

$$E||\widehat{Q}_K - I_K|| = E||\sum_i \frac{1}{n}(\tilde{P}_{Ki}\tilde{P}_{Ki}' - I_K)|| \leq C(\sqrt{\lambda_K^2\zeta_K^2\log(K)/n} + \lambda_K^2\zeta_K^2\log(K)/n).$$

Then we have $||\widehat{Q}_K - I_K|| = O_P(\sqrt{\lambda_K^2\zeta_K^2\log(K)/n})$ by Markov inequality.

For (A.2), we first look at the terms $\sqrt{\frac{1}{nV_K}}\tilde{P}_K(x)'\left(\widehat{Q}_K^{-1} - I_K\right)\tilde{P}^{K'}\varepsilon$. Conditional on the sample $X = [x_1, \cdots, x_n]$, this term has mean zero and variance,

$$\begin{aligned}
&\frac{1}{nV_K}\tilde{P}_K(x)'\left(\widehat{Q}_K^{-1} - I_K\right)\tilde{P}^{K'}E(\varepsilon\varepsilon'|X)\tilde{P}^{K}\left(\widehat{Q}_K^{-1} - I_K\right)\tilde{P}_K(x) \\
&\leq \frac{\bar{\sigma}^2}{V_K}\tilde{P}_K(x)'\left(\widehat{Q}_K^{-1} - I_K\right)\widehat{Q}_K\left(\widehat{Q}_K^{-1} - I_K\right)\tilde{P}_K(x) \\
&= \frac{\bar{\sigma}^2}{V_K}\tilde{P}_K(x)'\left(\widehat{Q}_K - I_K\right)\widehat{Q}_K^{-1}\left(\widehat{Q}_K - I_K\right)\tilde{P}_K(x) \\
&\leq \frac{\bar{\sigma}^2\tilde{P}_K(x)'\tilde{P}_K(x)}{V_K}\lambda_{max}(\widehat{Q}_K^{-1})||\left(\widehat{Q}_K - I_K\right)||^2 \\
&= O_P(\lambda_K^2\zeta_K^2\log(K)/n)
\end{aligned}$$

where the first and the last inequality uses $V_K \leq \bar{\sigma}^2\tilde{P}_K(x)'\tilde{P}_K(x)$, $V_K \geq \underline{\sigma}^2\tilde{P}_K(x)'\tilde{P}_K(x)$ by Assumption 3.2(ii), $||\widehat{Q}_K - I_K|| = O_P(\sqrt{\lambda_K^2\zeta_K^2\log(K)/n})$ by (A.1) and $\lambda_{max}(\widehat{Q}_K^{-1}) = (\lambda_{max}(\widehat{Q}_K))^{-1} = O_p(1)$ since all eigenvalues of $\widehat{Q}_K$ are bounded away from zero as $|\lambda_{min}(\widehat{Q}_K) - 1| \leq ||\widehat{Q}_K - I_K|| = o_p(1)$ by (A.1) and Assumption 3.2(iv)-(v). Then, by Chebyshev's inequality, we have that

$$\sqrt{\frac{1}{nV_K}}\tilde{P}_K(x)'\left(\widehat{Q}_K^{-1} - I_K\right)\tilde{P}^{K'}e = O_P(\sqrt{\lambda_K^2\zeta_K^2\log(K)/n}).$$

Next, consider the terms $\sqrt{\frac{1}{nV_K}}\tilde{P}_K(x)'\left(\widehat{Q}_K^{-1} - I_K\right)\tilde{P}^{K'}r_K$. Observe that $||\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{P}_{Ki}r_{Ki}|| = O_p(\ell_K c_K\sqrt{K})$ since

$$E[||\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{P}_{Ki}r_{Ki}]||^2] = E[\sum_{j=1}^K \tilde{P}_{ji}^2 r_{Ki}^2] \leq \ell_K^2 c_K^2 E[||\tilde{P}_{Ki}||^2] = \ell_K^2 c_K^2 K. \tag{A.4}$$

Combining (A.1) and (A.4) yields the results

$$|\sqrt{\frac{1}{nV_K}}\tilde{P}_K(x)'\left(\widehat{Q}_K^{-1}-I_K\right)\tilde{P}^{K'}r_K| \leq C||\widehat{Q}_K^{-1}||\cdot||\left(\widehat{Q}_K-I_K\right)||||\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{P}_{Ki}r_{Ki}||$$
$$= O_p(\sqrt{\frac{\lambda_K^2\zeta_K^2\log(K)}{n}}\ell_K c_K\sqrt{K})$$

by $||\frac{\tilde{P}_K(x)}{V_K^{1/2}}|| \asymp 1$ and using $||\widehat{Q}_K^{-1}|| = O_p(1)$.

We now prove (A.3). Consider $\sqrt{\frac{1}{nV_K}}\tilde{P}_K(x)'\tilde{P}^{K'}r_K$,

$$E[(\sqrt{\frac{1}{nV_K}}\tilde{P}_K(x)'\tilde{P}^{K'}r_K)^2] = E[(\frac{\tilde{P}_K(x)'\tilde{P}_{Ki}}{V_K^{1/2}}r_{Ki})^2] \leq (c_K\ell_K)^2$$

since $E[(\frac{\tilde{P}_K(x)'\tilde{P}_{Ki}}{V_K^{1/2}})^2] \asymp 1$ by Assumption 3.2(ii) and $E(r_{Ki})^2 \leq (\ell_K c_K)^2$ by Assumption 3.2(iii). Therefore, we have (A.3) by Chebyshev's inequality and using $E[\tilde{P}_{Ki}r_{Ki}] = 0$ from projection model. This completes the proof. *Q.E.D.*

Proof of Theorem 3.1. For any $\pi \in \Pi = [\underline{\pi}, 1]$, we first show the decomposition of the t-statistic in equation (3.2).

$$\begin{aligned}
T_n^*(\pi,\theta_0) &= T_n(\lfloor\pi\bar{K}\rfloor,\theta) \\
&= \sqrt{\frac{n}{V_\pi}}\tilde{P}_\pi(x)'(\widehat{\beta}_{\lfloor\pi\bar{K}\rfloor}-\beta_{\lfloor\pi\bar{K}\rfloor}) - \sqrt{\frac{n}{V_\pi}}r_\pi \\
&= \sqrt{\frac{1}{nV_\pi}}\tilde{P}_\pi(x)'\tilde{P}^{\lfloor\pi\bar{K}\rfloor'}(\varepsilon + r_{\lfloor\pi\bar{K}\rfloor'}) \\
&\quad + \sqrt{\frac{1}{nV_\pi}}\tilde{P}_\pi(x)'\left(\widehat{Q}_{\lfloor\pi\bar{K}\rfloor}^{-1}-I_{\lfloor\pi\bar{K}\rfloor}\right)\tilde{P}^{\lfloor\pi\bar{K}\rfloor'}(\varepsilon + r_{\lfloor\pi\bar{K}\rfloor'}) - \sqrt{\frac{n}{V_\pi}}r_\pi \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\tilde{P}_\pi(x)'\tilde{P}_{\pi i}\varepsilon_i}{V_\pi^{1/2}} + R_1(\lfloor\pi\bar{K}\rfloor) + R_2(\lfloor\pi\bar{K}\rfloor) - \sqrt{n}V_\pi^{-1/2}r_\pi
\end{aligned}$$

where $R_1(K), R_2(K)$ are defined in (A.2), (A.3).

By Lemma 1, we have $R_1(K) = O_p(\sqrt{\frac{\zeta_K^2\log K}{n}}(1+\ell_K c_K\sqrt{K})) = o_p(1)$, $R_2(K) = O_p(\ell_K c_K) = o_p(1)$ for any $K = \pi\bar{K} \in \mathcal{K}_n$ under Assumptions 3.1 and 3.2. Therefore we have following decomposition for any $\pi \in \Pi$,

$$T_n^*(\pi,\theta_0) = t_n^*(\pi) - \sqrt{n}V_\pi^{-1/2}r_\pi + o_p(1), \tag{A.5}$$

where

$$t_n^*(\pi) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\tilde{P}_\pi(x)' \tilde{P}_\pi(x_i)\varepsilon_i}{V_\pi^{1/2}}. \tag{A.6}$$

For given $n \geq 1$, $\pi \in \Pi$, define functions $f_{n,\pi} : (\mathcal{E} \times \mathcal{X}) \mapsto \mathbb{R}$,

$$f_{n,\pi}(\varepsilon, t) = \frac{\tilde{P}_\pi(x)' \tilde{P}_\pi(t)\varepsilon}{V_\pi^{1/2}(x)} = \frac{\tilde{P}_{\lfloor \bar{K}\pi \rfloor}(x)' \tilde{P}_{\lfloor \bar{K}\pi \rfloor}(t)\varepsilon}{V_{\lfloor \bar{K}\pi \rfloor}^{1/2}(x)}, (\varepsilon, t) \in \mathcal{E} \times \mathcal{X}. \tag{A.7}$$

Consider the class of measurable functions $\mathcal{F}_n = \{f_{n,\pi} : \pi \in \Pi\}$. Then, we consider following empirical process

$$\left\{ t_n^*(\pi) : \pi \in \Pi \right\} = \left\{ n^{-1/2} \sum_{i=1}^{n} f_{n,\pi}(\varepsilon_i, x_i) : \pi \in \Pi \right\}$$

which is indexed by classes of functions $\mathcal{F}_n$.

We want to show weak convergence of the empirical process $\{t_n^*(\cdot) : n \geq 1\}$ to a centered Gaussian process, $\mathbb{T}(\cdot)$ defined in the Theorem 3.1, in the space $\ell^\infty(\Pi)$ with totally bounded semimetric space (Π, ρ), where ρ is defined as $\rho(\pi_1, \pi_2) = |\pi_1 - \pi_2|$. Weak convergence results follows from marginal convergence to a multivariate normal distribution and asymptotic tightness. We closely follow Section 2.11.3 in van der Vaart and Wellner (1996) and verify conditions for the asymptotic tightness as in Theorem 2.11.22.

Note that the covariance kernel can be derived as follows

$$Ef_{n,\pi_1} f_{n,\pi_2} - Ef_{n,\pi_1} Ef_{n,\pi_2} = \frac{\tilde{P}_{\pi_1}(x)' E(\tilde{P}_{\pi_1}(x_i)\tilde{P}_{\pi_2}(x_i)'\varepsilon_i^2)\tilde{P}_{\pi_2}(x)}{V_{\pi_1}^{1/2} V_{\pi_2}^{1/2}}, \tag{A.8}$$

for any $\underline{\pi} \leq \pi_1 \leq \pi_2 \leq 1$. This term converges to the claimed covariance kernel $\Sigma(\pi_1, \pi_2)$ under Assumption 3.3(i). This covariance kernel can be bounded below and above some constant $0 < C_1, C_2 < \infty$ for all n,

$$0 < C_1 \leq \underline{\sigma}^2 \frac{V_{\pi_1}^{1/2}}{V_{\pi_2}^{1/2}} \leq \frac{\tilde{P}_{\pi_1}(x)' E(\tilde{P}_{\pi_1}(x_i)\tilde{P}_{\pi_2}(x_i)'\varepsilon_i^2)\tilde{P}_{\pi_2}(x)}{V_{\pi_1}^{1/2} V_{\pi_2}^{1/2}} \leq \bar{\sigma}^2 \frac{V_{\pi_1}^{1/2}}{V_{\pi_2}^{1/2}} \leq C_2 \tag{A.9}$$

by using $\underline{\sigma}^2 \tilde{P}_\pi(x)' \tilde{P}_\pi(x) \leq V_\pi \leq \bar{\sigma}^2 \tilde{P}_\pi(x)' \tilde{P}_\pi(x)$ from Assumption 3.2(ii). We also use the fact that $V_{\pi_1}^{1/2} \asymp V_{\pi_2}^{1/2} \asymp ||\tilde{P}_{\bar{K}}||$ for any $\pi_1, \pi_2 \in \Pi$ under Assumption 3.3(ii).

To show the finite dimensional convergence, by the Cramér-Wold device, it suffices to

show that for any $0 < \underline{\pi} \leq \pi_1 < \cdots < \pi_M \leq 1$,

$$\delta' t_n^* \xrightarrow{d} N(0, \delta'\Sigma\delta) \quad \forall \delta \in \mathbb{R}^M \tag{A.10}$$

where $t_n^* = (t_n^*(\pi_1), ..., t_n^*(\pi_M))'$, $\Sigma_{jl} = \lim_{n\to\infty} \Sigma_{jl,n}$, $\Sigma_{jl,n} \equiv \frac{\tilde{P}_{\pi_j}(x)' E(\tilde{P}_{\pi_j i}\tilde{P}'_{\pi_l i}\varepsilon_i^2)\tilde{P}_{\pi_l}(x)}{V_{\pi_j}^{1/2} V_{\pi_l}^{1/2}}$. To show (A.10) we will verify Lindberg's condition of the CLT for $\frac{1}{\sqrt{n}}\sum_{i=1}^n \omega_{ni} \xrightarrow{d} N(0,1)$, where $\omega_{ni} = (\delta'\Sigma_n\delta)^{-1/2}\sum_{j=1}^M \delta_j \frac{\tilde{P}_{\pi_j}(x)'\tilde{P}_{\pi_j i}\varepsilon_i}{V_{\pi_j}^{1/2}}$. Note that $E\omega_{ni} = 0$, and $\frac{1}{n}\sum_{i=1}^n E[\omega_{ni}^2] = 1$, since $E[\omega_{ni}^2] = (\delta'\Sigma_n\delta)^{-1}\delta' Var(f_n(\varepsilon_i, x_i))\delta = 1$, where $f_n(\varepsilon_i, x_i) = (f_{n,\pi_1}(\varepsilon_i, x_i), ..., f_{n,\pi_M}(\varepsilon_i, x_i))'$. By Assumption 3.2, we have $||\sum_{j=1}^M \delta_j \frac{\tilde{P}_{\pi_j}(x)'\tilde{P}_{\pi_j i}}{V_{\pi_j}^{1/2}}||_\infty \lesssim \zeta_{\bar{K}}\lambda_{\bar{K}}$. Moreover, $(\delta'\Sigma_n\delta)^{-1} \lesssim 1$. Therefore, for any $a > 0$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n E(|\omega_{ni}|^2 1\{|\omega_{ni}| > a\sqrt{n}\}) \\
&\lesssim M\sum_{j=1}^M E[|\frac{\tilde{P}_{\pi_j}(x)'\tilde{P}_{\pi_j i}\varepsilon_i}{V_{\pi_j}^{1/2}}|^2 1\{|\sum_{j=1}^M \delta_j \frac{\tilde{P}_{\pi_j}(x)'\tilde{P}_{\pi_j i}}{V_{\pi_j}^{1/2}}\varepsilon_i| > a\sqrt{n}\}] \\
&\leq M\sum_{j=1}^M E(|\frac{\tilde{P}_{\pi_j}(x)'\tilde{P}_{\pi_j i}}{V_{\pi_j}^{1/2}}|^2) \sup_x E[\varepsilon_i^2 1\{|\varepsilon_i| > a(\sqrt{n}/(\zeta_{\bar{K}}\lambda_{\bar{K}})\}|x_i = x],
\end{aligned}$$

where the last term goes to 0 under $n \to \infty$ by Assumption 3.2(ii), since $E[(\frac{\tilde{P}_\pi(x)'\tilde{P}_{\pi i}}{V_\pi^{1/2}})^2] \asymp 1$ for any π and $(\zeta_{\bar{K}}\lambda_{\bar{K}})/\sqrt{n} = o(1)$ by Assumption 3.2(iv). Thus, Lindberg condition is verified and therefore (A.10) holds by Lindberg-Feller CLT and Slutzky's Theorem. We show that the finite dimensional convergence to a Gaussian distribution with covariance kernel in the Theorem 3.1.

Now, we only need to show stochastic equicontinuity. Define $\alpha(x,\pi) \equiv \tilde{P}_\pi(x)/V_\pi^{1/2}(x) = \tilde{P}_\pi(x)/||\Omega_\pi^{1/2}\tilde{P}_\pi(x)||$. Note that $|f_{n,\pi}(\varepsilon, t)| = |\alpha(x,\pi)'P_\pi(t)\varepsilon| \leq C|f_{n,1}(\varepsilon,t)| \leq C|\varepsilon|\zeta_{\bar{K}}\lambda_{\bar{K}}$. We define envelope function $F_n(\varepsilon, t) \equiv |f_{n,1}(\varepsilon,t)| \vee 1$. Without loss of generality, we assume that $F_n \geq 1$. Note that $Ef_{n,\pi}^2 = 1$ for any π, thus $EF_n^2 = O(1)$. Moreover, Lindeberg conditions can be verified easily as follows. For any $a > 0$,

$$\begin{aligned}
E(F_n^2 1\{F_n > a\sqrt{n}\}) &= E[(\frac{\tilde{P}_1(x)'\tilde{P}_1(x_i)}{V_{\underline{\pi}}^{1/2}}\varepsilon_i)^2 1\{|\varepsilon_i| > a(\sqrt{n}/(\zeta_{\bar{K}}\lambda_{\bar{K}})\}] &\text{(A.11)}\\
&\leq \sup_x E[\varepsilon_i^2 1\{|\varepsilon_i| > a(\sqrt{n}/(\zeta_{\bar{K}}\lambda_{\bar{K}})\}|X_i = x] = o(1) &\text{(A.12)}
\end{aligned}$$

since $(\zeta_{\bar{K}}\lambda_{\bar{K}})/\sqrt{n} = o(1)$ and Assumption 3.2(ii). Moreover, for every $\delta_n \to 0$,

$$\sup_{\rho(\pi_1,\pi_2)<\delta_n} E(f_{n,\pi_1} - f_{n,\pi_2})^2 \to 0 \tag{A.13}$$

since $Ef_{n,\pi_1}f_{n,\pi_2} \to 1$ as $\rho(\pi_1,\pi_2) \to 0$.

Define $\kappa_{1,n} \equiv \sup_{\pi\neq\pi'} \frac{||\tilde{P}_{\pi'-\pi}(x)||}{||\pi'-\pi||}$ where $\tilde{P}_{\pi'-\pi}(x) = (\tilde{p}_{\lfloor \bar{K}\pi\rfloor+1}(x), ..., \tilde{p}_{\lfloor \bar{K}\pi'\rfloor}(x))'$. For sufficiently large n, $\kappa_{1,n} \lesssim ||\tilde{P}_{\pi'-\pi}(x)|| \lesssim V^{1/2}_{\pi'-\pi}(x)$ under Assumption 3.2 and 3.3(ii). Also define $\kappa_{2,n} \equiv \sup_{\pi\neq\pi'} \frac{|V_{\pi'}(x)-V_\pi(x)|}{||\pi'-\pi||}$.

Then, for any $\pi, \pi' \in \Pi = [\underline{\pi}, 1]$ such that $\pi < \pi'$, following holds for sufficiently large n,

$$|\alpha(x,\pi')'P_{\pi'}(t) - \alpha(x,\pi)'P_\pi(t)| = |\frac{\tilde{P}_{\pi'}(x)'\tilde{P}_{\pi'}(t)}{V^{1/2}_{\pi'}(x)} - \frac{\tilde{P}_\pi(x)'\tilde{P}_\pi(t)}{V^{1/2}_\pi(x)}| \tag{A.14}$$

$$\leq |\frac{\tilde{P}_{\pi'}(x)'\tilde{P}_{\pi'}(t) - \tilde{P}_\pi(x)'\tilde{P}_\pi(t)}{V^{1/2}_{\pi'}(x)}| + |\tilde{P}_\pi(x)'\tilde{P}_\pi(t)(\frac{1}{V^{1/2}_{\pi'}(x)} - \frac{1}{V^{1/2}_\pi(x)})| \tag{A.15}$$

$$\leq (\sup_\pi \frac{1}{|V^{1/2}_\pi(x)|})|\tilde{P}_{\pi'-\pi}(x)'\tilde{P}_{\pi'-\pi}(t)| + |\frac{\tilde{P}_\pi(x)'\tilde{P}_\pi(t)}{V^{1/2}_\pi(x)}(\frac{V_{\pi'}(x)-V_\pi(x)}{V^{1/2}_{\pi'}(x)(V^{1/2}_\pi(x)+V^{1/2}_{\pi'}(x))})| \tag{A.16}$$

$$\leq C_1(\sup_\pi \frac{1}{|V^{1/2}_\pi(x)|})\kappa_{1,n}\zeta_{\bar{K}}\lambda_{\bar{K}}||\pi'-\pi|| + C_2\zeta_{\bar{K}}\lambda_{\bar{K}}\frac{1}{\inf_\pi|V_\pi(x)|}\kappa_{2,n}||\pi'-\pi|| \tag{A.17}$$

$$\leq C_3\zeta_{\bar{K}}\lambda_{\bar{K}}||\pi'-\pi|| + C_4\zeta_{\bar{K}}\lambda_{\bar{K}}||\pi'-\pi|| = A\zeta_{\bar{K}}\lambda_{\bar{K}}||\pi'-\pi|| \tag{A.18}$$

where C_1, C_2, C_3, C_4, A are some constants do not depend on n. The third inequality uses the definition of $\kappa_{1,n}, \kappa_{2,n}$, $|\tilde{P}_{\pi'-\pi}(t)| \lesssim \zeta_{\bar{K}}\lambda_{\bar{K}}$ and $|\frac{\tilde{P}_\pi(x)'\tilde{P}_\pi(t)}{V^{1/2}_\pi(x)}| \lesssim \zeta_{\bar{K}}\lambda_{\bar{K}}$ under Assumption 3.1 and 3.2. The last inequality uses $\kappa_{1,n} \lesssim V^{1/2}_{\pi'-\pi}(x)$, $\kappa_{2,n} \lesssim \sup_\pi V_\pi(x)$, and $V_\pi(x) \asymp V_{\pi'}(x)$ for any $\pi, \pi' \in \Pi$ under Assumption 3.3(ii).

From this, we have

$$|f_{n,\pi'} - f_{n,\pi}| = |\varepsilon\alpha(x,\pi')'P_{\pi'}(t) - \varepsilon\alpha(x,\pi)'P_\pi(t)| \leq |\varepsilon|A\zeta_{\bar{K}}\lambda_{\bar{K}}||\pi'-\pi||. \tag{A.19}$$

Therefore, the class of functions $\mathcal{F}_n = \{f_{n,\pi} : \pi \in \Pi\}$ satisfies Lipschitz conditions for each n, and this implies that there are constants $A, V > 0$ such that

$$\sup_Q N(\epsilon||F_n||_{L^2(Q)}, \mathcal{F}_n, L^2(Q)) \leq (A/\epsilon)^V, 0 < \forall\epsilon \leq 1 \tag{A.20}$$

for each n. Then, following uniform-entropy condition holds for every $\delta_n \to 0$.

$$J(\delta_n, \mathcal{F}_n, L^2(Q)) = \int_0^{\delta_n} \sqrt{\log \sup_Q N(\epsilon||F_n||_{L^2(Q)}, \mathcal{F}_n, L^2(Q))} \longrightarrow 0. \tag{A.21}$$

Thus, by the Theorem 2.11.22 in van der Vaart and Wellner (1996), we have shown that the sequence $\{t_n^*(\pi) : \pi \in \Pi\}$ is asymptotically tight in $\ell^\infty(\Pi)$. Together with the definition of $\nu(\pi) = \lim_{n\to\infty} -\sqrt{n}V_\pi^{-1/2} r_\pi$ and the equation (A.5), we have $T_n^*(\pi, \theta_0) \Rightarrow \mathbb{T}(\pi) + \nu(\pi)$ for $\pi \in \Pi$. In addition, if Assumption 3.4 holds, then $|\sqrt{n}V_\pi^{-1/2} r_\pi| = O(\sqrt{n}V_\pi^{-1/2}\ell_{\lfloor \pi \bar{K} \rfloor} c_{\lfloor \pi \bar{K} \rfloor}) = o(1)$ for any $\pi \in \Pi$. Therefore, $T_n^*(\pi, \theta_0) \Rightarrow \mathbb{T}(\pi)$. This completes the proof.

Q.E.D.

A.2 Proof of Theorem 3.2

Proof. We prove the finite dimensional convergence using similar arguments to those used in the proof of Theorem 3.1. We repeat this here, as Assumption 3.5 impose different rates of K compare with the Assumption 3.1. If some elements of $|\nu_m| = +\infty$ under oversmoothing sequences, joint distribution of $(T_n(K_1, \theta_0), \cdots, T_n(K_M, \theta_0))'$ does not converge in distribution to a profer bounded random vector. Thus, continuous mapping theorem cannot be directly applied to obtain asymptotic distribution results. To circumvent this issue, remaining proof use the same type of argument as in Theorem 1 of Andrews and Guggenberger (2009) in the moment inequality literature.

By Lemma 1 and similar arguments as in Theorem 3.1, we have following decompositions for any $m = 1, 2, \cdots M$,

$$T_n(K_m, \theta_0) = t_n(m) + \nu_n(m) + o_p(1),$$

where $t_n(m) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\tilde{P}_{K_m}(x)' \tilde{P}_{K_m i} \varepsilon_i}{V_{K_m}^{1/2}}$ and $\nu_n(m) = -\sqrt{n}V_{K_m}^{-1/2} r_{K_m}(x)$ is defined in Assumption 3.2. To obtain joint asymptotic distribution of $t_n(m)$, we need to show

$$\delta' t_n \xrightarrow{d} N(0, \delta'\Sigma\delta), \quad \forall \delta \in \mathbb{R}^M \tag{A.22}$$

where $t_n = (t_n(1), ..., t_n(M))', \Sigma_{jl} = \lim_{n\to\infty} \Sigma_{jl,n}$, $\Sigma_{jl,n} \equiv \frac{\tilde{P}_{K_j}(x)' E(\tilde{P}_{K_j i} \tilde{P}'_{K_l i} \varepsilon_i^2) \tilde{P}_{K_l}(x)}{V_{K_j}^{1/2} V_{K_l}^{1/2}}$. Similarly to the proof of Theorem 3.1, we define $\omega_{ni} = (\delta'\Sigma_n\delta)^{-1/2} \sum_{j=1}^M \delta_j \frac{\tilde{P}_{K_j}(x)' \tilde{P}_{K_j i} \varepsilon_i}{V_{K_j}^{1/2}}$. Observe that $E\omega_{ni} = 0$, and $\frac{1}{n}\sum_{i=1}^n E[\omega_{ni}^2] = 1$, and

$$||\sum_{j=1}^M \frac{\tilde{P}_{K_j}(x)' \tilde{P}_{K_j}}{V_{K_j}^{1/2}}||_\infty \lesssim \sum_{j=1}^M \zeta_{K_j} \lambda_{K_j} \lesssim \zeta_{K_M} \lambda_{K_M}$$

by Assumptions 3.2 and 3.5. Lindberg's condition can be verified similarly as in the proof of Theorem 3.1. Therefore, finite dimensional convergence holds by Lindberg-Feller CLT and

Slutzky's Theorem.

Next, we let $G(\cdot)$ be a strictly increasing continuous distribution function on $\mathbb{R}$, for example standard normal cdf $\Phi(\cdot)$. For any m,

$$G_{n,m} = G(T_n(K_m, \theta_0)) = G(t_n(m) + \nu_n(m) + o_p(1)).$$

If $|\nu(m)| < \infty$, then we have

$$G_{n,m} \xrightarrow{d} G(Z_m + \nu(m)) \tag{A.23}$$

by finite dimensional CLT under Assumptions 3.2, 3.5 and the continuous mapping theorem. If $\nu(m) = +\infty$,

$$G_{n,m} \xrightarrow{p} 1 \tag{A.24}$$

since $t_n(m) = O_p(1)$, and $G(x) \to 1$ as $x \to \infty$, and by CLT. Moreover, if $\nu(m) = -\infty$

$$G_{n,m} \xrightarrow{p} 0 \tag{A.25}$$

as $G(x) \to 0$ as $x \to -\infty$. Since (A.23), (A.24), and (A.25) holds jointly, following holds for any strictly increasing continuous distribution function on $\mathbb{R}$, $G(\cdot)$,

$$G_n \equiv (G_{n,1}, \cdots, G_{n,M})' \xrightarrow{d} G_\infty \equiv (G(Z_1 + \nu(1)), \cdots, G(Z_M + \nu(M)))' \tag{A.26}$$

where $G_{n,m} = G(T_n(K_m, \theta_0))$, and $G(Z_m + \nu(m))$ denotes $G(+\infty) = 1$ when $\nu(m) = +\infty$, and $G(-\infty) = 0$ when $\nu(m) = -\infty$.

Next, we define $G^{-1}(\cdot)$ as the inverse of $G(\cdot)$. For $t = (t_1, \cdots, t_M)' \in \mathbb{R}^{M-1}_{[\pm\infty]} \times \mathbb{R}$, define $G_{(M)}(t) \equiv (G(x_1), \cdots, G(x_M))' \in [0,1]^{M-1} \times (0,1)$. For $y = (y_1, \cdots, y_M)' \in (0,1]^{M-1} \times (0, 1)$, define $G^{-1}_{(M)}(y) \equiv (G^{-1}(y_1), \cdots, G^{-1}(y_M))' \in \mathbb{R}^{M-1}_{[\pm\infty]} \times \mathbb{R}$. Define also $S^*(y)$ for $y \in (0, 1]^{M-1} \times (0,1)$,

$$S^*(y) \equiv S(G^{-1}_{(M)}(y)). \tag{A.27}$$

Note that $S^*(y)$ is continuous at all $y \in (0,1]^{M-1} \times (0,1)$ since $S(t)$ is continuous at all

$t \in \mathbb{R}^{M-1}_{[\pm\infty]} \times \mathbb{R}$. Then, we have

$$\begin{aligned} S(T_n(\theta_0)) &= S(G^{-1}_{(M)}(G_n)) \\ &= S^*(G_n) \\ &\xrightarrow{d} S^*(G_\infty) \\ &= S(G^{-1}_{(M)}(G_\infty)) = S(Z+\nu) \end{aligned}$$

where the first equality holds by the definition of $G^{-1}_{(M)}(\cdot)$, the second equality uses the definition of S^*. Convergence in the third line holds by (A.26), and the fourth and fifth equality uses the definition of S^*.

Q.E.D.

A.3 Proof of Corollary 4.1

Proof. Under Assumptions 3.1, 3.2, 3.3 and $\sup_\pi |\nu(\pi)| < \infty$, we have $T^*_n(\pi, \theta_0) \Rightarrow \mathbb{T}(\pi) + \nu(\pi)$ by Theorem 3.1. Then, $\mathrm{Inf}\, T_n(\theta_0) = \inf_{K \in \mathcal{K}_n} |T_n(K, \theta_0)| = \inf_{\pi \in \Pi} |T^*_n(\pi, \theta_0)| \xrightarrow{d} \inf_\pi |\mathbb{T}(\pi) + \nu(\pi)|$ holds by continuous mapping theorem. In addition, if Assumption 3.4 holds, $\mathrm{Inf}\, T_n(\theta_0) \xrightarrow{d} \inf_\pi |\mathbb{T}(\pi)|$ by Theorem 3.1.

For the second part of Corollary, we first define $S(t) \equiv \inf_m |t_m|$ for $t = (t_1, \cdots, t_M) \in \mathbb{R}^{M-1}_{[\pm\infty]} \times \mathbb{R}$. Note that $S(t)$ is continuous at all $t \in \mathbb{R}^{M-1}_{[\pm\infty]} \times \mathbb{R}$ under Assumption 3.5 (especially, assumption of at least one $|\nu_m| = O(1)$) by restricting the domain of functions appropriately. Then, we have

$$\mathrm{Inf}\, T_n(\theta_0) = S(T_n(\theta_0)) \xrightarrow{d} S(Z+\nu) = \inf_m |Z_m + \nu_m| \tag{A.28}$$

by Theorem 3.2. If $|\nu_m| = +\infty$, corresponding elements of $|Z_m + \nu_m| = +\infty$ by construction. This completes the proof of Corollary 4.1.

Q.E.D.

A.4 Proof of Corollary 4.2

Proof. We first provide (4.3) in Corollary 4.2.1. Under Assumptions 3.1-3.4, we have shown that $\mathrm{Inf}\, T_n(\theta_0) \xrightarrow{d} \xi_{\inf} = \inf_{\pi \in [\underline{\pi}, 1]} |\mathbb{T}(\pi)|$ in Corollary 4.1.1. Therefore,

$$\limsup_{n\to\infty} P(\mathrm{Inf}\, T_n(\theta_0) > c^{\inf}_{1-\alpha}) = \lim_{n\to\infty} P(\mathrm{Inf}\, T_n(\theta_0) > c^{\inf}_{1-\alpha}) = P(\xi_{\inf} > c^{\inf}_{1-\alpha}) = \alpha$$

where the first equality holds under subsequence $\{u_n\}$ of $\{n\}$ by the definition of lim sup, the second equality uses the Corollary 4.1.1 and the definition of $c_{1-\alpha}^{\inf}$ in (4.2). Moreover,

$$\limsup_{n\to\infty} P(\text{Inf}\, T_n(\theta_0) > z_{1-\alpha/2}) = P(\xi_{\inf} > z_{1-\alpha/2}) \leq P(|\mathbb{T}(\pi)| > z_{1-\alpha/2}) = \alpha$$

where the inequality uses $\xi_{\inf} = \inf_{\pi\in[\underline{\pi},1]} |\mathbb{T}(\pi)| \leq |\mathbb{T}(\pi)|$ and $\mathbb{T}(\pi) \stackrel{d}{=} N(0,1)$ for any single π.

Next, we prove Corollary 4.2.2. Under Assumptions 3.1-3.3 and $\sup_\pi |\nu(\pi)| < \infty$, we have $\text{Inf}\, T_n(\theta_0) \xrightarrow{d} \inf_{\pi\in[\underline{\pi},1]} |\mathbb{T}(\pi) + \nu(\pi)|$ with asymptotic bias $\nu(\pi)$. We have

$$\begin{aligned}
\limsup_{n\to\infty} P(\text{Inf}\, T_n(\theta_0) > c_{1-\alpha}^{\inf}) &= P(\inf_{\pi\in[\underline{\pi},1]} |\mathbb{T}(\pi) + \nu(\pi)| > c_{1-\alpha}^{\inf}) \\
&\leq \inf_\pi P(|\mathbb{T}(\pi) + \nu(\pi)| > c_{1-\alpha}^{\inf}) \\
&= \inf_\pi [1 - (P(Z \leq c_{1-\alpha}^{\inf} - |\nu(\pi)|) - P(Z \leq -c_{1-\alpha}^{\inf} - |\nu(\pi)|))] \\
&= \inf_\pi F(c_{1-\alpha}^{\inf}, |\nu(\pi)|) = F(c_{1-\alpha}^{\inf}, \inf_\pi |\nu(\pi)|)
\end{aligned}$$

where the first inequality uses $\inf_{\pi\in[\underline{\pi},1]} |\mathbb{T}(\pi) + \nu(\pi)| \leq |\mathbb{T}(\pi) + \nu(\pi)|$ for all π, the second equality uses $\mathbb{T}(\pi) \stackrel{d}{=} Z \sim N(0,1)$ and the definition of $F(\cdot)$. Finally, the last equality holds since $F(c, |\nu|)$ is monotone increasing function of $|\nu|$. Similarly,

$$\limsup_{n\to\infty} P(\text{Inf}\, T_n(\theta_0) > z_{1-\alpha/2}) = P(\inf_{\pi\in[\underline{\pi},1]} |\mathbb{T}(\pi) + \nu(\pi)| > z_{1-\alpha/2}) \leq F(z_{1-\alpha/2}, \inf_\pi |\nu(\pi)|).$$

Corollary 4.2.3 can be similarly derived with $\inf_m |\nu(m)| = 0$ under Assumption 3.5 and using the fact that $F(z_{1-\alpha/2}, 0) = \alpha$. This completes the proof. (If we further assume $\Sigma = I_M$ in Theorem 3.2, then $\limsup_{n\to\infty} P(\text{Inf}\, T_n(\theta_0) > c) = \prod_{m=M-M_1+1}^{M} F(c, |\nu(m)|)$ holds for any $0 < c < \infty$, by the asymptotic independence of Z_m, $m = 1, ..., M$ and when $|\nu(m)| = \infty$ for $m = 1, \cdots, M - M_1$ since $F(c, |\nu(m)|) = 1$ for $|\nu(m)| = \infty$.) *Q.E.D.*

A.5 Proof of Corollary 4.3

Proof. Under Assumptions 3.2, 3.4, 4.1, and 4.2, following finite dimensional convergence holds by Theorem 3.1,

$$T_n(\theta) = (T_n(K_1, \theta_0), \cdots, T_n(K_M, \theta_0))' \xrightarrow{d} Z = (Z_1, \cdots, Z_M)', \quad Z \sim N(0, \Sigma). \tag{A.29}$$

Under Assumptions 3.2, 3.4, 3.5, and 4.2, above also holds by Theorem 3.2. Note that $T_{n,\widehat{V}}(K,\theta) = \frac{\sqrt{n}(\widehat{\theta}_K - \theta_0)}{\widehat{V}_K^{1/2}} = \frac{V_K^{1/2}}{\widehat{V}_K^{1/2}} T_n(K,\theta)$. Then following holds for $A \equiv diag\{\frac{V_{K_1}^{1/2}}{\widehat{V}_{K_1}^{1/2}}, \cdots, \frac{V_{K_M}^{1/2}}{\widehat{V}_{K_M}^{1/2}}\}$,

$$(T_{n,\widehat{V}}(K_1,\theta_0), \cdots, T_{n,\widehat{V}}(K_M,\theta_0))' = AT_n(\theta) \xrightarrow{d} Z \tag{A.30}$$

by Assumption 4.3 and Slutzky Theorem and $A \xrightarrow{p} I_M$

Next consider $\widehat{c}_{1-\alpha}^{\inf}$ which is $(1-\alpha)$ quantile of $\inf_{m=1,\cdots,M} |Z_{m,\widehat{\Sigma}}|$ defined in (4.10),

$$\widehat{c}_{1-\alpha}^{\inf} = \inf\{x \in \mathbb{R} : P(\inf_{m=1,\cdots,M} |Z_{m,\widehat{\Sigma}}| \leq x) \geq 1-\alpha\}$$

where $Z_{\widehat{\Sigma}} = (Z_{1,\widehat{\Sigma}}, \cdots, Z_{M,\widehat{\Sigma}})' \sim N(0,\widehat{\Sigma})$, $\widehat{\Sigma}_{jj} = 1, \widehat{\Sigma}_{jl} = \widehat{V}_{K_j}^{1/2}/\widehat{V}_{K_l}^{1/2}$. Note that for any $j < l$,

$$\widehat{\Sigma}_{jl} = \frac{\widehat{V}_{K_j}^{1/2}}{\widehat{V}_{K_l}^{1/2}} = \frac{\widehat{V}_{K_j}^{1/2}}{V_{K_j}^{1/2}} \frac{V_{K_j}^{1/2}}{V_{K_l}^{1/2}} \frac{V_{K_l}^{1/2}}{\widehat{V}_{K_l}^{1/2}} \xrightarrow{p} \Sigma_{jl} \tag{A.31}$$

by Assumption 4.3. Therefore, $\widehat{\Sigma} \xrightarrow{p} \Sigma$, $Z_{\widehat{\Sigma}} \xrightarrow{d} Z_\Sigma$, and $\inf_{m=1,\cdots,M} |Z_{m,\widehat{\Sigma}}| \xrightarrow{d} \inf_{m=1,\cdots,M} |Z_{m,\Sigma}|$ hold. Thus, $\widehat{c}_{1-\alpha}^{\inf} \xrightarrow{p} c_{1-\alpha}^{\inf}$. *Q.E.D.*

A.6 Proof of Corollary 5.1

Proof. We first show Corollary 5.1.1. Note that $\text{Inf}\, T_n(\theta_0) = \inf_{K \in \mathcal{K}_n} |T_{n,\widehat{V}}(K,\theta)| \xrightarrow{d} \inf_m |Z_m|$ by Corollary 4.3. We have

$$\begin{aligned}\liminf_{n\to\infty} P(\theta_0 \in CI_{\texttt{inf}}^{\texttt{Robust}}) &= \liminf_{n\to\infty} P(\text{Inf}\, T_n(\theta_0) \leq c_{1-\alpha}^{inf} + o_p(1)) \\ &= P(\inf_m |Z_m| \leq c_{1-\alpha}^{inf}) = 1-\alpha\end{aligned}$$

where the first and the second equality holds by Corollary 4.3 and Corollary 4.1.1 under Assumptions 3.2, 3.4, 4.1, 4.2, and 4.3. Similarly,

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{\texttt{inf}}) = P(\inf_m |Z_m| \leq z_{1-\alpha/2}) \geq P(|Z_m| \leq z_{1-\alpha/2}) = 1-\alpha. \tag{A.32}$$

Corollary 5.1.2 and 5.1.3 can be similarly derived from Corollary 4.2.2 and 4.2.3, respectively. *Q.E.D.*

A.7 Proof of Corollary 6.1

Proof. Similar to the proof of Corollary 4.3, we can verify $\sup_{m=1,\cdots,M} |Z_{m,\widehat{\Sigma}}| \xrightarrow{d} \sup_{m=1,\cdots,M} |Z_{m,\Sigma}|$, $\widehat{c}^{\sup}_{1-\alpha} \xrightarrow{p} c^{\sup}_{1-\alpha}$, and $\operatorname{Sup} T_n(\theta_0) = \sup_m |T_{n,\widehat{V}}(K_m, \theta_0)| \xrightarrow{d} \sup_m |Z_{m,\Sigma}|$ either under Assumptions 3.2, 3.4, 4.1, 4.2, and 4.3 or under Assumptions 3.2, 3.4, 3.5, 4.2, and 4.3. Therefore, we have

$$\liminf_{n\to\infty} P(\theta_0 \in CI^{\texttt{Robust}}_{\texttt{pms}}) = \liminf_{n\to\infty} P(|T_{n,\widehat{V}}(\widehat{K}, \theta_0)| \leq \widehat{c}^{sup}_{1-\alpha}) \tag{A.33}$$

$$\geq \liminf_{n\to\infty} P(\operatorname{Sup} T_n(\theta_0) \leq \widehat{c}^{sup}_{1-\alpha}) \tag{A.34}$$

$$= P(\sup_m |Z_{m,\Sigma}| \leq c^{sup}_{1-\alpha}) = 1 - \alpha \tag{A.35}$$

where the first inequality uses $|T_{n,\widehat{V}}(\widehat{K}, \theta_0)| \leq \operatorname{Sup} T_n(\theta_0)$ for any $\widehat{K} \in \mathcal{K}_n$. *Q.E.D.*

A.8 Proof of Theorem 7.1

Proof. Conditional on $X = [x_1, \cdots, x_n]'$, following decomposition holds for any single sequence $K \in \mathcal{K}_n$

$$\sqrt{n}(\widehat{\theta}_K - \theta_0) = \widehat{\Gamma}_K^{-1} S_K,$$
$$\widehat{\Gamma}_K = \frac{1}{n}(W' M_K W), \quad S_K = \frac{1}{\sqrt{n}} W' M_K (g + \varepsilon)$$

where $g = [g_1, \cdots, g_n]', g_i = g_0(x_i), g_w = [g_{w1}, \cdots, g_{wn}]', g_{wi} = g_{w0}(x_i) = E[w_i|x_i], v = [v_1, \cdots, v_n]$. All remaining proofs contain conditional expectations (conditioning on X) hold almost surely (a.s.).

Under Assumption 7.1 and conditional homoskedastic error terms, $E[v_i^2|x_i] = E[v_i^2]$,

$$\widehat{\Gamma}_K = \Gamma_K + o_p(1), \quad \Gamma_K = (1 - K/n) E[v_i^2] \tag{A.36}$$

by Lemma 1 of Cattaneo, Jansson, and Newey (2015a). Moreover,

$$S_K = \frac{1}{\sqrt{n}} v' M_K \varepsilon + \frac{1}{\sqrt{n}} g_w' M_K g + \frac{1}{\sqrt{n}} (v' M_K g + g_w' M_K \varepsilon) \tag{A.37}$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} M_{K,ii} v_i \varepsilon_i - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1, j<i}^{n} P_{K,ij} (v_i \varepsilon_j + v_j \varepsilon_i) + o_p(1) \tag{A.38}$$

since $M_{K,ij} = -P_{K,ij}$ for $j < i$, $\frac{1}{\sqrt{n}} g_w' M_K g = O_p(\sqrt{n} \bar{K}^{-\gamma_g - \gamma_{g_w}}) = o_p(1)$, $\frac{1}{\sqrt{n}}(v' M_K g + g_w' M_K \varepsilon) = O_p(\bar{K}^{-\gamma_g} + \bar{K}^{-\gamma_{g_w}}) = o_p(1)$ by Lemma 2 of Cattaneo, Jansson and Newey (2015a)

under Assumption 7.1. Under conditional homoskedastic error $E[\varepsilon_i^2|w_i, x_i] = \sigma_\varepsilon^2$ following holds

$$T_n(K, \theta_0) = \sqrt{n} V_K^{-1/2}(\widehat{\theta}_K - \theta_0) = V_K^{-1/2} \Gamma_K^{-1} \frac{1}{\sqrt{n}} v' M_K \varepsilon + o_p(1) \xrightarrow{d} N(0, 1)$$

by Theorem 1 of Cattaneo, Jansson and Newey (2015a) which follows from Lemma A2 in Chao, Swanson, Hausman, Newey and Woutersen (2012).

For simplicity, here we only show the joint convergence of bivariate t-statistics, but the proof can be easily extended to multivariate case. For any $K_1 < K_2$ in $\mathcal{K}_n$, we show

$$\delta_1 T_n(K_1, \theta_0) + \delta_2 T_n(K_2, \theta_0) \xrightarrow{d} N(0, (\delta_1^2 + \delta_2^2 + 2\delta_1 \delta_2 v_{12})), \quad \forall (\delta_1, \delta_2) \in \mathbb{R}^2 \tag{A.39}$$

where $v_{12} = \lim_{n\to\infty} V_{K_1}^{1/2} / V_{K_2}^{1/2}$. We closely follows the proof of Lemma A2 in Chao et al. (2012). Define $Y_n, Y_{1,n}$ and $Y_{2,n}$ as follows

$$Y_n = \delta_1 Y_{1,n} + \delta_2 Y_{2,n}, \tag{A.40}$$

$$Y_{1,n} = \omega_{1,1n} + \sum_{i=2}^{n} y_{1,in}, \quad y_{1,in} = \omega_{1,in} + \bar{y}_{1,in}, \tag{A.41}$$

$$Y_{2,n} = \omega_{2,1n} + \sum_{i=2}^{n} y_{2,in}, \quad y_{2,in} = \omega_{2,in} + \bar{y}_{2,in}, \tag{A.42}$$

where $\omega_{1,in} = V_{K_1}^{-1/2} \Gamma_{K_1}^{-1} M_{K_1,ii} / \sqrt{n}$, $\bar{y}_{1,in} = \sum_{j<i} (u_{1,j} P_{K_1,ij} \varepsilon_i + u_{1,i} P_{K_1,ij} \varepsilon_j) / \sqrt{n}$, $u_{1,i} = V_{K_1}^{-1/2} \Gamma_{K_1}^{-1} v_i$ and $\omega_{2,in}, \bar{y}_{2,in}$ are similarly defined with appropriate terms $P_{K_2}, V_{K_2}, \Gamma_{K_2}$ with K_2. Similar to the proof of Lemma A2 in Chao et al. (2012), $\omega_{1,1n} = o_p(1), \omega_{2,1n} = o_p(1)$. Thus, we only need to show that following holds conditional on X with probability one

$$\sum_{i=2}^{n} (\delta_1 y_{1,in} + \delta_2 y_{2,in}) \xrightarrow{d} N(0, \delta_1^2 + \delta_2^2 + 2\delta_1 \delta_2 v_{12}). \tag{A.43}$$

It remains to provide Lindeberg-Feller condition.

$$\begin{aligned} E[(\sum_{i=2}^{n} \delta_1 y_{1,in} + \delta_2 y_{2,in})^2 | X] = \delta_1^2 E[(\sum_{i=2}^{n} y_{1,in})^2 | X] + \delta_2^2 E[(\sum_{i=2}^{n} y_{2,in})^2 | X] \\ + 2\delta_1 \delta_2 E[\sum_{i=2}^{n} \sum_{j=2}^{n} y_{1,in} y_{2,in} | X], \end{aligned} \tag{A.44}$$

where the first and second terms in (A.44) goes to δ_1^2, δ_2^2 a.s., respectively, as in the proof of Lemma A.2 in Chao et al. (2012). Note that $E[\omega_{1,in} \bar{y}_{2,in} | X] = 0, E[\omega_{2,in} \bar{y}_{1,in} | X] = 0$, and

$E[\omega_{1,1n}\omega_{2,in}|X] = 0, E[\omega_{2,1n}\omega_{1,in}|X] = 0$ for any $i > 1$. Followings are the key calculations for the asymptotic variance of leading terms in Y_n.

$$E[Y_{1,n}Y_{2,n}|X] = \frac{1}{n}V_{K_1}^{-1/2}\Gamma_{K_1}^{-1}E[v'M_{K_1}\varepsilon\varepsilon'M_{K_2}v|X]\Gamma_{K_2}^{-1}V_{K_2}^{-1/2} \tag{A.45}$$

$$= \frac{1}{n}V_{K_1}^{-1/2}\Gamma_{K_1}^{-1}\sigma_\varepsilon^2 E[v'M_{K_2}v|X]\Gamma_{K_2}^{-1}V_{K_2}^{-1/2} \tag{A.46}$$

$$= V_{K_1}^{-1/2}\Gamma_{K_1}^{-1}\sigma_\varepsilon^2\Gamma_{K_2}\Gamma_{K_2}^{-1}V_{K_2}^{-1/2} \tag{A.47}$$

$$= V_{K_1}^{1/2}/V_{K_2}^{1/2} \tag{A.48}$$

where the second equality uses conditional homoskedasticity $E[\varepsilon\varepsilon'|X,W] = \sigma_\varepsilon^2 I$ and $M_{K_1}M_{K_2} = M_{K_2}$, the third equality uses $tr(M_{K_2}) = n - K_2$ and $E[v^2|X] = E[v^2]$, and the last equality uses $V_{K_1} = \sigma_\varepsilon^2\Gamma_{K_1}^{-1}$. Therefore, we calculate components of last terms in (A.44) as follows

$$E[\sum_{i=2}^n\sum_{j=2}^n y_{1,in}y_{2,in}|X] = E[Y_{1,n}Y_{2,n}|X] - \sum_{i=2}^n E[\omega_{1,1n}y_{2,in}|X]$$

$$- \sum_{i=2}^n E[\omega_{2,1n}y_{1,in}|X] - E[\omega_{1,1n}\omega_{2,1n}|X] \tag{A.49}$$

$$= V_{K_1}^{1/2}/V_{K_2}^{1/2} - E[\omega_{1,1n}\omega_{2,1n}|X] \to v_{12} \quad a.s. \tag{A.50}$$

As in the proof of Lemma A.2 of Chao et al. (2012), we have

$$\sum_{i=2}^n E[(\delta_1 y_{1,in} + \delta_2 y_{2,in})^4|X] \lesssim \sum_{i=2}^n E[(y_{1,in})^4|X] + \sum_{i=2}^n E[(y_{2,in})^4|X] \to 0 \quad a.s. \tag{A.51}$$

Thus, by similar arguments following the proof of Lemma A.2 in Chao et al. (2012), we can apply the martingale central limit theorem. Then, by Slutzky theorem, joint convergence holds with the claimed covariance.

By Theorem 2 in Cattaneo, Jansson, and Newey (2015a), Assumption 4.3 holds with the following variance estimator for V_K

$$\widehat{V}_K = s^2\widehat{\Gamma}_K^{-1}, \quad s^2 = \frac{1}{n-1-K}\sum_{i=1}^n \widehat{\varepsilon}_i^2, \quad \widehat{\varepsilon}_i^2 = \sum_{j=1}^n M_{K,ij}(y_j - \widehat{\theta}_K w_j). \tag{A.52}$$

Then, we can show the coverage results using similar arguments to those used in the proof of Corollary 5.1 and 6.1. This completes the proof. *Q.E.D.*

B Figures and Tables

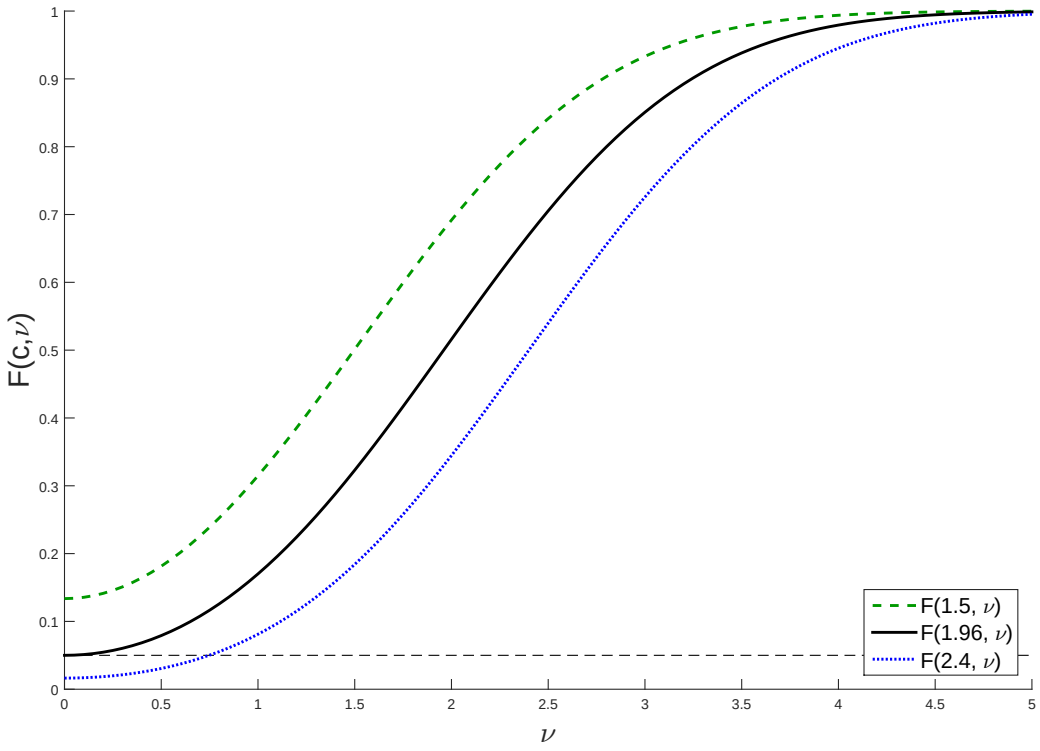


Figure 1: Plots of $F(c, \nu)$ as a function of ν for c = 1.5, 1.96, 2.4, where $F(c, |\nu|) = 1 - \Phi(c - |\nu|) + \Phi(-c - |\nu|)$ with the standard normal cumulative distribution function $\Phi(\cdot)$.

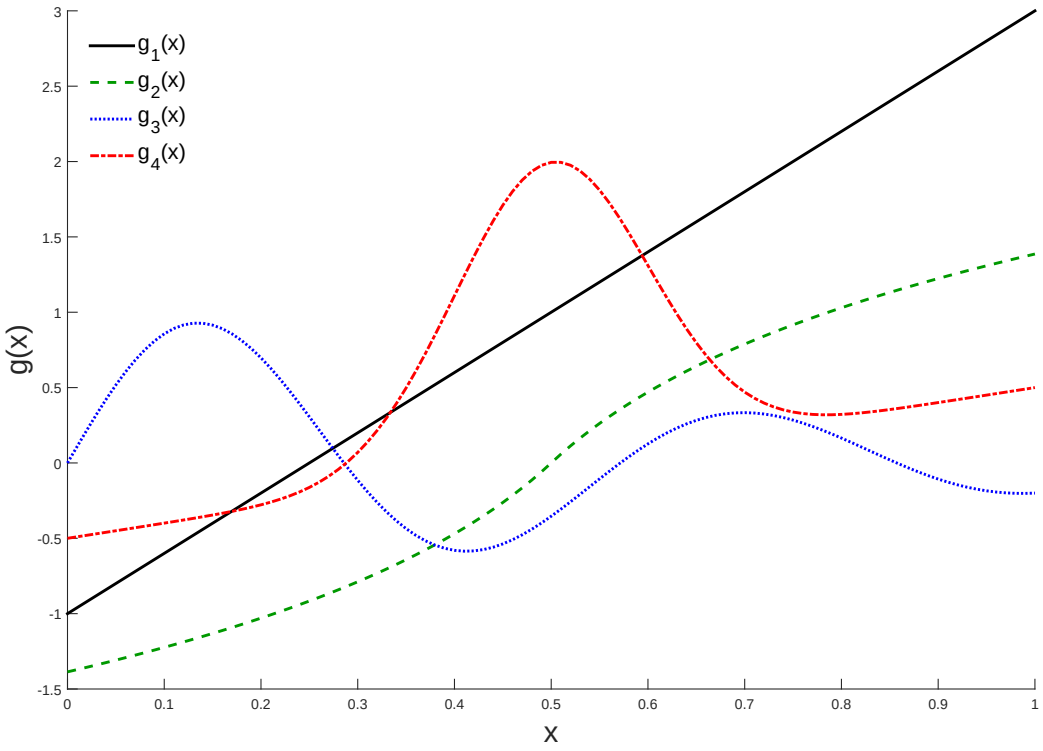
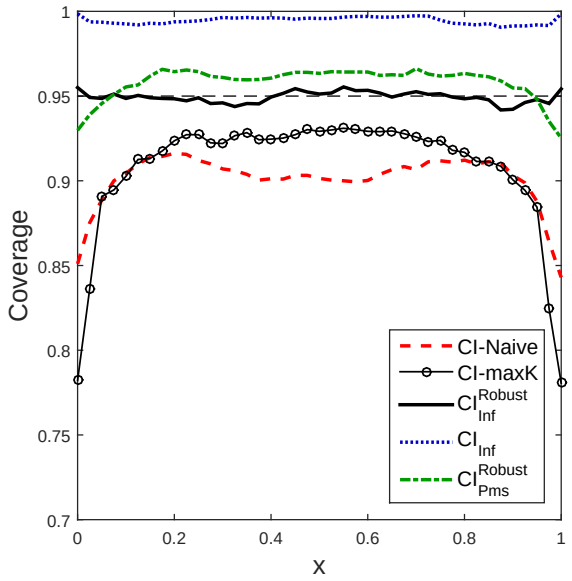


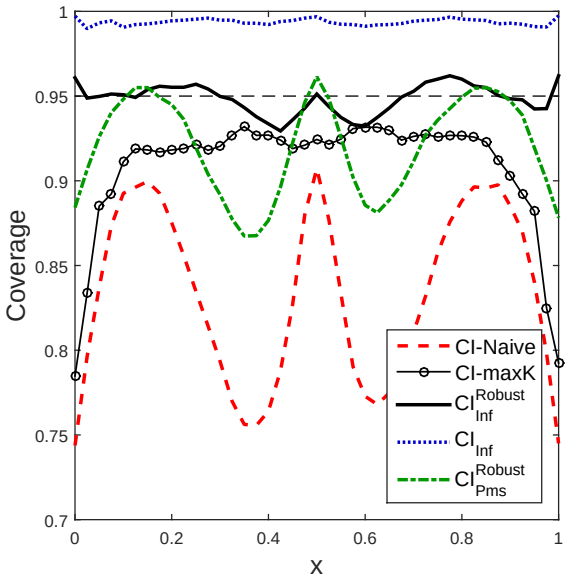
Figure 2: Different functions of $g(x)$ used in simulations (Section 8). Solid lines (Black) are $g_1(x) = 4x - 1$; Dashed lines (Green) are $g_2(x) = \ln(|6x - 3| + 1)sgn(x - 1/2)$; Dotted lines (Blue) are $g_3(x) = \sin(7\pi x/2)/[1 + 2x^2(sgn(x) + 1)]$; and Dash-dot lines (Red) are $g_4(x) = x - 1/2 + 5\phi(10(x - 1/2))$, where $\phi(\cdot)$ is the standard normal pdf.

Figure 3: Coverage - Polynomials
Nominal 95% Coverage of Various CIs for $g(x)$:
(1) $CI^{\texttt{Naive}}_{\texttt{pms}}$ with $\widehat{K}_{\texttt{cv}}$ (2) $CI_{\texttt{maxK}}$ with $\bar{K}$ (3) $CI^{\texttt{Robust}}_{\texttt{inf}}$ (4) $CI_{\texttt{inf}}$ (5) $CI^{\texttt{Robust}}_{\texttt{pms}}$ with $\widehat{K}_{\texttt{cv}}$

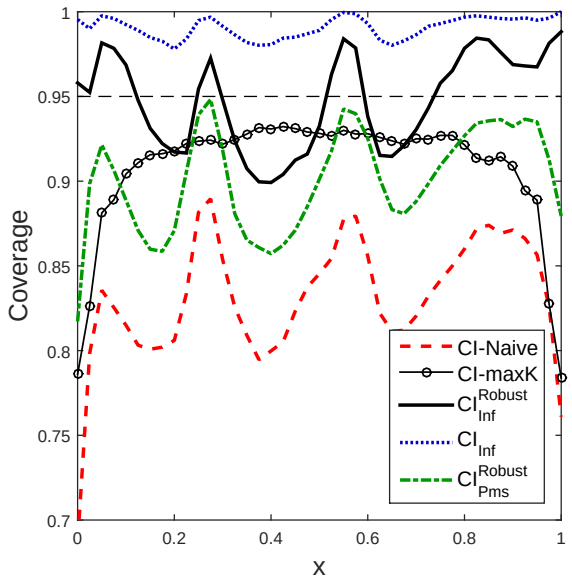
(a) $g_1(x) = 4x - 1$



(b) $g_2(x) = \ln(|6x - 3| + 1)sgn(x - 1/2)$



(c) $g_3(x) = \frac{\sin(7\pi x/2)}{1+2x^2(sgn(x)+1)}$



(d) $g_4(x) = x - 1/2 + 5\phi(10(x - 1/2))$

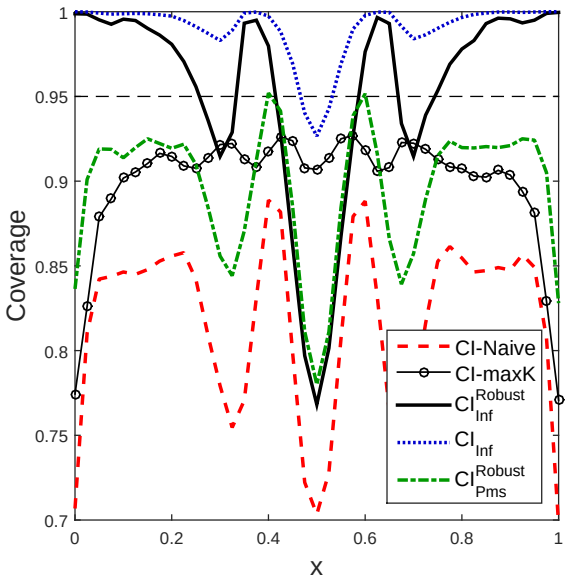
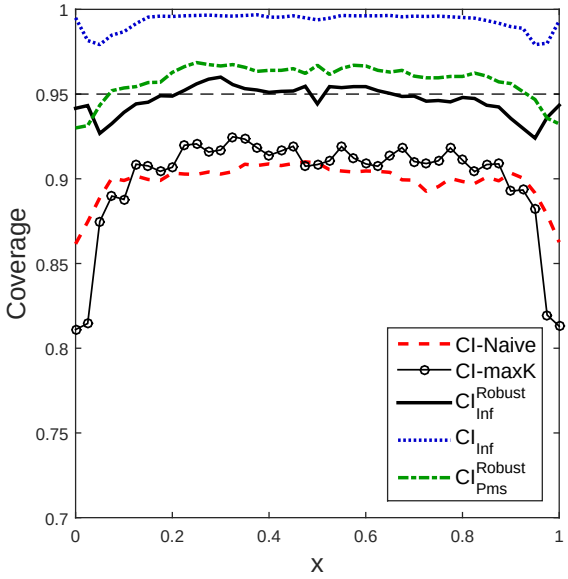
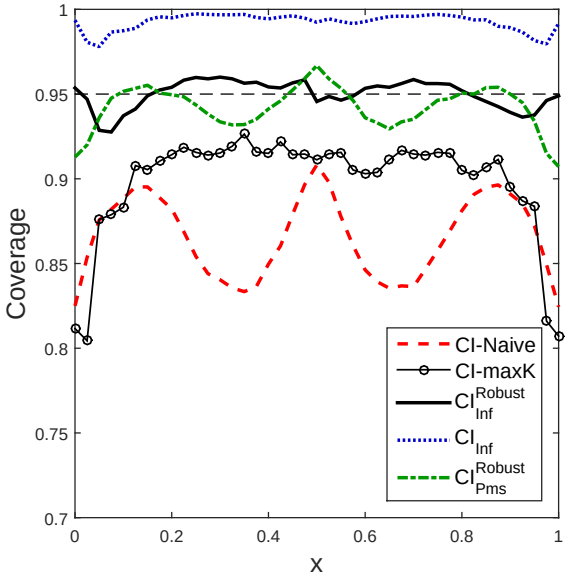


Figure 4: Coverage - Splines
Nominal 95% Coverage of Various CIs for $g(x)$:
(1) $CI^{\texttt{Naive}}_{\texttt{pms}}$ with $\widehat{K}_{\texttt{cv}}$ (2) $CI_{\texttt{maxK}}$ with $\bar{K}$ (3) $CI^{\texttt{Robust}}_{\texttt{inf}}$ (4) $CI_{\texttt{inf}}$ (5) $CI^{\texttt{Robust}}_{\texttt{pms}}$ with $\widehat{K}_{\texttt{cv}}$

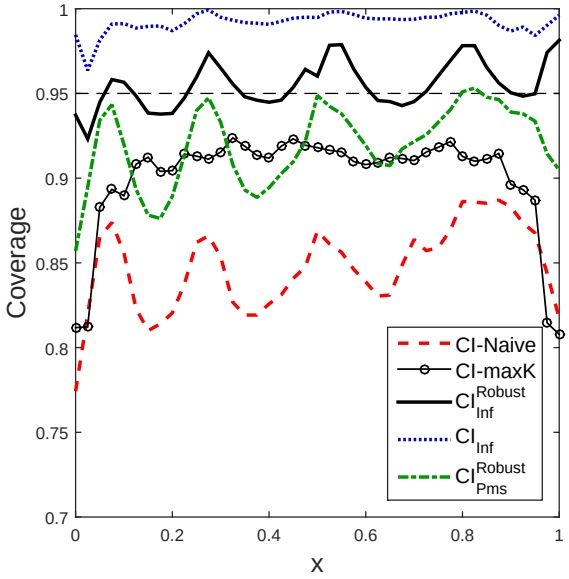
(a) $g_1(x) = 4x - 1$



(b) $g_2(x) = \ln(|6x - 3| + 1)sgn(x - 1/2)$



(c) $g_3(x) = \frac{\sin(7\pi x/2)}{1+2x^2(sgn(x)+1)}$



(d) $g_4(x) = x - 1/2 + 5\phi(10(x - 1/2))$

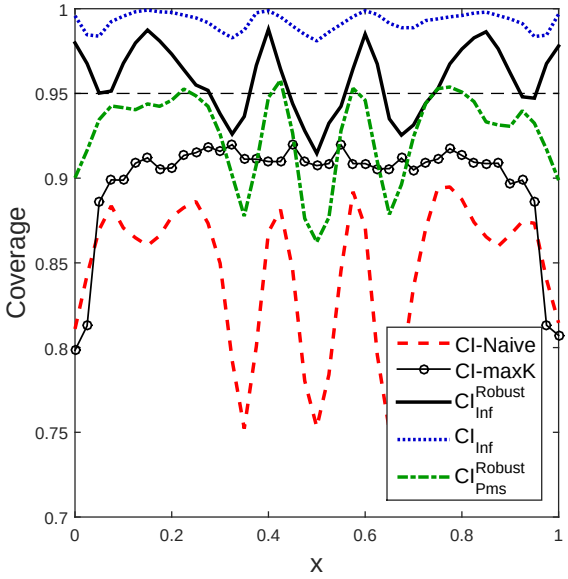
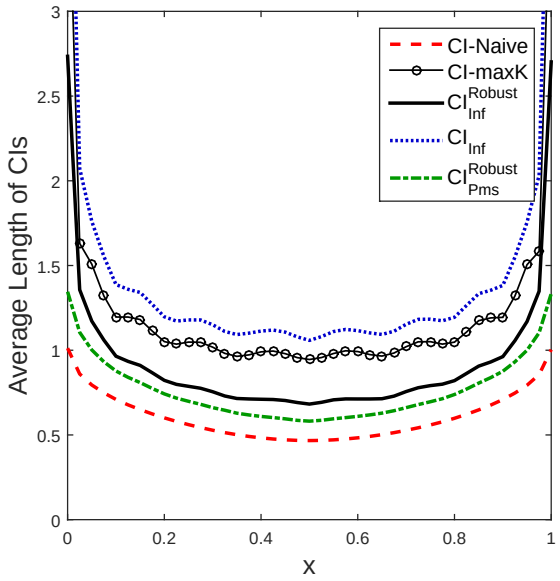
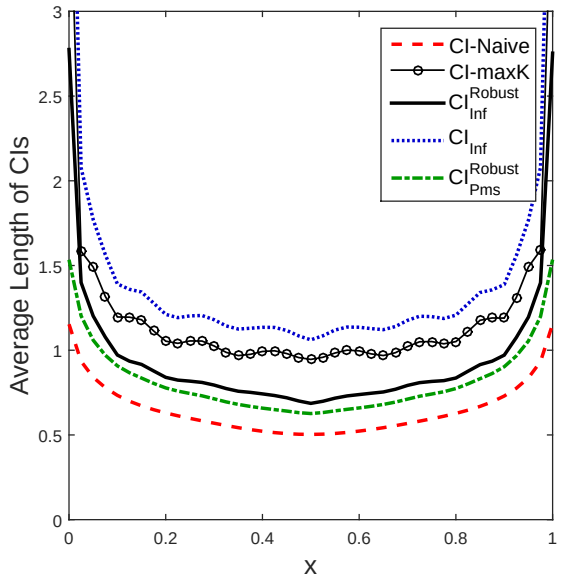


Figure 5: Length of CIs - Polynomials
Average lengths of nominal 95% CIs for $g(x)$:
(1) $CI^{\texttt{Naive}}_{\texttt{pms}}$ with $\widehat{K}_{\texttt{cv}}$ (2) $CI_{\texttt{maxK}}$ with $\bar{K}$ (3) $CI^{\texttt{Robust}}_{\texttt{inf}}$ (4) $CI_{\texttt{inf}}$ (5) $CI^{\texttt{Robust}}_{\texttt{pms}}$ with $\widehat{K}_{\texttt{cv}}$

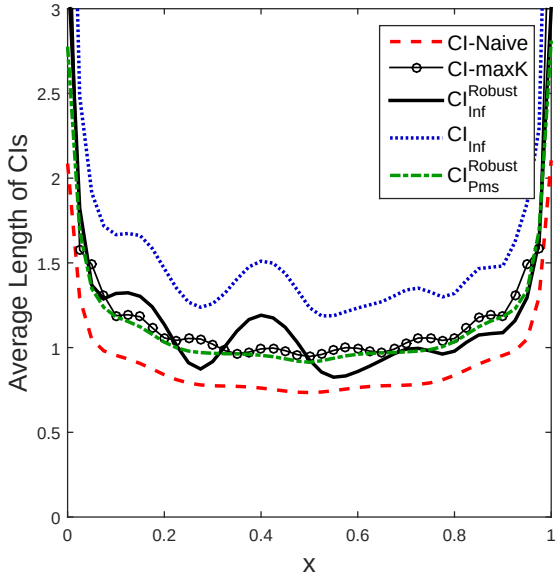
(a) $g_1(x) = 4x - 1$



(b) $g_2(x) = \ln(|6x - 3| + 1)sgn(x - 1/2)$



(c) $g_3(x) = \frac{\sin(7\pi x/2)}{1+2x^2(sgn(x)+1)}$



(d) $g_4(x) = x - 1/2 + 5\phi(10(x - 1/2))$

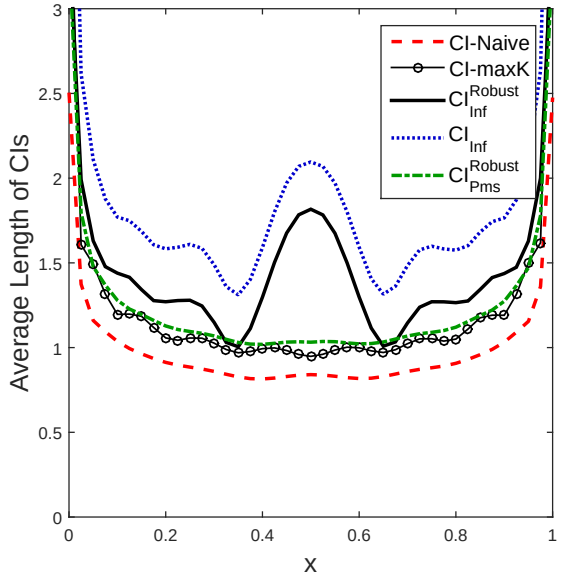
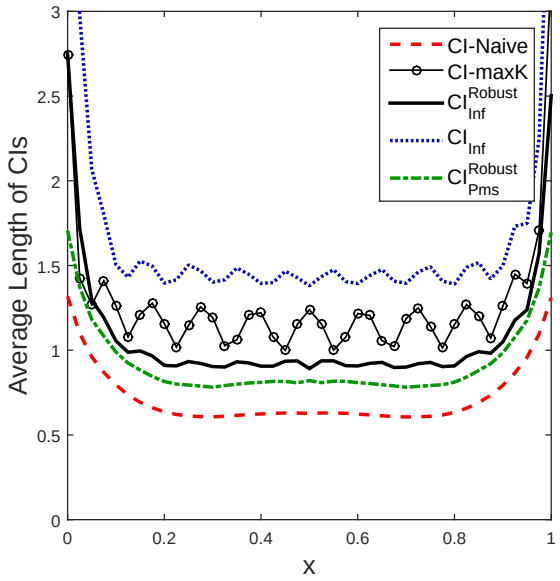
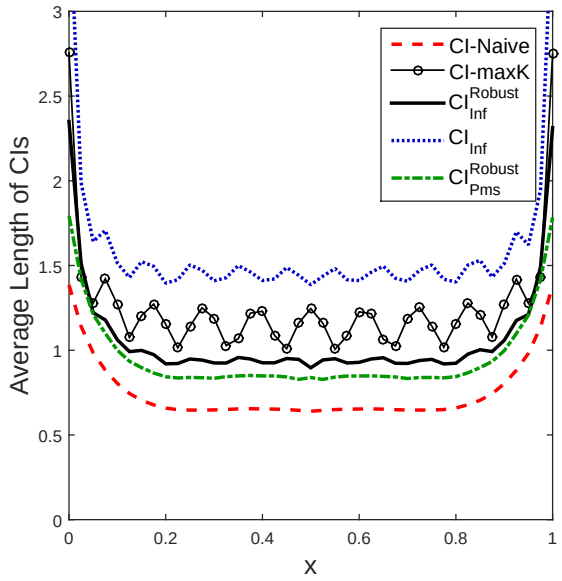


Figure 6: Length of CIs - Splines
Average lengths of nominal 95% CIs for $g(x)$:
(1) $CI^{\texttt{Naive}}_{\texttt{pms}}$ with $\widehat{K}_{\texttt{cv}}$ (2) $CI_{\texttt{maxK}}$ with $\bar{K}$ (3) $CI^{\texttt{Robust}}_{\texttt{inf}}$ (4) $CI_{\texttt{inf}}$ (5) $CI^{\texttt{Robust}}_{\texttt{pms}}$ with $\widehat{K}_{\texttt{cv}}$

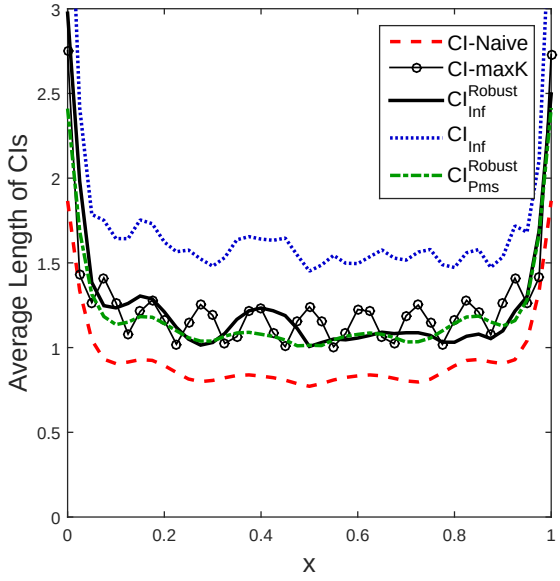
(a) $g_1(x) = 4x - 1$



(b) $g_2(x) = \ln(|6x - 3| + 1)sgn(x - 1/2)$



(c) $g_3(x) = \frac{\sin(7\pi x/2)}{1+2x^2(sgn(x)+1)}$



(d) $g_4(x) = x - 1/2 + 5\phi(10(x - 1/2))$

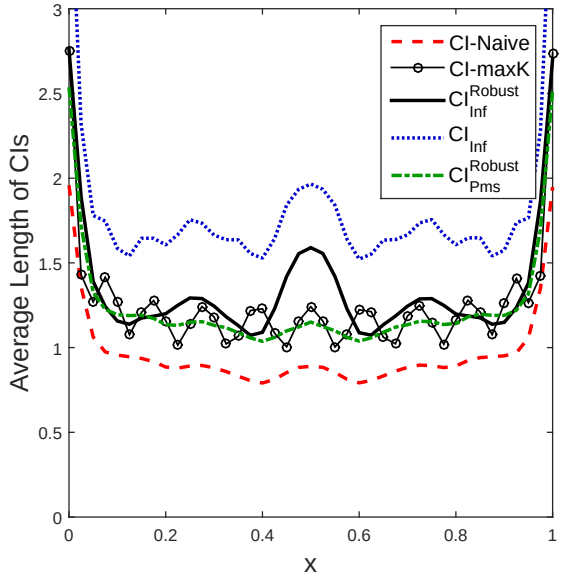
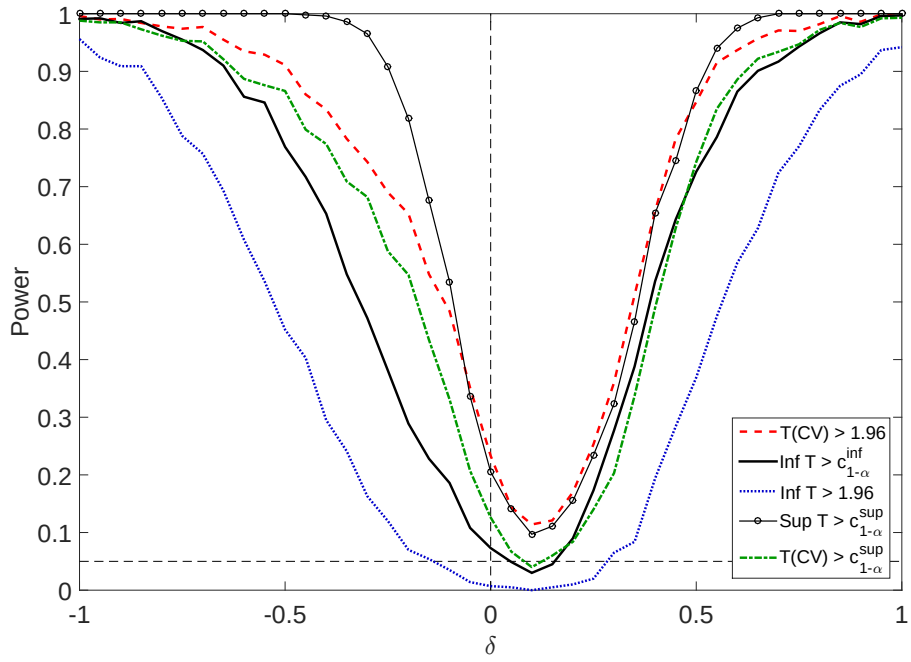


Figure 7: Power function against fixed alternatives. Design 2 : $g_2(x) = \ln(|6x-3|+1)sgn(x-1/2)$. $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_0 + \delta$, where $\theta_0 = g_2(x)$ at $x = 0.4$ for figure (a) and $x = 0.5$ for figure (b). Using Polynomials.

(a) $\theta_0 = g_2(x), x = 0.4$



(b) $\theta_0 = g_2(x), x = 0.5$

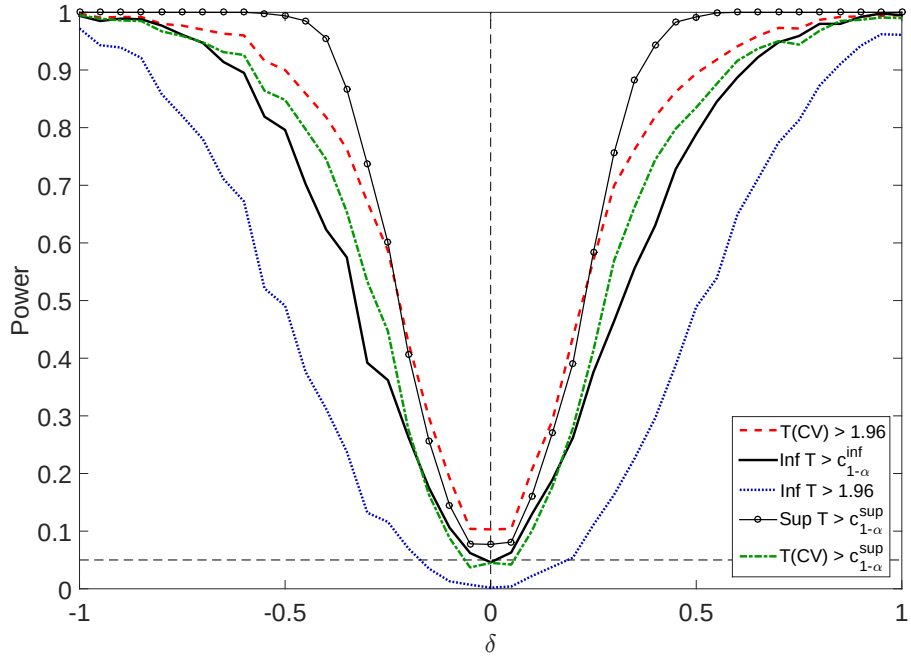


Table 1: Nonparametric Wage Elasticity of Hours of Work
Estimates in Blomquist and Newey (Table 1, 2002). Wage elasticity evaluated at the mean wage and income.

Additional Terms[1]	CV^2	$\widehat{E}_w$	$SE_{\widehat{E}_w}$	$CI_{\widehat{E}_w}$
$1, y_J, w_J$	0.00472	0.0372	0.0104	[0.0168, 0.0576]
$\Delta y \Delta w$	0.0313	0.0761	0.0128	[0.0510, 0.1012]
$\ell \Delta y$	0.0305	0.0760	0.0127	[0.0511, 0.1009]
y_J^2, w_J^2	0.0323	0.0763	0.0129	[0.0510, 0.1016]
$\Delta y^2, \Delta w^2$	0.0369	0.0543	0.0151	[0.0247, 0.0839]
$y_J w_J$	0.0364	0.0659	0.0197	[0.0273, 0.1045]
$\Delta y w$	0.0350	0.0628	0.0223	[0.0191, 0.1065]
$\ell^2 \Delta y$	0.0364	0.0636	0.0223	[0.0199, 0.1073]
y_J^3, w_J^3	0.0331	0.0845	0.0275	[0.0306, 0.1384]
$\ell \Delta y^2, \ell \Delta w^2, \ell \Delta y w$	0.0263	0.0775	0.0286	[0.0214, 0.1336]
$y_J^2 w_J, y_J w_J^2$	0.0252	0.0714	0.0289	[0.0148, 0.1280]
MLE estimates		0.123	0.0137	
Critical values: $\widehat{c}^{\inf}_{1-\alpha} = 0.9668$, $\widehat{c}^{\sup}_{1-\alpha} = 2.4764$				
Test $H_0 : E_w = 0$, $\quad \inf T_n(\theta_0) = 2.4706 > \widehat{c}^{\inf}_{1-\alpha}$				
$CI^{\texttt{Robust}}_{\texttt{inf}} = [0.0271, 0.1111]$ $CI_{\texttt{inf}} = [0.0148, 0.1384]$, $\quad CI^{\texttt{Robust}}_{\texttt{pms}} = [0.0169, 0.0916]$				

[1] y : non-labor income, w : marginal wage rates, ℓ: the end point of the segment in a piecewise linear budget set. $\ell^m \Delta y^p w^q$ denotes $\sum_j \ell_j^m (y_j^p w_j^q - y_{j+1}^p w_{j+1}^q)$.

[2] CV denotes cross-validation criteria defined in Blomquist and Newey (2002, p.2464).

C Supplementary material

The supremum of the t-statistics and confidence intervals uniform in the number of series terms

In this supplementary material, we consider the supremum of the t-statistics over all series terms and discuss more about inference methods based on this test statistic. In another direction, this paper also derives the robust inference method after searching over different specifications for nonparametric series estimation.

Suppose a researcher reports only 'favorable' subset of positive results and hiding large different specifications which show overall mixed results or pretending not to search. These practices may lead to distorted inference and the misleading conclusion if we take variability of the first step specification search into account. For example, if a researcher computes many t-statistics and chooses the largest one, then the usual standard normal critical value must be adjusted to control size. The importance of specification search (or data mining/ data snooping) has long been alerted in various other contexts (see Leamer (1983), White (2000), Romano and Wolf (2005), Hansen (2005), and recent papers by Varian (2014), Athey and Imbens (2015), and Armstrong and Kolesár (2015)). Considering the supremum statistic is quite natural to control the size of the joint test in multiple testing literatures.

Specification search is widely used in estimating the parametric model in a less clear way. Although nonparametric series estimation gives a systematic way of doing specification search by restricting the domain of search as $K \in [\underline{K}, \bar{K}]$, little justification has been done, especially for the inference problems. Here, we introduce the tests based on the supremum of the t-statistics over all series terms using the critical values from its asymptotic distribution. We show that this also controls size with undersmoothing conditions. This tests can be used to construct CIs which are uniform in K that have a correct coverage. That is, all confidence intervals using the critical value from supremum t-statistics jointly cover the true parameter at the nominal level, asymptotically. This robust inference method is one way to improve the credibility of inference by admitting search over large sets of different models in nonparametric regression and doing some corrections as usual in multiple testing literatures.

We consider a following 'supremum' t-statistic

$$\mathrm{Sup}\, T_n(\theta) = \sup_{K \in \mathcal{K}_n} |T_n(K, \theta)|. \tag{C.1}$$

The supremum of the t-statistics is appropriate in the context of multiple testing and is known to control the size of the family wise error rate (FWE). We may consider the specification search over large sets of $\mathcal{K}_n$ as simultaneously testing a single hypothesis H_0

based on different test statistics $T_n(K,\theta)$ over $K \in \mathcal{K}_n$. Multiple testing setup is more natural when we focus on the pseudo-true parameter θ_K, i.e., the best linear approximation for $g_0(x)$. One can consider simultaneous testing of individual hypothesis $H_{K,0} : \theta_K = \theta_0$ vs $H_{K,1} : \theta_K \neq \theta_0$ for different $K \in \mathcal{K}_n$. Controlling FWE corresponds to control the following probability asymptotically, $FWE = P(\text{reject at least one hypothesis } H_{K,0}, K \in \mathcal{K}_n) \leq \alpha$.

To derive the asymptotic size of the test and coverage of CI based on the $\operatorname{Sup} T_n(\theta)$, we first provide asymptotic null limiting distribution of the supremum statistics analogous to the Corollary 1 for the infimum test statistic, $\operatorname{Inf} T_n(\theta)$.

Corollary C.1. *1. Under Assumptions 3.1, 3.2, 3.3, and* $\sup_\pi |\nu(\pi)| < \infty$, $\operatorname{Sup} T_n(\theta_0) \xrightarrow{d} \sup_{\pi \in [\underline{\pi},1]} |\mathbb{T}(\pi) + \nu(\pi)|$, *where* $\mathbb{T}(\pi)$ *is the mean zero Gaussian process defined in Theorem 3.1. In addition, if Assumption 3.4 holds, then* $\operatorname{Sup} T_n(\theta_0) \xrightarrow{d} \xi_{\sup} = \sup_{\pi \in [\underline{\pi},1]} |\mathbb{T}(\pi)|$.

2. Suppose Assumptions 3.2 and 3.5 hold. In addition, if $\sup_m |\nu(m)| < \infty$ *are satisifed, then* $\operatorname{Sup} T_n(\theta_0) \xrightarrow{d} \sup_{m=1,\cdots,M} |Z_m + \nu(m)|$ *where* Z_m *is an element of* $M \times 1$ *normal vector* $Z \sim N(0,\Sigma)$ *and* $\nu = (\nu(1),\cdots,\nu(M))'$ *defined in Theorem 3.2. If* $\sup_m |\nu(m)| = \infty$, *then* $\operatorname{Sup} T_n(\theta_0) \xrightarrow{p} \infty$.

Corollary C.1.2 shows that $\operatorname{Sup} T_n(\theta_0)$ converges in probability to infinity under alternative set Assumption 3.5. This implies that the supremum of the t-statistics can be sensitive to those oversmoothing sequences (small K) with high bias. Next Corollary provides the asymptotic size of the test based on $\operatorname{Sup} T_n(\theta)$ similar to Corollary 4.2.

Corollary C.2. *1. Under Assumptions 3.1-3.4, following holds*

$$\limsup_{n\to\infty} P(\operatorname{Sup} T_n(\theta_0) > c^{\sup}_{1-\alpha}) = \alpha. \tag{C.2}$$

2. Under Assumptions 3.1-3.3, and $\sup_\pi |\nu_\pi| < \infty$, *following holds*

$$\limsup_{n\to\infty} P(\operatorname{Sup} T_n(\theta_0) > c^{\sup}_{1-\alpha}) \geq F(c^{\sup}_{1-\alpha}, \sup_\pi |\nu(\pi)|) \tag{C.3}$$

where $F(c,|\nu|) = 1 - \Phi(c - |\nu|) + \Phi(-c - |\nu|)$ *with standard normal cumulative distribution function* $\Phi(\cdot)$.

3. Under Assumptions 3.2, 3.5, and $\sup_m |\nu(m)| = \infty$, $\limsup_{n\to\infty} P(\operatorname{Sup} T_n(\theta_0) > c) = 1$ *for any* $0 < c < \infty$.

Contrary to the $\operatorname{Inf} T_n(\theta)$ test statistic, Corollary C.2.2 shows that the test based on $\operatorname{Sup} T_n(\theta)$ can be sensitive to the large asymptotic bias, and this leads to the over-rejection

of the test. Suppose $F(c^{\sup}_{1-\alpha}, q) = \alpha$ for some $q > 0$. If $\sup_\pi |\nu(\pi)| > q$, then the asymptotic size is strictly greater than α. This also can be seen from the results in C.2.3. If $|\nu(m)| = \infty$ for any m, then the asymptotic size of the test is equal to 1.

Next, we define $CI_{\texttt{sup}}$ based on $\mathrm{Sup}\, T_n(\theta)$ and the critical value $\widehat{c}^{\sup}_{1-\alpha}$ in Section 6.

$$\begin{aligned} CI_{\texttt{sup}} &\equiv \{\theta : \sup_{K \in \mathcal{K}_n} |T_{n,\widehat{V}}(K, \theta)| \leq \widehat{c}^{\sup}_{1-\alpha}\} \\ &= \bigcap_{K \in \mathcal{K}_n} \{\theta : |T_{n,\widehat{V}}(K, \theta)| \leq \widehat{c}^{\sup}_{1-\alpha}\} = [\sup_K(\widehat{\theta}_K - \widehat{c}^{\sup}_{1-\alpha} s(\widehat{\theta}_K)), \inf_K(\widehat{\theta}_K + \widehat{c}^{\sup}_{1-\alpha} s(\widehat{\theta}_K))]. \end{aligned} \tag{C.4}$$

Note that $CI_{\texttt{sup}}$ is an intersection of all CIs in $\mathcal{K}_n$ using critical value $\widehat{c}^{\sup}_{1-\alpha}$.

Corollary C.3. *1. Under Assumptions 3.2, 4.1, 4.2, and 4.3,*

$$\liminf_{n\to\infty} P(\theta_K \in [\widehat{\theta}_K \pm \widehat{c}^{\sup}_{1-\alpha} s(\widehat{\theta}_K)] \quad \forall K \in \mathcal{K}_n) = 1 - \alpha. \tag{C.5}$$

In addition, if Assumption 3.4 (undersmoothing) holds,

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{\texttt{sup}}) = \liminf_{n\to\infty} P(\theta_0 \in CI_K = [\widehat{\theta}_K \pm \widehat{c}^{\sup}_{1-\alpha} s(\widehat{\theta}_K)] \quad \forall K \in \mathcal{K}_n) = 1 - \alpha. \tag{C.6}$$

2. Under Assumptions 3.2, 4.1, 4.2, 4.3, and $\sup_m |\nu(m)| < \infty$*,*

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{\texttt{sup}}) \leq 1 - F(c^{\sup}_{1-\alpha}, \sup_m |\nu(m)|). \tag{C.7}$$

3. Under Assumptions 3.2, 3.5, 4.3, and $\sup_m |\nu(m)| = \infty$*,* $\liminf_{n\to\infty} P(\theta_0 \in CI_{\texttt{sup}}) = 0$*.*

By using an appropriate critical value from the distribution of $\mathrm{Sup}\, T_n(\theta)$, (C.5) gives asymptotic coverage of the uniform confidence intervals over $K \in \mathcal{K}_n$ for the pseudo-true value θ_K. (C.6) gives asymptotic coverage probability of $CI_{\texttt{sup}}$ for the true value θ_0 with undersmoothing assumption, which is same as joint coverage of uniform confidence intervals over $K \in \mathcal{K}_n$.

Corollary C.3.2 and C.3.3 show that the coverage can be sensitive to the asymptotic bias. Especially, uniform coverage results based on $\mathrm{Sup}\, T_n(\theta)$ in (C.6) can be highly sensitive to some small $K \in \mathcal{K}$ which has a large asymptotic bias, so that the coverage probability can be far below than the nominal level. Recall that $CI_{\texttt{sup}}$ is constructed by the intersection of

all confidence intervals in $\mathcal{K}_n$ using larger critical value $\widehat{c}^{\sup}_{1-\alpha}$ than the normal critical value. Intersection can give tighter CI, however, if one of the estimators has a large bias, resulting CI can be too narrow to cover the true parameter. In the worst scenario, the intersection can be empty sets so that the coverage of uniform CIs can be 0. This was formally stated in C.3.3. Under Assumption 3.5, if $|\nu(m)| = \infty$ for some m then asymptotic coverage probability of $CI_{\texttt{sup}}$ is exactly 0.

C.1 Proof of the results in Section C

C.1.1 Proof of Corollary C.1

Proof. The first part follows from Theorem 3.1 and continuous mapping theorem similar to the proof of Corollary 4.1. For the second part of Corollary C.1, consider $S(t) = \sup_m |t_m|$ for $t = (t_1, \cdots, t_M)$ similarly as in the proof of Corollary 4.1. We have

$$\operatorname{Sup} T_n(\theta_0) = \sup_m |T_n(K_m, \theta_0)| = S(T_n(\theta_0)). \tag{C.8}$$

If $\sup_m |\nu(m)| < \infty$, $S(t)$ is continuous at all $t \in \mathbb{R}^M$. Therefore, following holds

$$\operatorname{Sup} T_n(\theta_0) \xrightarrow{d} S(Z + \nu) = \sup_m |Z_m + \nu(m)| \tag{C.9}$$

by Theorem 3.2. If $|\nu_m| = +\infty$ for some m, then then $|T_n(K_m, \theta_0)| \xrightarrow{p} +\infty$, therefore $\operatorname{Sup} T_n(\theta_0) \xrightarrow{p} +\infty$. *Q.E.D.*

C.1.2 Proof of Corollary C.2

Proof. First, we observe that $|T_n(\widehat{K}, \theta_0| \leq \operatorname{Sup} T_n(\theta_0)$ for any $\widehat{K} \in \mathcal{K}_n$. Then we have

$$\limsup_{n\to\infty} P(|T_n(\widehat{K}, \theta)| > c^{\sup}_{1-\alpha}) \leq \limsup_{n\to\infty} P(\operatorname{Sup} T_n(\theta_0) > c^{\sup}_{1-\alpha}) = P(\xi_{\sup} > c^{\sup}_{1-\alpha}) = \alpha$$

by Corollary C.1.1. Next, without assuming Assumption 3.4, we have

$$\begin{aligned}
\limsup_{n\to\infty} P(\operatorname{Sup} T_n(\theta_0) > c^{\sup}_{1-\alpha}) &= P(\sup_{\pi\in[\underline{\pi},1]} |\mathbb{T}(\pi) + \nu(\pi)| > c^{\sup}_{1-\alpha}) \\
&= 1 - P(\sup_{\pi\in[\underline{\pi},1]} |\mathbb{T}(\pi) + \nu(\pi)| \leq c^{\sup}_{1-\alpha}) \\
&\geq \sup_{\pi} [1 - P(|\mathbb{T}(\pi) + \nu(\pi)| \leq c^{\sup}_{1-\alpha})] \\
&= \sup_{\pi} F(c^{\sup}_{1-\alpha}, |\nu(\pi)|) = F(c^{\sup}_{1-\alpha}, \sup_{\pi} |\nu(\pi)|)
\end{aligned}$$

where the first inequality uses $P(\sup_{\pi\in[\underline{\pi},1]}|\mathbb{T}(\pi)+\nu(\pi)|\leq c^{\sup}_{1-\alpha})\leq P(|\mathbb{T}(\pi)+\nu(\pi)|\leq c^{\sup}_{1-\alpha})$ for all π. The third and last equality use the definition of F and monotone increasing property of $F(c,|\nu|)$ with respect to $|\nu|$.

Next, we consider Corollary C.2.3 under alternative set assumption. If $\sup_m|\nu(m)|=\infty$, then $\mathrm{Sup}\,T_n(\theta_0)\xrightarrow{p}+\infty$ by Corollary C.1.2. Thus, $\limsup_{n\to\infty}P(\mathrm{Sup}\,T_n(\theta_0)>c)=1$ since $F(c,\infty)=1$ for any $0<c<\infty$. *Q.E.D.*

C.1.3 Proof of Corollary C.3

Proof. This follows from Corollary 4.3 and Corollary C.2 similar to the proof of Corollary 6.1. Recall that the t-statistic can be written as

$$T_{n,\widehat{V}}(K,\theta_0)=\frac{\sqrt{n}(\widehat{\theta}_K-\theta_0)}{\widehat{V}_K^{1/2}}=\frac{\sqrt{n}(\widehat{\theta}_K-\theta_K)}{\widehat{V}_K^{1/2}}+\frac{\sqrt{n}r_K}{\widehat{V}_K^{1/2}} \tag{C.10}$$

First, consider (C.5),

$$\liminf_{n\to\infty}P(\theta_K\in[\widehat{\theta}_K\pm\widehat{c}^{\sup}_{1-\alpha}s(\widehat{\theta}_K)]\quad\forall K\in\mathcal{K}_n) \tag{C.11}$$

$$=\liminf_{n\to\infty}P(|\frac{\sqrt{n}(\widehat{\theta}_K-\theta_K)}{\widehat{V}_K^{1/2}}|\leq\widehat{c}^{sup}_{1-\alpha}\quad\forall K\in\mathcal{K}_n)=\liminf_{n\to\infty}P(\sup_K|\frac{\sqrt{n}(\widehat{\theta}_K-\theta_K)}{\widehat{V}_K^{1/2}}|\leq\widehat{c}^{sup}_{1-\alpha}) \tag{C.12}$$

$$=P(\sup_m|Z_m|\leq c^{sup}_{1-\alpha})=1-\alpha \tag{C.13}$$

where the last equality follows from Theorem 3.1 and Corollary 4.3 under Assumptions 3.2, 4.1, 4.2, and 4.3. In addition, if Assumption 3.4 holds, we have that

$$\liminf_{n\to\infty}P(\theta_0\in CI_{\mathtt{sup}})=\liminf_{n\to\infty}P(\mathrm{Sup}\,T_n(\theta_0)\leq\widehat{c}^{sup}_{1-\alpha}) \tag{C.14}$$

$$(=\liminf_{n\to\infty}P(|T_{n,\widehat{V}}(K,\theta_0)|\leq\widehat{c}^{sup}_{1-\alpha}\quad\forall K\in\mathcal{K}_n)) \tag{C.15}$$

$$=P(\sup_m|Z_m|\leq c^{sup}_{1-\alpha})=1-\alpha. \tag{C.16}$$

This completes the first part of Corollary C.3. The second part can be shown similarly to the proof of Corollary C.2.2. For the last part, if $\sup_m|\nu(m)|=\infty$, then $\liminf_{n\to\infty}P(\theta_0\in CI_{\mathtt{sup}})=0$ by Corollary C.2.3.

Q.E.D.