Essays on Nonparametric Inference and Instrument Selection

by

Byunghoon Kang

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Economics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2016

Date of final oral examination: 05/05/2016

The dissertation is approved by the following members of the Final Oral Committee:

Bruce E. Hansen, Professor, Economics, Chair

Jack R. Porter, Professor, Economics

Xiaoxia Shi, Assistant Professor, Economics

Joachim Freyberger, Assistant Professor, Economics

Chunming Zhang, Professor, Statistics

To my wife, Yena.

Acknowledgments

Over the past five years, I have received support and benefited from numerous people in this amazing world. First, I am deeply indebted to my advisor Bruce Hansen for his guidance and encouragement. Without his continuous support and help, I may never have completed this dissertation. He was always so generous with his time, and, of course, patience. He has been a good mentor, teacher, and advisor in every stage of the dissertation. What I learned from him will certainly be a valuable asset throughout my entire life. I would also like to thank Jack Porter for his thoughtful discussions and suggestions. He always helped me go through with a more intuitive way and think with a broader perspective. I would like to thank him for being such a great placement director over the past few years, and of course helping me get a job. I am also grateful to Xiaoxia Shi. She always has been an influential mentor from the very beginning of my graduate studies. Her enthusiasm as a great scholar and a teacher in Econometrics helped me a lot in various ways.

Other than my committee members, special mention goes to Joachim Freyberger for always being very sharp and get to the point during our conversations. I am also thankful to Chunming Zhang for her time and being a member of the final oral examination. I acknowledge helpful conversations with my friends, other graduate students and numerous seminar speakers at UW-Madison in many ways. I also acknowledge support by Kwanjeong Educational Foundation Graduate Research Fellowship and Leon Mears Dissertation Fellowship from UW-Madison.

I would like to thank my parents and my family as well. Although my son, Joonwoo has

been little help to write this dissertation, he gave me a whole new, different perspective on my life. Last, but foremost, I would like to show my gratitude and respect to my wife, Yena Nam. She decided to sacrifice everything to stay with me. Without her support, unlimited patience and unconditional love, this hasn't been an enjoyable journey at all. I dedicate this dissertation to her.

Contents

Contents **iv**

List of Tables **vi**

List of Figures **vii**

Abstract **viii**

1 Inference in Nonparametric Series Estimation with Data-Dependent Number of Series Terms **1**

1.1 Introduction 1

1.1.1 Related Literature 5

1.1.2 Notation 8

1.2 Nonparametric Series Regression 9

1.3 Asymptotic Distribution of the Joint t-statistics 11

1.3.1 Weak Convergence of t-statistic Process 11

1.3.2 Alternative Set with Different Rates 15

1.4 Test Statistic 17

1.4.1 Asymptotic Distribution of the Test Statistic 18

1.4.2 Asymptotic Size of the Test Statistic 19

1.4.3 Critical Values 22

1.5 Confidence Intervals 25

1.6 Post-Model Selection Inference 28
1.7 Extension: Partially Linear Model 30
1.8 Simulations 35
1.9 Illustrative Empirical Application : Nonparametric Estimation of Labor Supply Function and Wage Elasticity with Nonlinear Budget Set 39
1.10 Conclusion 42
1.11 References 43
1.12 Proofs 48
1.13 Figures and Tables 69
1.14 Supplementary Material 78

2 Higher Order Approximation of IV Estimators with Locally Invalid Instruments 85

2.1 Introduction 85
2.1.1 Related Literature 88
2.2 Linear IV Model With Locally Invalid Instruments 91
2.3 Higher-Order MSE Approximation with Locally Invalid Instruments 92
2.4 Higher Order MSE Approximation under Drifting Sequences Faster than $N^{1/2}$ 99
2.5 MSE Approximation of 2SLS under $K = O(\sqrt{N})$ 102
2.6 Invalidity-Robust Criteria to Choose Instruments 103
2.7 Conclusions 108
2.8 References 109
2.9 Proofs 115

List of Tables

1.1 Nonparametric Wage Elasticity of Hours of Work Estimates in Blomquist and Newey (2002, Table 1) 77

List of Figures

1.1 Different functions of $g(x)$ 69
1.2 Plots of $F(c, \nu)$ 70
1.3 Coverage - Polynomials 71
1.4 Coverage - Splines 72
1.5 Length of CIs - Polynomials 73
1.6 Length of CIs - Splines 74
1.7 Power function against fixed alternatives 75
1.8 Patterns of t-statistics with K 76

Abstract

My dissertation consists of two chapters on nonparametric inference and model selection in econometric models.

Researchers in economics and social science need reliable models and statistical tools to quantify economic relationships and uncertainty associated with data. In practice, researchers often evaluate their object of interests with various specifications in the first stage of analysis or select model by some criteria. Unfortunately, commonly used statistical methods may fail to assess uncertainty inherent in the first step specification search. Moreover, some existing model selection criteria may be fragile due to model misspecification errors. All these methods can lead to misleading conclusions without valid, robust corrections. To quantify and test economic theories more accurately in such cases, researchers and policy makers need more reliable and robust methods. My research investigates these issues and provides practical methods in empirical research with rigorous theoretical justifications.

First chapter provides new inference methods in nonparametric series regression with data dependent number of series terms. Nonparametric series estimation have increased their popularity as it gives flexible method addressing potential misspecification of the parametric model. However, implementation in practice requires a choice of *the number of series terms* and the estimation and inference may largely depend on its choice. 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 chapter 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 a uniform asymptotic theory for the t-statistic process indexed by the number of series terms. Using this result, I show that the test based on the infimum of the t-statistics and its asymptotic critical value controls the asymptotic size with the undersmoothing condition. We can construct a valid confidence interval (CI) by test statistic inversion that has correct asymptotic coverage probability. Even when asymptotic bias terms are present without the undersmoothing condition, I show that the CI based on the infimum of the t-statistics bounds the 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 existing methods with possibly ad hoc choice of series terms.

Second chapter provides instrument selection criteria in instrumental variable (IV) regression model when there is a large set of instruments with potential invalidity. Economic data identified by IV model sometimes involve large sets of potential instruments and debates about their validity. Existing methods for instrument selection are largely based on *a priori* assumption of an instrument's validity and/or based on the first-order asymptotics, which may lead to a large finite sample bias with many and invalid instruments. First, I derive higher-order mean square error (MSE) approximation for two-stage least squares (2SLS), limited information maximum likelihood (LIML), modified Fuller (FULL) and bias-adjusted 2SLS (B2SLS) estimator allowing locally invalid instruments. Based on the approximation to the higher-order MSE, I propose an invalidity-robust instrument selection criteria (IRC) that capture two sources of finite sample bias at the same time: bias from using many instruments and bias from invalid instruments. I also show optimality result of choice of instruments based on the criteria of Donald and Newey (2001) under certain locally invalid instruments specification.

Chapter 1

Inference in Nonparametric Series Estimation with Data-Dependent Number of Series Terms

1.1 Introduction

Nonparametric series estimation has received attention in both theoretical econometrics and applied economics. I consider the following nonparametric regression model;

$$\begin{aligned} y_i &= g_0(x_i) + \varepsilon_i, \\ E(\varepsilon_i|x_i) &= 0 \end{aligned} \tag{1.1.1}$$

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 (1.1.1) 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 se-

ries 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 (LS). 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, the required K may vary with different data sets to accommodate the smoothness and nonlinearity of the 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* (or *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 variance. 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 the bias-correction available to date. Ignoring asymptotic bias with this undersmoothing assumption, one can apply the conventional confidence interval (CI) using the standard normal critical value, with estimate and standard error based on some choice of "large" K. However, the asymptotic theory does not provide specific guidelines for choosing a "sufficiently large" number of series terms to make the bias small in practice. Some 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. 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 $K \in [\underline{K}, \bar{K}]$ with the given sample sizes n. For example, a researcher may use quadratic, cubic, or quartic terms in the polynomial regression, or try a different number of knots in the 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 or Akaike information criterion (AIC)) and some rule-of-thumb methods that are suggested for inference, also require eval-

uating estimates with different Ks. If some researchers evaluate different estimators with different number of terms, it is not clear how this randomness affects inference.

In this paper, I construct inference methods in nonparametric series regression with data-dependent number of series terms. I consider the testing problem for a regression function at a point and introduce tests based on the *infimum of the studentized t-statistics* over different series terms. Tests based on the infimum t-statistics and searching for the small t-statistic have a similar motivation to the one on which the undersmoothing condition is theoretically based: searching for the "large" K that has a small bias and large variance. 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, especially for inference (for example, see Newey (2013), Newey, Powell and Vella (2003)). Here, I formally justify this conventional wisdom by introducing the infimum test statistic, and provide an inference method based on its asymptotic distribution.

For this, I first provide a uniform asymptotic theory for the t-statistic process indexed by the number of series terms. Existing asymptotic normality of the t-statistic in the literature holds under a deterministic sequence of $K \to \infty$ as the sample size n increases. The main contribution of this paper is to derive the asymptotic distribution theory for the entire sequences of t-statistics over a range of K.

Using this result, I show that the test based on the infimum of the t-statistics and its asymptotic critical value control the asymptotic size (null rejection probability) of the test with the undersmoothing condition for all Ks in a set. Allowing asymptotic bias without the undersmoothing condition, I also analyze the effect of bias on the size of the tests. Even when asymptotic bias terms are present, the tests based on the infimum t-statistic bound the size distortions, in the sense that the asymptotic size of the tests is bounded above by the asymptotic size of a single t-statistic with 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 true parameter 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. I find that our proposed CI performs well in Monte Carlo experiments; coverage probability of the CI based on the infimum t-statistics is close to the nominal level in various simulation setups. I also find that this CI bounds the coverage distortions even when asymptotic bias is present. As an illustrative example, I revisit nonparametric estimation of wage elasticity of the expected labor supply, as in Blomquist and Newey (2002).

As a by-product of the joint asymptotic distribution results, 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, this 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. I expect this lead to a tighter CI than those based on the Bonferroni-type critical value, as we 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 $n^{1/2}$ (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, in which K increases much slower than the sample size n. To fully account for joint dependency of the t-statistics with

the different sequences of Ks in the partially linear model setup, this requires a different approximation theory than the nonparametric regression setup. 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 are able to account for the dependency of t-statistics with different Ks. I also propose methods to construct CIs that are similar to the nonparametric regression setup and provide their asymptotic coverage properties. Potential empirical applications include, but are not limited to, estimation of the treatment effect model with series approximations.

1.1.1 Related Literature

The literature on the nonparametric series estimation is vast, but data-dependent series term selection and its impact on estimation or inference is 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. 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, derived under a deterministic sequence for K.

Recent papers by Horowitz (2014), Chen and Christensen (2015a) develop novel 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. In this paper, we focus on the inference problem rather than estimation with the similar issues arising from cross-validation.

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, and Horowitz and Lee (2012) suggested increasing K until variance suddenly increases. They discuss these methods work well in practice and simulation for the inference. Here, with the similar ideas, we can account the randomness introduced in the first step specification search by introducing the infimum test statistic and provide formal inference methods based on its asymptotic distribution results.

Several important papers have investigated the asymptotic properties of series (and sieves) estimators, including papers by Andrews (1991a), Eastwood and Gallant (1991), Newey (1997), Huang (2003a), Chen (2007), Belloni, Chernozhukov, Chetverikov, and Kato (2015), and Chen and Christensen (2015b), among many others. They focused on (pointwise and uniform) convergence rates, asymptotic normality for series estimators, and inference on (linear and 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.

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. 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. Hall and Horowitz (2013) proposed CIs using first-stage bootstrap methods to account for the bias of 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 of nonparametric estimators

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 formula compare with kernel estimation. In recent paper, Hansen (2014) introduce a bias-robust CI using the critical value from a non-central normal distribution with an estimated asymptotic bias.

A recent paper that is concurrent with this paper, Armstrong and Kolesár (2015) considered inference methods in kernel estimation. 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 usual correction for the multiple testing problem. 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)). Although this paper has analogous results with Armstrong and Kolesár (2015) considering supremum of the t-statistics (see Section 1.14), the main focus of this paper is asymptotic bias and undersmoothing condition, which may be crucial in series estimation. Compare with the new tests based on the infimum t-statistics, inference based on the supremum t-statistic can be sensitive to the bias problems, i.e., supremum t-statistics may pick estimator with huge bias under the null that lead to over-rejection of the test.[1]

The outline of the paper is as follows. I first introduce basic nonparametric series regression setup in Section 1.2. In Section 1.3, I provide an empirical process theory for the t-statistic sequences over a set. Section 1.4 introduces infimum of the t-statistic and describes the asymptotic null distributions of the test statistic. Then, I provide the asymptotic size

[1]We may also consider other types of t-statistics that is robust to the bias issues such as "median" 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.

results of the test and implementation procedure for the critical value. Section 1.5 introduces CIs based on the infimum test statistic and provides their coverage properties. Section 1.6 analyzes valid post-model selection inference in this setup. Section 1.7 extends our inference methods to the partially linear model setup. Section 1.8 includes Monte Carlo experiments in various setups. Section 1.9 illustrates proposed inference methods using the nonparametric estimation of wage elasticity of the expected labor supply, as in Blomquist and Newey (2002), then Section 1.10 concludes. Section 1.12 and 1.13 include all proofs, figures and tables. Section 1.14 discuss inference procedures based on the supremum of the t-statistics.

1.1.2 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 a 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 $\mathbb{R}_{[\pm\infty]} = \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$.

1.2 Nonparametric Series Regression

In this section, I first introduce the nonparametric series regression setup. 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 pointwise inference in x and I will omit the dependence on x if there is no confusion.

We consider sequence of approximating model indexed by 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. Basis functions may come from set of large number of potential regressors and/or their nonlinear transformations.

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

$$\widehat{g}_K(x) = P_K(x)'\widehat{\beta}_K, \qquad \widehat{\beta}_K = (P^{K'}P^K)^{-1}P^{K'}Y \tag{1.2.1}$$

where $P_{Ki} \equiv P_K(x_i) = (p_1(x_i), p_2(x_i), \cdots, p_K(x_i))', P^K = [P_{K1}, \cdots, P_{Kn}]', Y = (y_1, \cdots y_n)'$. We can think of $\widehat{g}_K(x)$ as an estimator of the best linear approximation for $g_0(x)$, $P_K(x)'\beta_K$, where β_K can be defined as the best linear projection coefficients, $\beta_K \equiv (E[P_{Ki}P_{Ki}'])^{-1}E[P_{Ki}y_i]$. For some $x \in \mathcal{X}$, define the approximation error using K series terms as $r_K(x) = g_0(x) - P_K(x)'\beta_K$. 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

$$y_i = P_{Ki}'\beta_K + \varepsilon_{Ki}, \quad E[P_{Ki}\varepsilon_{Ki}] = 0 \tag{1.2.2}$$

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 asymptotic variance formula

$$\begin{aligned} V_K &\equiv V_K(x) = P_K(x)'Q_K^{-1}\Omega_K Q_K^{-1}P_K(x), \\ Q_K &= E(P_{Ki}P_{Ki}'), \quad \Omega_K = E(P_{Ki}P_{Ki}'\varepsilon_i^2) \end{aligned} \tag{1.2.3}$$

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

We are interested in (two-sided) testing for θ

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0. \tag{1.2.4}$$

The studentized t-statistic for H_0 is

$$T_n(K, \theta_0) \equiv \frac{\sqrt{n}(\widehat{g}_K(x) - g_0(x))}{V_K^{1/2}} = \frac{\sqrt{n}(\widehat{\theta}_K - \theta_0)}{V_K^{1/2}}. \tag{1.2.5}$$

Under standard regularity conditions (will be discussed in Section 1.3) including an under-smoothing rate for *deterministic sequence* $K \to \infty$ as $n \to \infty$, the asymptotic distribution of the t-statistic is well known

$$T_n(K, \theta_0) \xrightarrow{d} N(0, 1). \tag{1.2.6}$$

See, for example, Andrews (1991a), Newey (1997), Belloni et al. (2015), Chen and Christensen (2015b). In the next section, I formally develop an asymptotic distribution theory of the t-statistic (1.2.5) in K over a set $\mathcal{K}_n$.

1.3 Asymptotic Distribution of the Joint t-statistics

1.3.1 Weak Convergence of t-statistic Process

In this section, I provide asymptotic distribution theory of the joint t-statistics over a set. 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 parameter π, which is a 'fraction' of the largest series terms $\bar{K}$.

Assumption 1.1. *(Set of number of series terms) Let $\mathcal{K}_n$ as*

$$\mathcal{K}_n = \{K \in \mathbb{N}^+ : K \in [\underline{K}, \bar{K}]\}$$

where $\underline{K} \equiv \lfloor \underline{\pi} \bar{K} \rfloor$ for $\underline{\pi} \in (0,1)$ and $\mathbb{N}^+$ is the set of all positive integers.

The standard inference methods in this setup typically consider singleton set $\mathcal{K}_n = \{K\}$. Assumption 1.1 considers range of number of series terms and considers (infinite) sequence of models indexed by $\pi \in \Pi = [\underline{\pi}, 1]$ using $K = \lfloor \pi \bar{K} \rfloor$ series terms. Note that $\mathcal{K}_n$ is indexed by sample size n, as I will impose rate conditions for the largest $\bar{K} \equiv \bar{K}(n)$ in the next Assumption 1.2.

We now consider the sequence of t-statistics $T_n(K, \theta)$ defined in (1.2.5) for $K \in \mathcal{K}_n$. Under Assumption 1.1, I define the *t-statistic process*, $T_n^*(\pi, \theta)$, indexed by $\pi \in \Pi = [\underline{\pi}, 1]$ as

$$T_n^*(\pi, \theta) \equiv T_n(\lfloor \pi \bar{K} \rfloor, \theta). \tag{1.3.1}$$

$T_n^*(\pi, \theta)$ is the t-statistic using $K = \lfloor \pi \bar{K} \rfloor$ number of series terms. Note that $T_n^*(\pi, \theta)$ is a step function of π.

In addition to imposing the set assumption, I impose mild regularity conditions that are standard in nonparametric series regression literature and are satisfied by well-known basis functions. I closely follow assumptions in the recent paper by Belloni et al. (2015), Chen and

Christensen (2015b) and impose rate conditions of K uniformly over $\mathcal{K}_n$. Other regularity conditions in the literature (e.g., Newey (1997)) can also be imposed here with different rate conditions of K.

Define $\zeta_K \equiv \sup_{x\in\mathcal{X}} ||P_K(x)||$ as the largest normalized length of the regressor vector for each $K \in \mathcal{K}_n$, and $\lambda_K \equiv \lambda_{min}(Q_K)^{-1/2}$ for $K \times K$ design matrix $Q_K = E(P_{Ki}P_{Ki}')$.

Assumption 1.2. *(Regularity conditions)*

(1) $\{y_i, x_i\}_{i=1}^n$ *are i.i.d random variables satisfying the model* (1.1.1).

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

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

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

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

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

Assumption 1.2-(2) imposes moment conditions and standard uniform integrability conditions. Assumption 1.2-(3) is satisfied with various basis functions. If the support $\mathcal{X}$ is a cartesian product of compact connected intervals (e.g. $\mathcal{X} = [0,1]^{d_x}$), 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 1.2-(3) vary with 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 series/sieves basis). When the probability density function of x_i is uniformly bounded above

and bounded away from zero over compact support $\mathcal{X}$ and orthonormal basis is used, then we have $\lambda_K \lesssim 1$ (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 1.2-(5) can be replaced by the specific bounds of ζ_K, c_K, ℓ_K. For example, for the power series, Assumption 1.2-(5) reduced to $\sup_{K\in\mathcal{K}_n} \sqrt{K^2(\log K)/n}(1 + K^{3/2-p/d_x}) = \sqrt{\bar{K}^2(\log \bar{K})/n}(1 + \bar{K}^{3/2-p/d_x}) \to 0$ with the Assumption 1.1.

Together with the Assumption 1.2, set $\mathcal{K}_n$ in Assumption 1.1 considers the sequence of models that has the same rate of K, i.e., $K \asymp K'$ for any $K, K' \in \mathcal{K}_n$. Dimension of $\mathcal{K}_n$ is $|\mathcal{K}_n| = \lfloor \bar{K}(1 - \underline{\pi})\rfloor + 1 \to \infty$ as $n \to \infty$. Assumption 1.1 does not consider all different sequences of K satisfying asymptotic normality of series estimators, however, these are appropriate sequences to be able to develop joint distributions of the t-statistics. As studentized t-statistic is normalized by variance terms V_K which increases differently with different rates of K, two t-statistics with different rates are asymptotically independent, thus hard to incorporate dependency (see Section 1.3.2 for formal results). Therefore, this set assumption is important for our theory to provide uniform central limit theorem of the t-statistic process.

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 $r_\pi \equiv r_\pi(x) = r_{\lfloor \bar{K}\pi\rfloor}(x)$. Asymptotic 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 1.1 and 1.2, the t-statistic process under H_0 can be decomposed as follows

$$T_n^*(\pi, \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{P_\pi(x)' P_{\pi i} \varepsilon_i}{V_\pi^{1/2}} - \sqrt{n} V_\pi^{-1/2} r_\pi + o_p(1), \quad \pi \in \Pi \tag{1.3.2}$$

where $\sqrt{n} V_\pi^{-1/2} r_\pi$ is a bias term due to approximation errors. I define the asymptotic bias for the sequence of models indexed by π as the limit of the second term

$$\nu(\pi) \equiv \lim_{n\to\infty} -\sqrt{n} V_\pi^{-1/2} r_\pi. \tag{1.3.3}$$

Under the following undersmoothing condition, the asymptotic bias $\nu(\pi)$ is 0. To assess the effect of bias on inference, we will consider distinctions between imposing undersmoothing condition and not.

Assumption 1.3. *(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 $c_K\ell_K \lesssim K^{-p/d_x}K$ for the power series, Assumption 1.3 can be replaced by $\sup_{K\in\mathcal{K}_n} |\sqrt{n}K^{1/2-p/d_x}| = \sqrt{n}\bar{K}^{1/2-p/d_x} \to 0$ since pointwise standard error $V_K^{1/2} \propto \sqrt{K}$.

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

Theorem 1.1. *Under Assumptions 1.1, 1.2 and* $\sup_\pi |\nu(\pi)| < \infty$,

$$T_n^*(\pi, \theta_0) \Rightarrow \mathbb{T}(\pi) + \nu(\pi) \tag{1.3.4}$$

where $\mathbb{T}(\pi)$ *is a mean zero Gaussian process on* $\ell^\infty(\Pi)$ *with covariance function* $\Sigma(\pi_1, \pi_2) = \lim_{n\to\infty} \Sigma_n(\pi_1, \pi_2)$, *where*

$$\Sigma_n(\pi_1, \pi_2) = \frac{P_{\pi_1}(x)' E(P_{\pi_1 i} P'_{\pi_2 i} \varepsilon_i^2) P_{\pi_2}(x)}{V_{\pi_1\wedge\pi_2}^{1/2} V_{\pi_1\vee\pi_2}^{1/2}} \tag{1.3.5}$$

for any $\pi_1, \pi_2 \in \Pi$, *and* $\nu(\pi)$ *is defined in* (1.3.3). *In addition, if Assumption 1.3 is satisfied, then*

$$T_n^*(\pi, \theta_0) \Rightarrow \mathbb{T}(\pi), \quad \pi \in \Pi. \tag{1.3.6}$$

Theorem 1.1 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 asymptotic null distribution of the t-statistic process in (1.3.4) is equal to a mean zero Gaussian process $\mathbb{T}(\pi)$ plus the asymptotic bias $\nu(\pi)$.

Remark 1. (Covariance function) Under conditional homoskedasticity, $E(\varepsilon_i^2|x_i = x) = \sigma^2$, the covariance function of the limiting Gaussian process reduces to the simple form

$$\Sigma(\pi_1, \pi_2) = \lim_{n\to\infty} \frac{V_{\pi_1\wedge\pi_2}^{1/2}}{V_{\pi_1\vee\pi_2}^{1/2}} \tag{1.3.7}$$

for any $\pi_1, \pi_2 \in \Pi$. This is well defined since the rates of V_{π_1} and V_{π_2} are the same, i.e., $V_{\pi_1} \asymp V_{\pi_2}$. For example, if we consider polynomial basis $P_K(x) = (1, x^1, \cdots, x^{K-1})'$ and the point $x = 1$, then $\Sigma(\pi_1, \pi_2) = \lim_{n\to\infty} K_{\pi_1\wedge\pi_2}^{1/2}/K_{\pi_1\vee\pi_2}^{1/2} = (\frac{\pi_1\wedge\pi_2}{\pi_1\vee\pi_2})^{1/2}$ and it only depends on π_1, π_2.

Remark 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 analogous results for the partially linear model in Section 1.7.

Remark 3. (Rate conditions) Note that the asymptotic bias $|\nu(\pi)|$ in (1.3.3) is zero if $\bar{K}$ increases faster than the optimal MSE rate (undersmoothing), is non-zero but finite if $\bar{K}$ increases at the optimal MSE rate, and $|\nu(\pi)|$ is infinity if $\bar{K}$ increases slower than the optimal MSE rate (oversmoothing). Theorem 1.1 does not allow oversmoothing rates $|\nu(\pi)| = \infty$, as we require $\sup_\pi |\nu(\pi)| < \infty$.

1.3.2 Alternative Set with Different Rates

Next, we provide different approximations to the sequence of t-statistics with an alternative set $\mathcal{K}_n$ constructed to allow different rates of Ks. This alternative set assumption considers

broader range of Ks than the Assumption 1.1, as it considers different rates and allows oversmoothing rates of K which increases slower than the optimal rate.

Assumption 1.4. *(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* $K_m \equiv \tau n^{\phi_m}$ *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* $n \to \infty$.

If $|\nu(m)| = \infty$ for some m, alternative set assumption considers rates of K from oversmoothing to undersmoothing with different ϕ_m. Here, $\underline{K}$ can increases slowly and $\bar{K}$ satisfies undersmoothing rates. Undersmoothing assumption for the $\bar{K}$, i.e. $\nu(M) = 0$, may be restrictive. However, this is merely a modeling device considering broad rage of K and taking some large enough $\bar{K}$ so that satisfy undersmoothing.

Note that Assumption 1.4 only considers finite K sequences, i.e., $|\mathcal{K}_n| = M$. In finite samples, we only consider finite set $\mathcal{K}_n$, so the difference between Assumption 1.1 and 1.4 only matters in large samples. Different rate conditions lead to different approximations to the sequence of t-statistics. As shown in Theorem 1.1, we can incorporate dependency of the t-statistics under Assumption 1.1. However, as it considers the sequence of K with the same growth rates which only differ in constant π, Theorem 1.1 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 $K \in \mathcal{K}_n$. Although, Assumption 1.4 only considers the finite sequence of t-statistics, it is useful to consider the effect of bias on inference problems that will be considered in Section 1.4.

Similar to Theorem 1.1, if we impose $\sup_m |\nu(m)| < \infty$, then the joint t-statistics converge in distribution to a normal distribution with the asymptotic bias terms under the alternative set assumption. However, joint t-statistics do not converge in distribution to a bounded random vector if some of the elements $|\nu(m)| = \infty$ with oversmoothing sequences.

This matters when we obtain the asymptotic distribution of the test statistic that is some continuous transformation of the joint t-statistics, and I will discuss this in Section 1.4.

Theorem 1.2. *Under Assumptions 1.2, 1.4 and* $\sup_m |\nu(m)| < \infty$*,*

$$(T_n(K_1, \theta_0), \cdots, T_n(K_M, \theta_0))' \xrightarrow{d} Z + \nu$$

where $Z = (Z_1, \cdots, Z_M)' \sim N(0, I_M)$ *and* $\nu = (\nu(1), \cdots, \nu(M))'$ *are* $M \times 1$ *vectors.*

If $\sup_m |\nu(m)| = \infty$*, then following holds for any strictly increasing continuous distribution function on* $\mathbb{R}$*,* $G(\cdot)$*,*

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

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* $\nu(m) = -\infty$*.*

1.4 Test Statistic

In this section, I introduce an *infimum* test statistic and analyze its asymptotic null distribution based on Theorem 1.1 and 1.2. Then, I provide the asymptotic size result of the tests, and methods to obtain the critical value for our inference procedures.

I consider following test statistic

$$\text{Inf}\, T_n(\theta) \equiv \inf_{K \in \mathcal{K}_n} |T_n(K, \theta)|. \tag{1.4.1}$$

As I denoted in the introduction, there are several reasons to consider $\text{Inf}\, T_n$ in series regression context. First of all, small t-statistic centered at the true value corresponds to the approximation with K that has a small bias and large variance, which is good for the coverage

(for the size as well) as what undersmoothing assumption does for eliminating asymptotic bias, theoretically. 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 (2003)).

1.4.1 Asymptotic Distribution of the Test Statistic

Asymptotic null limiting distribution of the infimum test statistic follows immediately from Theorem 1.1 and 1.2.

Corollary 1.1. *1. If Assumptions 1.1, 1.2 and* $\sup_\pi |\nu(\pi)| < \infty$ *are satisfied, then* $\operatorname{Inf} T_n(\theta_0) \xrightarrow{d} \inf_{\pi\in[\underline{\pi},1]} |\mathbb{T}(\pi) + \nu(\pi)|$*, where* $\mathbb{T}(\pi)$ *is the mean zero Gaussian process defined in Theorem 1.1. In addition, if Assumption 1.3 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 1.2 and 1.4, $\operatorname{Inf} T_n(\theta_0) \xrightarrow{d} \inf_{m=1,\cdots,M} |Z_m + \nu(m)|$*, where* Z_m *is an element of* $M \times 1$ *normal vector* $Z \sim N(0, I_M)$ *and* $\nu = (\nu(1), \cdots, \nu(M))'$ *is defined in Theorem 1.2.*

Corollary 1.1-1 derives the asymptotic null limiting distribution of $\operatorname{Inf} T_n(\theta)$ under $\mathcal{K}_n$ with same rates of K (Assumption 1.1). Corollary 1.1-2 provides the asymptotic distribution under alternative $\mathcal{K}_n$ with different rates of K (Assumption 1.4).

Whether some $|\nu(m)|$ are bounded or not, Corollary 1.1-2 shows that $\operatorname{Inf} T_n(\theta_0)$ converge in distribution to the bounded random variable. Under H_0, $\operatorname{Inf} T_n(\theta)$ exclude all small Ks corresponding to oversmoothing (where the bias is of larger order than the variance) 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 1.4.2 (for size results) and Section 1.5 (for coverage results).

1.4.2 Asymptotic Size of the Test Statistic

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 1.1-1, i.e., solves

$$P(\xi_{\inf} > c^{\inf}_{1-\alpha}) = \alpha \tag{1.4.2}$$

for $0 < \alpha < 1/2$. The asymptotic null distribution, $\xi_{\inf}$, can be completely defined by covariance kernel of the limiting Gaussian process $\mathbb{T}(\pi)$ in Theorem 1.1. Since the limiting process can not be written as some transformation of Brownian motion process, the asymptotic critical value cannot be tabulated, in general. However, critical value can be obtained by standard Monte Carlo method or by the weighted bootstrap method. I will discuss approximation of the critical value in Section 1.4.3.[2]

Next, we define $z_{1-\alpha/2}$ as $(1-\alpha/2)$ quantile of standard normal distribution function, which also solves $P(|Z| > z_{1-\alpha/2}) = \alpha$ where $Z \sim N(0,1)$. Next corollary provides the asymptotic size results of the tests based on $\operatorname{Inf} T_n(\theta)$ follow from the asymptotic distribution results in Corollary 1.1.

Corollary 1.2. *1. Under Assumptions 1.1, 1.2 and 1.3, following holds with critical values $c^{\inf}_{1-\alpha}$ defined in* (1.4.2) *and the normal critical value $z_{1-\alpha/2}$,*

$$\limsup_{n\to\infty} P(\operatorname{Inf} T_n(\theta_0) > c^{\inf}_{1-\alpha}) = \alpha, \qquad \limsup_{n\to\infty} P(\operatorname{Inf} T_n(\theta_0) > z_{1-\alpha/2}) \le \alpha. \tag{1.4.3}$$

2. Suppose Assumptions 1.1 and 1.2 hold. If $\sup_\pi |\nu_\pi| < \infty$, then following inequality

[2]Without imposing the undersmoothing assumption, asymptotic distribution of $\operatorname{Inf} T_n(\theta_0)$ in Corollary 1.1-1 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. This approach is a difficult problem that is beyond the scope of this paper. See Hansen (2014) for this important direction with single $\Pi = \pi$ and the critical value from $|N(0,1) + \widehat{\nu}(\pi)|$.

holds

$$\limsup_{n\to\infty} P(\mathrm{Inf}\, T_n(\theta_0) > c^{\mathrm{inf}}_{1-\alpha}) \leq F(c^{\mathrm{inf}}_{1-\alpha}, \inf_{\pi} |\nu(\pi)|), \tag{1.4.4}$$

$$\limsup_{n\to\infty} P(\mathrm{Inf}\, T_n(\theta_0) > z_{1-\alpha/2}) \leq F(z_{1-\alpha/2}, \inf_{\pi} |\nu(\pi)|), \tag{1.4.5}$$

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

3. *Under Assumptions 1.2 and 1.4, following holds*

$$\limsup_{n\to\infty} P(\mathrm{Inf}\, T_n(\theta_0) > c^{\mathrm{inf}}_{1-\alpha}) = \prod_{m=1}^{M} F(c^{\mathrm{inf}}_{1-\alpha}, |\nu(m)|), \tag{1.4.6}$$

$$\limsup_{n\to\infty} P(\mathrm{Inf}\, T_n(\theta_0) > z_{1-\alpha/2}) \leq \alpha. \tag{1.4.7}$$

Corollary 1.2-1 shows that the tests based on the infimum test statistic asymptotically control size with the undersmoothing condition. As $\mathrm{Inf}\, T_n(\theta_0) \leq |T_n(K, \theta_0)|$ and $|T_n(K, \theta_0)| \xrightarrow{d} |N(0, 1)|$ for any single $K \in \mathcal{K}_n$, the test based on $\mathrm{Inf}\, T_n(\theta)$ using normal critical value also controls the asymptotic size.

Without undersmoothing assumption, Corollary 1.2-2 derives the upper bounds of the asymptotic null rejection probability of the tests based on $\mathrm{Inf}\, T_n(\theta)$. Equations (1.4.4) and (1.4.5) show that the asymptotic size is bounded above by the asymptotic size of a single t-statistic with the smallest asymptotic bias. Note that $F(c, |\nu|)$ is a monotone decreasing function of c. Typically $c^{\mathrm{inf}}_{1-\alpha} < z_{1-\alpha/2}$ holds, so that $F(z_{1-\alpha/2}, 0) = \alpha < F(c^{\mathrm{inf}}_{1-\alpha}, 0)$. Moreover, $F(c, |\nu|)$ is a monotone increasing function of $|\nu|$ (see also Hall and Horowitz (2013), Hansen (2014) for the similar function F and Figure 1.2 for the plots of F as a function of $|\nu|$ with different c). Also note that the right hand side of (1.4.5) is exactly equal to α if the smallest bias is 0, $\inf_{\pi} |\nu(\pi)| = 0$.

Corollary 1.2-3 shows the asymptotic size results of the test under the alternative set

assumption (Assumption 1.4). Equation (1.4.6) gives useful information about the effect of asymptotic bias on the asymptotic size by allowing 'large' asymptotic bias $|\nu(m)| = \infty$. First, the asymptotic size of the test based on $\text{Inf}\, T_n(\theta)$ is not affected by K_m such that $|\nu(m)| = \infty$ (oversmoothing), as $F(c, \infty) = 1$ for any bounded $c > 0$. 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 $c^{\text{inf}}_{1-\alpha} = z_{1-\alpha^{1/M}/2}$ follows from Theorem 1.2. In this special case, the asymptotic size is a decreasing function of the number of undersmoothing sequences M_1, and is equal to α when $|\nu(m)| = 0$ for all m, similar to Corollary 1.2-1. Second, the asymptotic size is an increasing function of bias term $|\nu(m)|$, as $F(c, |\nu|)$ is an increasing function of $|\nu|$. Third, (1.4.6) also gives the bound of the asymptotic size similar to Corollary 1.2-2, as $\prod_{m=1}^{M} F(c, |\nu(m)|) \leq F(c, \inf |\nu(m)|) = F(c, 0)$. Using this upper bound, equation (1.4.7) shows that the test based on $\text{Inf}\, T_n(\theta_0)$ with normal critical value controls size asymptotically.

Note that the asymptotic size result in (1.4.7) relies on the inequality $\text{Inf}\, T_n(\theta_0) \leq |T_n(\bar{K}, \theta_0)|$ and the fact that $T_n(\bar{K}, \theta_0) \xrightarrow{d} N(0, 1)$ under Assumption 1.4. If we know that $\bar{K}$ satisfies undersmoothing condition and others not, then there's no point of searching over different K; we may just use $\bar{K}$ for the inference. This may work well if $\bar{K}$ coincides with some infeasible size-optimal sequence $K^*(n)$ that minimizes $|P(T_n(K, \theta_0) > z_{1-\alpha/2}) - \alpha|$. However, in practice, choice of $\bar{K}$ can be ad hoc. Heuristically, if we use too large $\bar{K}$, then the power of the test based on $T_n(\bar{K}, \theta)$ with the normal critical value can be low, as $T_n(\bar{K}, \theta)$ can be very small with large variance $V_{\bar{K}}$ under alternatives. However, the test based on $\text{Inf}\, T_n(\theta_0)$ and its asymptotic critical value $c^{\text{inf}}_{1-\alpha}$ may have better power, as this test compare with the smaller critical value than the normal critical value. Further, our theory still provides the bound of the asymptotic size in (1.4.7) without any undersmoothing conditions on $K \in \mathcal{K}_n$, as $F(z_{1-\alpha/2}, \inf_m |\nu(m)|)$. Asymptotic distribution result in Corollary 1.1-2 is still valid, as long as at least one $\nu(m)$ is bounded, i.e., $|\nu(M)| = O(1)$.

In sum, $\text{Inf}\, T_n(\theta)$ leads to the tests that control the asymptotic size or bound the size distortions. One possible concern is that the low power property of the test using $\text{Inf}\, T_n(\theta)$ compare with the other statistics (e.g., the supremum of the t-statistics). Investigating local power comparisons of the level α test based on the other statistics, and the effect of asymptotic bias on subsequent power function are very important, but these are beyond the scope of this paper. However, I want to emphasize that inference based on the other transformation of the t-statistics can be highly sensitive to the bias problems, thus may lead to over-rejection of the test (see Section 1.14 for the inference based on the supremum test statistic). I will discuss the length of CIs based on the infimum test statistic in Section 1.5 and calibrate the length of CIs in various simulation setup in Section 1.8.

1.4.3 Critical Values

In this section, I discuss detail descriptions to approximate critical value defined in (1.4.2). Here, I suggest using simple Monte Carlo method to obtain critical value. To make implementation procedures simple and feasible, I impose following set assumption and conditional homoskedasticity.

Assumption 1.5. *(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* $K_m = \lfloor \pi_m \bar{K} \rfloor$ *for constant* π_m, $0 < \underline{\pi} = \pi_1 < \pi_2 < \cdots < \pi_M = 1$, *and fixed* M.

Assumption 1.6. *(Conditional homoskedasticity)* $E(\varepsilon_i^2 | x_i = x) = \sigma^2$.

Assumption 1.5 is a finite dimensional version of Assumption 1.1, and is different with an alternative set (Assumption 1.4) that considers different rate of Ks. As we have shown in Theorem 1.1, if Assumptions 1.2, 1.3, 1.5 and 1.6 are satisfied, then following finite

dimensional convergence of the t-statistics holds

$$(T_n(K_1,\theta_0),\cdots,T_n(K_M,\theta_0))' \xrightarrow{d} Z=(Z_1,\cdots,Z_M)', \quad Z\sim N(0,\Sigma), \tag{1.4.8}$$

where Σ is a variance-covariance matrix defined in (1.3.7), $\Sigma_{jl} = \lim_{n\to\infty} V_{K_j}^{1/2}/V_{K_l}^{1/2}$ for any $j<l$. Under Assumptions 1.2, 1.3 and 1.4, (1.4.8) also holds with $\Sigma = I_M$ follows by Theorem 1.2. 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

$$\begin{aligned}
&\widehat{V}_K = P_K(x)'\widehat{Q}_K^{-1}\widehat{\Omega}_K\widehat{Q}_K^{-1}P_K(x),\\
&\widehat{Q}_K = \frac{1}{n}\sum_{i=1}^n P_{Ki}P_{Ki}', \quad \widehat{\Omega}_K = \frac{1}{n}\sum_{i=1}^n P_{Ki}P_{Ki}'\widehat{\varepsilon}_{Ki}^2.
\end{aligned} \tag{1.4.9}$$

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

$$\begin{aligned}
&\widehat{c}_{1-\alpha}^{\inf} \equiv (1-\alpha) \text{ quantile of } \inf_{m=1,\cdots,M}|Z_{m,\widehat{\Sigma}}|,\\
&\text{where } Z_{\widehat{\Sigma}} = (Z_{1,\widehat{\Sigma}},\cdots,Z_{M,\widehat{\Sigma}})' \sim N(0,\widehat{\Sigma}), \quad \widehat{\Sigma}_{jj}=1, \widehat{\Sigma}_{jl} = \widehat{V}_{K_j}^{1/2}/\widehat{V}_{K_l}^{1/2}.
\end{aligned} \tag{1.4.10}$$

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 $\{\operatorname{Inf} T_n^b = \inf_m |Z_{m,\widehat{\Sigma}}^b| : b=1,\cdots,B\}$.[3]

I impose following additional assumption about the variance estimator $\widehat{V}_K$ to show the validity of using Monte-Carlo simulation critical values.

Assumption 1.7. $\sup_{K\in\mathcal{K}_n} |\frac{\widehat{V}_K}{V_K} - 1| = o_p(1)$ *as* $n, K\to\infty$.

[3]Conditional homoskedasticity assumption is only for a simpler implementation. Based on the general covariance function defined in (1.3.5), we can construct $\widehat{\Sigma}$ under general heteroskedastic error; $\widehat{\Sigma}_{j,l} = \frac{\widehat{V}_{K_{jl}}}{\widehat{V}_{K_j}^{1/2}\widehat{V}_{K_l}^{1/2}}$ 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 1.7 holds under the regularity conditions (Assumption 1.2) with an additional assumption. I discuss sufficient conditions to this holds in the Section 1.12 (see proof of Corollary 1.3).

Next, we consider following t-statistic $T_{n,\widehat{V}}(K,\theta)$ replacing variance of the series estimator V_K with $\widehat{V}_K$

$$T_{n,\widehat{V}}(K,\theta_0) \equiv \frac{\sqrt{n}(\widehat{\theta}_K - \theta_0)}{\widehat{V}_K^{1/2}}. \tag{1.4.11}$$

Following corollary shows the first-order joint asymptotic distribution of $T_{n,\widehat{V}}(K,\theta_0)$ to those of $T_n(K,\theta_0)$ in (1.4.8) for $K \in \mathcal{K}_n$. It also provides the validity of Monte Carlo critical values $\widehat{c}^{\inf}_{1-\alpha}$ defined in (1.4.10).

Corollary 1.3. *Under Assumptions 1.2, 1.3,1.5, 1.6 and 1.7,* $\widehat{c}^{\inf}_{1-\alpha} \xrightarrow{p} c^{\inf}_{1-\alpha}$ *holds where* $\widehat{c}^{\inf}_{1-\alpha}$ *are defined in* (1.4.10) *and* $c^{\inf}_{1-\alpha}$ *are the* $(1-\alpha)$ *quantile of the asymptotic null distribution* $\inf_{m=1,\cdots,M} |Z_m|$ *with* $Z = (Z_1,\cdots,Z_M)' \sim N(0,\Sigma)$, Σ *defined in* (1.3.7). *This also holds under Assumptions 1.2, 1.3 and 1.4 with* $\Sigma = I_M$.

Alternatively, we can use weighted bootstrap method to approximate asymptotic critical value. 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

$$\begin{aligned} &\hat{\beta}^b_K = \arg\min_b \frac{1}{n}\sum_{i=1}^n \omega_i (y_i - P'_{Ki} b)^2, \quad \hat{g}^b_K(x) = P_K(x)'\hat{\beta}^b_K, \\ &T^b_n(K) = \frac{\sqrt{n}(\hat{g}^b_K(x) - \hat{g}_K(x))}{\widehat{V}_K^{1/2}}. \end{aligned} \tag{1.4.12}$$

Then, construct $\operatorname{Inf} T^b_n = \inf_K |T^b_n(K)|$. Repeat this B times (1000 or 5000) and define $\widehat{c}^{\inf,WB}_{1-\alpha}$ as conditional $1-\alpha$ quantile of $\{\operatorname{Inf} T^b_n : b = 1,\cdots,B\}$ given the data. The idea behind the

weighted bootstrap methods may work 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, validity of the weighted bootstrap is beyond the scope of this paper and will be pursued for the future work.

1.5 Confidence Intervals

Now, I introduce CIs 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 $H_1 : \theta \neq \theta_0$. Thus, we collect all values of θ where the test statistic $\text{Inf}\, T_n(\theta)$ defined in Section 1.4 does not exceed its critical value. I first define $CI^{\texttt{Robust}}_{\texttt{inf}}$ based on $\text{Inf}\, T_n(\theta)$ and critical value $\widehat{c}^{\,\text{inf}}_{1-\alpha}$ defined in Section 1.4.3.

$$\begin{aligned} CI^{\texttt{Robust}}_{\texttt{inf}} &\equiv \{\theta : \inf_{K\in\mathcal{K}_n} |T_{n,\widehat{V}}(K,\theta)| \leq \widehat{c}^{\,\text{inf}}_{1-\alpha}\} \\ &= \{\theta : |T_{n,\widehat{V}}(K,\theta)| > \widehat{c}^{\,\text{inf}}_{1-\alpha}, \forall K\}^C = \bigcup_{K\in\mathcal{K}_n} \{\theta : |T_{n,\widehat{V}}(K,\theta)| \leq \widehat{c}^{\,\text{inf}}_{1-\alpha}\} \\ &= [\inf_K(\widehat{\theta}_K - \widehat{c}^{\,\text{inf}}_{1-\alpha} s(\widehat{\theta}_K)), \sup_K(\widehat{\theta}_K + \widehat{c}^{\,\text{inf}}_{1-\alpha} s(\widehat{\theta}_K))] \end{aligned} \tag{1.5.1}$$

where $s(\widehat{\theta}_K) \equiv \sqrt{\widehat{V}_K/n}$ is a standard error of series estimator 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 critical value $\widehat{c}^{\,\text{inf}}_{1-\alpha}$. $CI^{\texttt{Robust}}_{\texttt{inf}}$ can be constructed as the lower and the upper end point of confidence intervals for all $K \in \mathcal{K}_n$ using $\widehat{c}^{\,\text{inf}}_{1-\alpha}$.

Note that the last equality in (1.5.1) holds only when there is no dislocated CI, i.e., intersection is nonempty for any two CIs using $\widehat{c}^{\,\text{inf}}_{1-\alpha}$. 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. However, in general, there is no guarantee that the union of the confidence intervals are connected. Dislocated confidence interval may

show some evidence of bias for the specific model, but using superset can widen $CI^{\texttt{Robust}}_{\texttt{inf}}$ in this case. Although this paper does not consider data-dependent set $\mathcal{K}_n$, i.e., data-dependent choice of $\underline{K}$ and $\bar{K}$, possible large length of CI can be avoidable if $\underline{K}$ is reasonably large and this is exactly the condition needed in the next Corollary 1.4 to have a correct coverage. I will also discuss the coverage property of $CI^{\texttt{Robust}}_{\texttt{inf}}$ even with large bias of some models. Note that possible large length of the $CI^{\texttt{Robust}}_{\texttt{inf}}$ is also related to the possible low power property of the test.

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

$$\begin{aligned} CI_{\texttt{inf}} &\equiv \{\theta : \inf_{K\in\mathcal{K}_n} |T_{n,\widehat{V}}(K,\theta)| \le z_{1-\alpha/2}\} \\ &= [\inf_K(\widehat{\theta}_K - z_{1-\alpha/2}s(\widehat{\theta}_K)), \sup_K(\widehat{\theta}_K + z_{1-\alpha/2}s(\widehat{\theta}_K))] \end{aligned} \tag{1.5.2}$$

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

Next Corollary shows valid coverage property of the above CIs, and it follows from Corollary 1.2 and 1.3.

Corollary 1.4. *1. Under Assumptions 1.2, 1.3, 1.5, 1.6, and 1.7,*

$$\liminf_{n\to\infty} P(\theta_0 \in CI^{Robust}_{inf}) = 1-\alpha, \qquad \liminf_{n\to\infty} P(\theta_0 \in CI_{inf}) \ge 1-\alpha \tag{1.5.3}$$

2. Under Assumptions 1.2, 1.5, 1.6, 1.7, and $\sup_m |\nu(m)| < \infty$ *where* $\nu(m) \equiv \nu(\pi_m)$ *as in* (1.3.3),

$$\liminf_{n\to\infty} P(\theta_0 \in CI^{Robust}_{inf}) \ge 1 - F(c^{\inf}_{1-\alpha}, \inf_m |\nu(m)|) \tag{1.5.4}$$

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{inf}) \ge 1 - F(z_{1-\alpha/2}, \inf_m |\nu(m)|) \tag{1.5.5}$$

3. Under Assumptions 1.2, 1.4, and 1.7,

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{inf}^{Robust}) = 1 - \prod_{m=1}^{M} F(c_{1-\alpha}^{\inf}, |\nu(m)|) \tag{1.5.6}$$

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{inf}) \geq 1 - \alpha \tag{1.5.7}$$

Corollary 1.4-1 shows the validity of $CI_{\texttt{inf}}^{\texttt{Robust}}$ and $CI_{\texttt{inf}}$, i.e., asymptotic coverage of CIs are greater than or equal to $1-\alpha$. Note that the Corollary 1.4-1 requires undersmoothing condition, i.e., there is no asymptotic bias for all Ks in $\mathcal{K}_n$.

Without undersmoothing condition, Corollary 1.4-2 and 1.4-3 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 smallest bias, similarly to the asymptotic size results in Corollary 1.2. Equation (1.5.6) also implies that the asymptotic coverage of $CI_{\texttt{inf}}^{\texttt{Robust}}$ does not affected by oversmoothing sequence (small K_m) such that $|\nu(m)| = \infty$, as $F(c, \infty) = 1$. Similar to the asymptotic size results, the asymptotic coverage is equal to $1 - \alpha^{M_1/M}$ when $|\nu(m)| = \infty$ for $m = 1, \cdots, M - M_1$ and $|\nu(m)| = 0$ for the others, and is an increasing function of the number of undersmoothing sequences. The asymptotic coverage is equal to $1-\alpha$ when $|\nu(m)| = 0$ for all m. Although the coverage is a decreasing function of $|\nu(m)|$, (1.5.6) implies that it is bounded below by $1 - F(c_{1-\alpha}^{\inf}, 0)$. Furthermore, (1.5.7) shows that $CI_{\texttt{inf}}$ using normal critical value achieve nominal coverage probability $1-\alpha$. $CI_{\texttt{inf}}$ and $CI_{\texttt{inf}}^{\texttt{Robust}}$ bound coverage distortions even when asymptotic bias terms are present for several Ks in a set, in this sense they 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 (1.5.3) and (1.5.7) holds with inequality, not equality. Therefore, 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 the $\bar{K}$. In contrast, $CI_{\texttt{inf}}^{\texttt{Robust}}$ may have shorter length by using smaller critical value than the normal critical value.

1.6 Post-Model Selection Inference

In this section, I provide methods to construct a valid CI that gives correct coverage even after selecting he number of series terms using any type of selection rules.

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

$$|T_n(\widehat{K},\theta)|, \quad \widehat{K} \in \mathcal{K}_n \tag{1.6.1}$$

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

$$CI^{\texttt{Naive}}_{\texttt{pms}} \equiv \{\theta : |T_n(\widehat{K},\theta)| \leq z_{1-\alpha/2}\} = [\widehat{\theta}_{\widehat{K}} - z_{1-\alpha/2}s(\widehat{\theta}_{\widehat{K}}), \widehat{\theta}_{\widehat{K}} + z_{1-\alpha/2}s(\widehat{\theta}_{\widehat{K}})]. \tag{1.6.2}$$

The conventional method of using normal critical value in (1.6.2) 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 $\widehat{K}$ (e.g., $\widehat{K} = \widehat{K}_{\texttt{cv}}$ selected by cross-validation). 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 (or other asymptotically equivalent criteria such as AIC). 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 undersmoothing condition

does, then the asymptotic distribution may have bias terms and resulting naive CI may have large coverage distortions.

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

$$\text{Sup}\, T_n(\theta) \equiv \sup_{K \in \mathcal{K}_n} |T_n(K, \theta)|. \tag{1.6.3}$$

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 $\text{Sup}\, T_n(\theta_0) \xrightarrow{d} \xi_{\text{sup}} \equiv \sup_{\pi \in [\underline{\pi}, 1]} |\mathbb{T}(\pi)|$ under the same assumptions as in Corollary 1-1. 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 the $c^{\text{inf}}_{1-\alpha}$ defined in (1.4.2), I define asymptotic critical value $c^{\text{sup}}_{1-\alpha}$ as $1-\alpha$ quantile of ξ_{sup}. We can approximate this critical value by using Monte Carlo simulation based method similarly as in Section 1.4.3. To be specifically, I define

$$\widehat{c}^{\text{sup}}_{1-\alpha} \equiv (1-\alpha) \text{ quantile of } \sup_{m=1,\cdots,M} |Z_{m,\widehat{\Sigma}}|, \tag{1.6.4}$$

where $Z_{\widehat{\Sigma}} = (Z_{1,\widehat{\Sigma}}, \cdots, Z_{M,\widehat{\Sigma}})' \sim N(0, \widehat{\Sigma})$ and $\widehat{\Sigma}$ are defined in (1.4.10). Under the same assumptions as in Corollary 1.3, we can also verify $\widehat{c}^{\text{sup}}_{1-\alpha} \xrightarrow{p} c^{\text{sup}}_{1-\alpha}$.

Next, I define the following robust post-selection CI using the critical value $\widehat{c}^{\text{sup}}_{1-\alpha}$ rather than the normal critical value $z_{1-\alpha/2}$ compare to the $CI^{\texttt{Naive}}_{\texttt{pms}}$ in (1.6.1),

$$CI^{\texttt{Robust}}_{\texttt{pms}} \equiv [\widehat{\theta}_{\widehat{K}} - \widehat{c}^{\text{sup}}_{1-\alpha} s(\widehat{\theta}_{\widehat{K}}), \widehat{\theta}_{\widehat{K}} + \widehat{c}^{\text{sup}}_{1-\alpha} s(\widehat{\theta}_{\widehat{K}})], \quad \widehat{K} \in \mathcal{K}_n. \tag{1.6.5}$$

For example, we can construct $CI^{\texttt{Robust}}_{\texttt{pms}}$ with $\widehat{K}_{\texttt{cv}}$ selected by cross-validation among $\mathcal{K}_n$ using the critical value $\widehat{c}^{\text{sup}}_{1-\alpha}$.

Next Corollary shows that the robust post-selection $CI^{\texttt{Robust}}_{\texttt{pms}}$ guarantees asymptotic coverage as $1-\alpha$.

Corollary 1.5. *Under Assumptions 1.2, 1.3, 1.5, 1.6, and 1.7,*

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{pms}^{Robust}) \geq 1-\alpha. \tag{1.6.6}$$

Even though Corollary 1.5 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 type of data-dependent selection criteria or any selection rules researchers might want to use. Here, I impose an undersmoothing (Assumption 1.3) and therefore $CI_{\texttt{pms}}^{\texttt{Robust}}$ does not dealing with the bias problem explicitly. However, it accommodates bias by enlarging confidence interval using larger critical values $\widehat{c}_{1-\alpha}^{\text{sup}}$ than the normal critical value. Moreover, we also expect $\widehat{c}_{1-\alpha}^{\text{sup}}$ 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.

1.7 Extension: Partially Linear Model

In this section, I provide inference methods for the partially linear model (PLM) similar to those in 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. I consider following partially linear model

$$y_i = \theta_0 w_i + g_0(x_i) + \varepsilon_i, \qquad E(\varepsilon_i | w_i, x_i) = 0. \tag{1.7.1}$$

We are interested in inference on treatment/policy effect θ_0 after approximating unknown function $g_0(x_i)$ by regressors $p(x_i)$ among a set of potential control variables. Number of regressors could 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. Parametric part w_i is always included in the model, however, we are unsure which covariates/transformations of x_i should be used.

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

$$y_i = \theta_0 w_i + P_{Ki}'\beta_K + r_{Ki} + \varepsilon_i, \tag{1.7.2}$$

where the approximation error r_{Ki} is defined similarly as in Section 1.2. Then, similar to nonparametric regression model, 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}

$$\widehat{\theta}_K = (W'M_K W)^{-1} W'M_K Y \tag{1.7.3}$$

where $W = (w_1, \cdots, w_n)', M_K = I_K - P^K(P^{K'}P^K)^{-1}P^{K'}, P^K = [P_{K1}, \cdots, P_{Kn}]', Y = (y_1, \cdots, y_n)'$. Estimator for β_K is given by $(\widehat{\theta}_K, \widehat{\beta}_K')' = (H^{K'}H^{K'})^{-1}H^{K'}Y$ where $H^K = [W, P^K]$. For notational simplicity, I use the similar notation as defined in nonparametric regression setup.

The asymptotic normality and valid inference for the partially linear model has been developed in the literature. Donald and Newey (1994) derived the asymptotic normality of $\widehat{\theta}_K$ under standard rate conditions $K/n \to 0$. 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. 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 that grows as fast as sample size n, is particularly useful for our purpose. Under $K/n \to c$, the limiting normal distribution has a larger variance than the standard asymptotic variance derived under $K/n \to 0$, and

the adjusted variance depends on the number of terms K. 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 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 all asymptotically equivalent. This is also related to the well known results of the series based two-step semiparametric estimation (see Newey (1994b)). However, using the large sample approximation that allow the number of series can grow with the same rate of sample size, we can construct a joint distribution of the t-statistics with different sequence of models. This provides a useful approximation theory to fully account the dependency of the t-statistics with different Ks.

I impose an assumption that are same as in Cattaneo, Jansson, and Newey (2015a) uniformly over the model $K \in \mathcal{K}_n$, where $\mathcal{K}_n$ is same as in the Assumption 1.5. Let $v_i \equiv w_i - g_{w0}(x_i)$ where $g_{w0}(x_i) \equiv E[w_i|x_i]$. Then, by construction $E[v_i|x_i] = 0$.

Assumption 1.8. *(Regularity conditions for Partially Linear Model: Assumption PLM in Cattaneo, Jansson, and Newey (2015a))*

1. $\{y_i, w_i, x_i\}$ *are i.i.d random variables satisfying the model* (1.7.1).

2. *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* $E[v_i^4|x_i] \leq C$.

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

4. *For all* $K \in \mathcal{K}_n$, *there exists* γ_g, γ_{g_w},

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

Assumption 1.8 does not require $K/n \to 0$ which is required to get asymptotic normality in the literature (e.g., Donald and Newey (1994)). Assumption 1.8-(4) typically holds for the

polynomials and splines basis, similar to the nonparametric setup. For example, Assumption 1.8-(4) 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 1.5 and 1.8,

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_K - \theta) &= (\frac{1}{n}W'M_KW)^{-1}\frac{1}{\sqrt{n}}W'M_KY \\ &= \widehat{\Gamma}_K^{-1}(\frac{1}{\sqrt{n}}\sum_i v_i M_{ii}^K \varepsilon_i + \frac{1}{\sqrt{n}}\sum_{i=1}^n \sum_{j=1, j\neq i}^n v_i M_{ij}^K \varepsilon_j) + o_p(1)\end{aligned} \tag{1.7.4}$$

where $\widehat{\Gamma}_K = W'M_KW/n$. Note that under $K/n \to 0$, the leading term $\frac{1}{\sqrt{n}}\sum_i v_i M_{ii}^K \varepsilon_i = \frac{1}{\sqrt{n}}\sum_i v_i\varepsilon_i + o_p(1)$. Thus t-statistics $T_K(\theta)$ are asymptotically equivalent under any sequences $K \to \infty$ satisfying the standard rate conditions. However, under the faster rate conditions on K imposed here, the second term is not negligible and converges to bounded random variables. Cattaneo, Jansson, and Newey (2015a) apply 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 $T_n(K, \theta)$. Under Assumptions 1.5, 1.8 and undersmoothing condition $nK^{-2(\gamma_g + \gamma_{g_w})} \to 0$, we get following asymptotic distributional results for a deterministic sequence of K assuming conditional homoskedasticity.

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

where V_K coincides with the standard asymptotic variance formula 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 $K/n \nrightarrow 0$.

Next theorem is the main result for the partially linear model setup, analogous to nonparametric setup. Theorem 1.3 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 1.5 and 1.6.

Theorem 1.3. *Suppose Assumptions 1.5 and 1.8 hold. Also,* $n\bar{K}^{-2(\gamma_g+\gamma_{gw})} \to 0$ *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 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_{jl} *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 1 for* $j = l$. *Moreover, under Assumptions 1.5, 1.7 and 1.8, coverage probability holds for the following CIs*

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{inf}^{Robust}) = 1 - \alpha, \quad \liminf_{n\to\infty} P(\theta_0 \in CI_{inf}) \geq 1 - \alpha \tag{1.7.5}$$

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{pms}^{Robust}) \geq 1 - \alpha \tag{1.7.6}$$

where $CI_{inf}^{Robust}, CI_{inf}$, *and* CI_{pms}^{Robust} *are similarly defined as in Section 1.5 and 1.6 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 1.3 derives the joint asymptotic distribution of the $T_n(K, \theta_0)$ over $K \in \mathcal{K}_n$ for the parametric part in partially linear model. Note that the variance-covariance matrix Σ is same as in nonparametric model setup (see equation (1.3.7) or (1.4.8)). Variance-covariance matrix Σ_{jl} for any $j \neq l$ can be reduced under the condition $\bar{K}/n \to c$,

$$\Sigma_{jl} = \lim_{n\to\infty} \frac{V_{K_{j\wedge l}}^{1/2}}{V_{K_{j\vee l}^{1/2}}} = \lim_{n\to\infty} \frac{(1 - K_{j\wedge l}/n)^{-1/2} V^{1/2}}{(1 - K_{j\vee l}/n)^{-1/2} V^{1/2}} = \lim_{n\to\infty} \frac{(1 - \pi_{j\wedge l}\bar{K}/n)^{-1/2}}{(1 - \pi_{j\vee l}\bar{K}/n)^{-1/2}} = (\frac{1 - c\pi_{j\vee l}}{1 - c\pi_{j\wedge l}})^{1/2}. \tag{1.7.7}$$

Theorem 1.3 also shows the asymptotic coverage property of CIs similar to Corollary 1.4 in the nonparametric setup. The lower bounds of the asymptotic coverage for $CI^{\texttt{Robust}}_{\texttt{inf}}, CI_{\texttt{inf}}$ can be also derived without undersmoothing assumption ($n\bar{K}^{-2(\gamma_g+\gamma_{g_w})} \to 0$), thus omitted here.

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

$$\widehat{V}_K = s^2\widehat{\Gamma}_K^{-1}, \quad s^2 = \frac{1}{n-1-K}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2, \quad \widehat{\varepsilon}_i^2 = \sum_{j=1}^{n} M_{K,ij}(y_j - \widehat{\theta}_K w_j). \tag{1.7.8}$$

For consistency results and more discussions, see section 3.2 (Theorem 2) of Cattaneo, Jansson, and Newey (2015a) and also Cattaneo, Jansson, and Newey (2015b) under heteroskedasticity.

1.8 Simulations

This section investigates the small sample performance of the proposed methods in Sections 1.5 and 1.6. We are mainly interested in empirical coverage of CIs for the true value of $g(x)$ over the support of x for various functions $g(x)$ and different basis.

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

$$\begin{aligned} y_i &= g(x_i) + \varepsilon_i, \\ x_i = \Phi(x_i^*), \begin{pmatrix} x_i^* \\ \varepsilon_i \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right) \end{aligned} \tag{1.8.1}$$

where $\Phi(\cdot)$ is the standard normal cdf 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 standard normal pdf. The

functions $g_1(x)$ and $g_2(x)$ are used in Newey and Powell (2003), Chen and Christensen (2015) and we label them as linear and nonlinear designs. $g_3(x)$ and $g_4(x)$ are rescaled version of Hall and Horowitz (2013), and we denote these as highly nonlinear designs. See Figure 1.1 for the shape of all 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 from my experience.

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 with different orders and quadratic splines with evenly placed knots. In either case, K denotes the 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 pointwise coverage properties 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 1.5. Finally, I examine robust post-selection CI, $CI^{\texttt{Robust}}_{\texttt{pms}}$ with $\widehat{K}_{\texttt{cv}}$, defined in Section 1.6. The critical values, $\widehat{c}^{\,\inf}_{1-\alpha}$ and $\widehat{c}^{\,\sup}_{1-\alpha}$ are constructed using the Monte-Carlo method described in Sections 1.4.3 and Section 1.6.

Figure 1.3 reports nominal 95% coverage probability of all five CIs. Overall, $CI^{\texttt{Robust}}_{\texttt{inf}}$ performs very well across the different simulation designs. Its empirical coverage is close to the nominal 95% level at many points over the support. $CI_{\texttt{inf}}$ using normal critical value also performs well, as coverage is no less than the nominal level at almost all points. However, $CI_{\texttt{inf}}$ 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 value. $CI_{\texttt{maxK}}$ slightly undercovers at many points, and works quite poorly especially at the boundary. $CI^{\texttt{Robust}}_{\texttt{pms}}$ with the adjustment of using larger critical value $\widehat{c}^{\,\text{sup}}_{1-\alpha}$ than normal critical value seems also work well, but does not solve bias problem completely (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 very 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 1.3-(a) for $CI_{\texttt{maxK}}$ are mostly due to the small sample size. However, given the small sample size, coverage $CI^{\texttt{Robust}}_{\texttt{inf}}$ is still fairly close to 95%.

For the slightly nonlinear function $g_2(x)$, coverage of all confidence intervals except $CI_{\texttt{inf}}$ is less than 0.95 at some 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 is consistent with our theory that $CI^{\texttt{Robust}}_{\texttt{inf}}$ bounds the size distortions even when there are large biases for all polynomial approximations over $K \in \mathcal{K}_n$. In 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 large bias. Possibly poor coverage property at this point was also described in Hall and Horowitz (2013, Figure 3). In this case, regression spline seems much better for approximating this local point. Figure 1.4 shows the coverage probability of CIs using quadratic splines with different number of knots. As we can see from Figure 1.4, $CI^{\texttt{Robust}}_{\texttt{inf}}$ with splines works better to achieve correct coverage for $g_2(x = 0.4), g_4(x = 0.5)$, and for other different functions as well.

In Figure 1.5, I compare the length of the five CIs for the polynomial series. In the linear and nonlinear designs, rank of the length in a narrower order is as follows; $CI^{\texttt{Naive}}_{\texttt{pms}} < CI^{\texttt{Robust}}_{\texttt{pms}} \leq CI^{\texttt{Robust}}_{\texttt{inf}} < CI_{\texttt{maxK}} < CI_{\texttt{inf}}$. This is what we expected as $CI^{\texttt{Naive}}_{\texttt{pms}}$ is too narrow, $CI_{\texttt{maxK}}$ is somewhat wide because of large variance using $\bar{K}$. For the highly nonlinear design, $CI^{\texttt{Robust}}_{\texttt{inf}}$ and $CI_{\texttt{inf}}$ become wider at some points where estimates are relatively sensitive

across K. Length of $CI_{\texttt{maxK}}$ is similar for $g_3(x)$ or shorter for $g_4(x)$ compare than $CI^{\texttt{Robust}}_{\texttt{inf}}$. Figure 1.6 compares the length of CIs for the splines, and it shows similar patterns with polynomial approximation. Given that $CI^{\texttt{Robust}}_{\texttt{inf}}$ has a similar or only a slightly wider length than the others, we want to highlight that it has better or similar coverage probability at most points than $CI_{\texttt{maxK}}, CI^{\texttt{Naive}}_{\texttt{pms}}$ and $CI^{\texttt{Robust}}_{\texttt{pms}}$, as in Figure 1.4.

We expect 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^*: It 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 $CI_{\texttt{maxK}}$ may be too wide because of large variance, or the normal distribution may be a poor approximation with $\bar{K}$ in small sample size. In contrast, $CI^{\texttt{Robust}}_{\texttt{inf}}$ and $CI_{\texttt{inf}}$ is least affected with those small K with large bias, and performs quite well even in small sample size.

In sum, $CI^{\texttt{Robust}}_{\texttt{inf}}$ seems to work well in various simulation experiments. It is the only method close to nominal coverage and it is least affected by biases. $CI_{\texttt{inf}}$ also performs well, but it can be conservative. In some simulation results, coverage of $CI^{\texttt{Robust}}_{\texttt{pms}}$ is close to the nominal level, thus it is also advisable to report.

In addition to length comparisons, I also provide power of the different test statistics. In Figure 1.7, 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. Of course, the power depends on 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 following rejection probability based on $\operatorname{Inf} T_n(\theta), \operatorname{Sup} T_n(\theta)$, 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) $P(\operatorname{Inf} T_n(\theta_0 + \delta) > \widehat{c}^{\inf}_{1-\alpha})$; (3) $P(\operatorname{Inf} T_n(\theta_0 + \delta) > z_{1-\alpha/2})$; (4) $P(\operatorname{Sup} T_n(\theta_0 + \delta) > \widehat{c}^{\sup}_{1-\alpha})$; (5) $P(|T_n(\widehat{K}_{\texttt{cv}},$

$\theta_0 + \delta)| > \widehat{c}^{\sup}_{1-\alpha})$. As expected, Figure 1.7-(a) and (b) show that the tests based on $\text{Inf}\, T_n(\theta)$ are the only method to control size or bound the size distortions when bias exists for some Ks.

In Figure 1.8, I report typical t-statistic patterns as a function of number of series K. Specifically, I plot $E[T_n(K, \theta_0)]$ evaluated at the true value $\theta_0 = g_0(x)$ as a function of K. I also calculate $K_{\texttt{inf}} \equiv \arg\min_K |T_n(K, \theta_0)|$ that minimizes t-statistics evaluated at the unknown true function $g_0(x)$, which is infeasible, but feasible in simulations. I also plot the median values of $\widehat{K}_{\texttt{cv}}$ selected by cross-validation as vertical lines. We can easily see that typical t-statistic patterns shows asymmetric V-shape: decrease rapidly with K, but increases slightly. Moreover minimizer of the $\text{Inf}\, T_n(\theta)$ is not always coincide with the largest K in $\mathcal{K}_n$. This is because series estimation with large K also leads to unreliable estimates due to the estimation variance similar to the problem of using too small bandwidth in kernel estimation. In each simulations, $K_{\texttt{inf}}$ is likely to be larger than $\widehat{K}_{\texttt{cv}}$ when bias are large, and $K_{\texttt{inf}}$ is equal or slightly larger than $\widehat{K}_{\texttt{cv}}$ when bias terms are small.

1.9 Illustrative Empirical Application : Nonparametric Estimation of Labor Supply Function and Wage Elasticity with Nonlinear Budget Set

In this section, I illustrate robust inference procedures by revisiting a paper by Blomquist and Newey (2002). 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 only requires estimated variance for different specifications that are reported in the table of Blomquist and Newey (2002). It is quite straightforward to construct the proposed CI without any replication of the data sets in this case and this is

one of the computational advantages of our procedure.

Understanding how tax and policy affect individual labor supply has been central issues in labor economics (see Hausman (1985) and Blundell and MaCurdy (1999), among many others). Focusing on the conditional mean of hours of work given the individual budget set, Blomquist and Newey (2002) estimate labor supply function 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 (e.g. maximum likelihood estimation (MLE)).

Specifically, they consider following models by exploiting additive structure follows from the utility maximization with piecewise linear budget sets.

$$h_i = g(x_i) + \varepsilon_i, \quad E(\varepsilon_i | x_i) = 0, \tag{1.9.1}$$

$$g(x_i) = g_1(y_J, w_J) + \sum_{j=1}^{J-1} [g_2(y_j, w_j, \ell_j) - g_2(y_{j+1}, w_{j+1}, \ell_j)], \tag{1.9.2}$$

where h_i is the hours of the ith individual and $x_i = (y_1, \cdots, y_J, w_1, \cdots, w_J, \ell_1, \cdots, \ell_J)$ is the budget set that can be represented by 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, I use the similar notations with theirs. Equation (1.9.2) for the conditional mean function follows from Theorem 2.1 of Blomquist and Newey (2002), and this additive structure greatly reduce dimensionality. They consider following power series for $g(x)$

$$p_k(x) = (y_J^{p_1(k)} w_J^{q_1(k)}, \sum_{j=1}^{J-1} \ell_j^{m(k)} (y_j^{p_2(k)} w_j^{q_2(k)} - y_{j+1}^{p_2(k)} w_{j+1}^{q_2(k)})). \tag{1.9.3}$$

Using the data from the Swedish "Level of Living" survey in 1973, 1980 and 1990, they pool the data from three waves and use the data from 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

$$E_w = \bar{w}/\bar{h}[\frac{\partial g(w,\cdots,w,\bar{y},\cdots,\bar{y})}{\partial w}]|_{w=\bar{w}}, \tag{1.9.4}$$

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

Table 1.1 is exactly 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 each row. For example, estimates in the second raw use the term in the first row $(1, y_J, w_J)$ with additional terms $(\Delta y, \Delta w)$. Here, $\ell^m \Delta y^p w^q$ denotes approximating term $\sum_i \ell_j^m (y_j^p w_j^q - y_{j+1}^p w_{j+1}^q)$. They also report cross-validation criteria, CV, for each model specification. In their formula, series terms are chosen to maximize CV, which minimizes asymptotic MSE. In addition to their original table, I also report CI for each specification. As we can see from the table, it is ambiguous which large K should be used for the inference. We do not have compelling reason 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 $CI_{\texttt{inf}}$, $CI^{\texttt{Robust}}_{\texttt{pms}}$ defined in Sections 1.5 and 1.6. One nice feature of the new method is that we can construct critical values and CIs without any replication of the data under homoskedastic assumption. Monte Carlo methods defined in (1.4.10) only requires variance estimates, thus we can simply construct critical value from estimated standard errors. If we have the dataset, then we could also implement critical value based on general variance forms under heteroskedasticity or bootstrap critical value. Using Monte-Carlo method, estimated critical values are $\widehat{c}^{\text{inf}}_{1-\alpha} = 0.9668, \widehat{c}^{\text{sup}}_{1-\alpha} = 2.4764$, respectively.

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, $CI = [0.0273, 0.1045]$ using all the additional terms up to the 6th row. Morover, $CI^{\texttt{Robust}}_{\texttt{inf}}$ is substantially tighter than

$CI_{\texttt{maxK}} = [0.0148, 0.1280]$ using the largest number of series terms $\bar{K}$ 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 naive confidence interval to $[0.0169, 0.0916]$.

1.10 Conclusion

This paper considers the construction of inference methods with data-dependent number of series terms in 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-statistic sequences 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 least 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 methods in the partially linear model setup.

1.11 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.

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.

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.

Newey, W. K. (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica*, 58(4), 809-837.

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.

Newey, W. K., F. Hsieh, J. Robins (2003): "Undersmoothing and Bias Corrected Functional Estimation," *Working Paper*.

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

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.

1.12 Proofs

In this section, we define additional notations for the empirical process theory used in the proof of Theorem 1.1. Given measurable space $(S, \mathcal{S})$, let $\mathcal{F}$ as a class of measurable functions $f : \mathcal{S} \to \mathbb{R}$. We define $N(\varepsilon, \mathcal{F}, L_2(Q))$ 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 $||f||_{Q,2} = (\int |f|^2 dQ)^{1/2}$ and measure Q. Uniform entropy numbers relative to L_2 are defined as $\sup_Q \log N(\epsilon||F||_{Q,2}, \mathcal{F}, L_2(Q))$ where 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 to discuss 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 universal constant that does not depend on n.

Proof of Theorem 1.1

For any sequence $\{K = \lfloor \pi \bar{K} \rfloor : n \geq 1\} \in \prod_{n=1}^{\infty} \mathcal{K}_n$ under Assumptions 1.1 and 1.2, we first define orthonormalized vector of basis functions $\tilde{P}_K(x)$.

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

where $\tilde{P}_{Ki} = \tilde{P}_K(x_i)$ is similarly defined as in the original basis functions. We observe that

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

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 as unknown and deal with non-orthonormalized series terms here.

Next, we re-define, with abuse of notation, pseudo-true value β_K in (1.2.2) with 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 $\varepsilon_{Ki} = r_{Ki} + \varepsilon_i$, $r_K(x) = g_0(x) - \tilde{P}_K(x)'\beta_K$. Define $r_K \equiv (r_{K1}, \cdots r_{Kn})', r_{Ki} = r_K(x_i)$. 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 1.1. Versions of proof of Lemma 1.1 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 1.1. Note that different rate conditions can be used in Assumption 1.2, but lead to different bounds in (1.12.1)-(1.12.3) for the following Lemma 1.1.

Lemma 1.1. *Under Assumptions 1.1 and 1.2, for any $K \in \mathcal{K}_n$, following holds*

$$||\widehat{Q}_K - I_K|| = O_p(\sqrt{\frac{\zeta_K^2\lambda_K^2 \log K}{n}}), \tag{1.12.1}$$

$$R_1(K) \equiv \sqrt{\frac{1}{nV_K}}\tilde{P}_K(x)'\left(\widehat{Q}_K^{-1} - I_K\right)\tilde{P}^{K\prime}(\varepsilon + r_K) = O_p(\sqrt{\frac{\lambda_K^2\zeta_K^2 \log K}{n}}(1 + \ell_K c_K\sqrt{K})), \tag{1.12.2}$$

$$R_2(K) \equiv \sqrt{\frac{1}{nV_K}}\tilde{P}_K(x)'\tilde{P}^{K\prime}r_K = O_p(\ell_K c_K). \tag{1.12.3}$$

To provide (1.12.1) in Lemma 1.1, we first introduce matrix Bernstein inequality in Tropp (2015). See also Lemma 2.1 of Chen and Christensen (2015b).

Lemma 1.2 (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,

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

Proof of Lemma 1.1.

To provide bound in (1.12.1), we apply Lemma 1.2 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 Lemma 1.2, 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 (1.12.2), 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,

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

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 1.2-(2), $||\widehat{Q}_K - I_K|| = O_P(\sqrt{\lambda_K^2 \zeta_K^2 \log(K)/n})$ by (1.12.1) and $||\widehat{Q}_K^{-1}|| \lesssim 1$ by Assumption 1.2-(4). Then by Chebyshev's inequality, we have that

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

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

$$E[||\frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{P}_{Ki} r_{Ki}]||^2] = E[\sum_{j=1}^K \tilde{P}_{ji}^2 r_{Ki}^2] \leq \ell_K^2 c_K^2 E[||\tilde{P}_{Ki}||^2] = \ell_K^2 c_K^2 K. \tag{1.12.4}$$

Combining (1.12.1) and (1.12.4) yields the results

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

by Assumption 1.2-(4).

We now prove (1.12.3). Consider $\sqrt{\frac{1}{nV_K}} \tilde{P}_K(x)' \tilde{P}^{K\prime} r_K$,

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

since $E[(\frac{\tilde{P}_K(x)' \tilde{P}_{Ki}}{V_K^{1/2}})^2] \asymp 1$ by Assumption 1.2-(2) and $E(r_{Ki})^2 \leq \ell_K c_K$ by Assumption 1.2-(3). Therefore, we have (1.12.3) by Chebyshev's inequality and $E[\tilde{P}_{Ki} r_{Ki}] = 0$. This completes the proof. *Q.E.D.*

Proof of Theorem 1.1. For any $\pi \in \Pi = [\underline{\pi}, 1]$, we first show the decomposition of the t-

statistic in equation (1.3.2).

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

where $R_1(K), R_2(K)$ are defined in (1.12.2), (1.12.3).

By Lemma 1.1, 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 $K = \lfloor \pi \bar{K} \rfloor \in \mathcal{K}_n$ under Assumptions 1.1 and 1.2. Therefore we have following decomposition for any $\pi \in \Pi$

$$T_n^*(\pi,\theta_0) = t_n^*(\pi) - \sqrt{n} V_\pi^{-1/2} r_\pi + o_p(1), \tag{1.12.5}$$

where

$$t_n^*(\pi) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\tilde{P}_\pi(x)' \tilde{P}_{\pi i} \varepsilon_i}{V_\pi^{1/2}}. \tag{1.12.6}$$

Now we show weak convergence of the empicial process $\{t_n^*(\cdot) : n \geq 1\}$ to the mean zero Gaussian process $\mathbb{T}(\cdot)$ defined in the Theorem 1.1. Let $\mathcal{F}_n = \{f_{n,\pi} : \pi \in \Pi\}$ be a sequence of classes of measurable functions $f_{n,\pi} : (\mathcal{E} \times \mathcal{X})$ to $\mathbb{R}$ indexed by π,

$$f_{n,\pi}(\varepsilon, t) = \frac{\tilde{P}_\pi(x)' \tilde{P}_\pi(t) \varepsilon}{V_\pi^{1/2}(x)} = \frac{\tilde{P}_{\lfloor \bar{K} \pi \rfloor}(x)' \tilde{P}_{\lfloor \bar{K} \pi \rfloor}(t) \varepsilon}{V_{\lfloor \bar{K} \pi \rfloor}^{1/2}(x)}, (\varepsilon, t) \in \mathcal{E} \times \mathcal{X}. \tag{1.12.7}$$

Consider empirical process $\{t_n^*(\pi) : \pi \in \Pi\} = \{n^{-1/2} \sum_{i=1}^n f_{n,\pi}(\varepsilon_i, x_i) : \pi \in \Pi\}$ indexed by classes of functions $\mathcal{F}_n = \{f_{n,\pi} : \pi \in \Pi\}$. We want to show weak convergence of the stochastic

process 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 Gaussian process 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

$$Ef_{n,\pi_1} f_{n,\pi_2} - Ef_{n,\pi_1} Ef_{n,\pi_2} = \frac{\tilde{P}_{\pi_1}(x)' E(\tilde{P}_{\pi_1}(x_i)\tilde{P}_{\pi_2}(x_i)'\varepsilon_i^2)\tilde{P}_{\pi_2}(x)}{V_{\pi_1}^{1/2} V_{\pi_2}^{1/2}}. \tag{1.12.8}$$

This term converges to the claimed covariance function $\Sigma(\pi_1, \pi_2)$. This covariance kernel can be bounded below and above some constant $0 < c, C < \infty$ for all n,

$$c \leq \underline{\sigma}^2 \frac{V_{\pi_1}^{1/2}}{V_{\pi_2}^{1/2}} \leq \frac{\tilde{P}_{\pi_1}(x)' E(\tilde{P}_{\pi_1}(x_i)\tilde{P}_{\pi_2}(x_i)'\varepsilon_i^2)\tilde{P}_{\pi_2}(x)}{V_{\pi_1}^{1/2} V_{\pi_2}^{1/2}} \leq \bar{\sigma}^2 \frac{V_{\pi_1}^{1/2}}{V_{\pi_2}^{1/2}} \leq C \tag{1.12.9}$$

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 1.2-(2). We also use the fact that $V_{\pi_1}^{1/2} \asymp V_{\pi_2}^{1/2} \asymp ||\tilde{P}_{\bar{K}}||$ for any π_1, π_2 under Assumption 1.1 and 1.2. Thus, there exist $c, C > 0$ which is independent of n, π such that $0 < c \leq \frac{V_{\pi_1}^{1/2}}{V_{\pi_2}^{1/2}} \leq C < 1$.

We first show the finite dimensional convergence. It suffices to show that

$$\begin{pmatrix} t_n^*(\pi_1) \\ t_n^*(\pi_2) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & v_{12} \\ v_{12} & 1 \end{pmatrix} \right) \quad \forall \pi_1 < \pi_2, \tag{1.12.10}$$

where $v_{12} = \lim_{n\to\infty} v_{12,n}$, $v_{12,n} \equiv \frac{\tilde{P}_{\pi_1}(x)' E(\tilde{P}_{\pi_1 i}\tilde{P}'_{\pi_2 i}\varepsilon_i^2)\tilde{P}_{\pi_2}(x)}{V_{\pi_1}^{1/2} V_{\pi_2}^{1/2}}$. By Cramér-Wold device, above holds if for any $\pi_1 < \pi_2$,

$$(\delta_1^2 + \delta_2^2 + 2\delta_1\delta_2 v_{12,n})^{-1}\delta_1 t_n^*(\pi_1) + \delta_2 t_n^*(\pi_2) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \omega_{ni} \xrightarrow{d} N(0,1) \quad \forall (\delta_1, \delta_2) \in \mathbb{R}^2 \tag{1.12.11}$$

where $\omega_{ni} = (\delta_1^2 + \delta_2^2 + 2\delta_1\delta_2 v_{12,n})^{-1}(\delta_1 \frac{\tilde{P}_{\pi_1}(x)'\tilde{P}_{\pi_1 i}\varepsilon_i}{V_{\pi_1}^{1/2}} + \delta_2 \frac{\tilde{P}_{\pi_2}(x)'\tilde{P}_{\pi_2 i}\varepsilon_i}{V_{\pi_2}^{1/2}})$.

To show (1.12.11), we need to verify Lindberg's condition. Note that $E\omega_{ni} = 0$, and $\frac{1}{n}\sum_{i=1}^n E[\omega_{ni}^2] = 1$, since

$$E[\omega_{ni}^2] = (\delta_1^2 + \delta_2^2 + 2\delta_1\delta_2 v_{12,n})^{-1}(\delta_1^2 E[(\frac{\tilde{P}_{\pi_1}(x)'\tilde{P}_{\pi_1 i}\varepsilon_i}{V_{\pi_1}^{1/2}})^2] + \delta_2^2 E[(\frac{\tilde{P}_{\pi_2}(x)'\tilde{P}_{\pi_2 i}\varepsilon_i}{V_{\pi_2}^{1/2}})^2]$$
$$+ \delta_1\delta_2 E[(\frac{\tilde{P}_{\pi_1}(x)'\tilde{P}_{\pi_1 i}\varepsilon_i}{V_{\pi_1}^{1/2}})(\frac{\tilde{P}_{\pi_2}(x)'\tilde{P}_{\pi_2 i}\varepsilon_i}{V_{\pi_2}^{1/2}})]) = 1.$$

By Assumption 1.2, we have $||\delta_1 \frac{\tilde{P}_{\pi_1}(x)'\tilde{P}_{\pi_1 i}}{V_{\pi_1}^{1/2}} + \delta_2 \frac{\tilde{P}_{\pi_2}(x)'\tilde{P}_{\pi_2 i}}{V_{\pi_2}^{1/2}}||_\infty \lesssim \zeta_{\bar{K}}\lambda_{\bar{K}}$. Therefore, for any $a > 0$,

$$\frac{1}{n}\sum_{i=1}^n E(|\omega_{ni}|^2 1\{|\omega_{ni}| > a\sqrt{n}\})$$
$$\lesssim 2E[|\frac{\tilde{P}_{\pi_1}(x)'\tilde{P}_{\pi_1 i}\varepsilon_i}{V_{\pi_1}^{1/2}}|^2 1\{|\omega_{ni}| > a\sqrt{n}\}] + 2E[|\frac{\tilde{P}_{\pi_2}(x)'\tilde{P}_{\pi_2 i}\varepsilon_i}{V_{\pi_2}^{1/2}}|^2 1\{|\omega_{ni}| > a\sqrt{n}\}]$$
$$\leq 2(E[|\frac{\tilde{P}_{\pi_1}(x)'\tilde{P}_{\pi_1 i}\varepsilon_i}{V_{\pi_1}^{1/2}}|^2 + E[|\frac{\tilde{P}_{\pi_2}(x)'\tilde{P}_{\pi_2 i}\varepsilon_i}{V_{\pi_2}^{1/2}}|^2) \sup_x E[\varepsilon_i^2 1\{|\varepsilon_i| > a(\sqrt{n}/(\zeta_{\bar{K}}\lambda_{\bar{K}})\}|x_i = x],$$

where the last term goes to 0 under $n \to \infty$ by Assumption 1.2-(2), since $E[(\frac{\tilde{P}_\pi(x)'\tilde{P}_{\pi i}\varepsilon_i}{V_\pi^{1/2}})^2] \asymp 1$ for any π by Assumption 1.2-(2) and $(\zeta_{\bar{K}}\lambda_{\bar{K}})/\sqrt{n} = o(1)$ by Assumption 1.2-(4). Thus, Lindberg condition is verified and (1.12.11) holds by Lindberg-Feller CLT. Therefore (1.12.10) holds by Slutzky's Theorem. We show that the finite dimensional convergence to a Gaussian distribution with covariance kernel in the Theorem 1.1.

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 $|f_{n,\pi}(\varepsilon,t)| = |\alpha(x,\pi)'P_\pi(t)\varepsilon| \leq |\frac{V_1^{1/2}}{V_{\underline{\pi}}^{1/2}} f_{n,1}(\varepsilon,t)|$ We define envelope function $F_n(\varepsilon,t) \equiv |\frac{V_1^{1/2}}{V_{\underline{\pi}}^{1/2}} f_{n,1}(\varepsilon,t)| \vee 1$. 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 $a > 0$,

$$E(F_n^2 1\{F_n > a\sqrt{n}\}) = E[(\frac{\tilde{P}_1(x)'\tilde{P}_1(x_i)}{V_{\underline{\pi}}^{1/2}}\varepsilon_i)^2 1\{\varepsilon_i^2|\varepsilon_i| > a(\sqrt{n}/(\zeta_{\bar{K}}\lambda_{\bar{K}})\}] \quad (1.12.12)$$

$$\leq \sup_x E[\varepsilon_i^2 1\{|\varepsilon_i| > a(\sqrt{n}/(\zeta_{\bar{K}}\lambda_{\bar{K}})\}|X_i = x] = o(1) \quad (1.12.13)$$

since $(\zeta_{\bar{K}}\lambda_{\bar{K}})/\sqrt{n} = o(1)$ by Assumption 1.2-(2). Moreover, for every $\delta_n \to 0$,

$$\sup_{\rho(\pi_1,\pi_2)<\delta_n} E(f_{n,\pi_1} - f_{n,\pi_2})^2 \to 0 \quad (1.12.14)$$

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

Define also $\kappa \equiv \sup_{\pi\neq\pi'} \frac{V_\pi(x)-V_{\pi'}(x)}{||\pi'-\pi||}$. For any $\pi, \pi' \in \Pi = [\underline{\pi}, 1]$ such that $\pi < \pi'$,

$$|\alpha(x,\pi')'P_{\pi'}(t) - \alpha(x,\pi)'P_\pi(t)| = |\frac{\tilde{P}_{\pi'}(x)'\tilde{P}_{\pi'}(t)}{V_{\pi'}^{1/2}(x)} - \frac{\tilde{P}_\pi(x)'\tilde{P}_\pi(t)}{V_\pi^{1/2}(x)}| \quad (1.12.15)$$

$$\leq |\frac{\tilde{P}_{\pi'}(x)'\tilde{P}_{\pi'}(t) - \tilde{P}_\pi(x)'\tilde{P}_\pi(t)}{V_{\pi'}^{1/2}(x)}| + |\tilde{P}_\pi(x)'\tilde{P}_\pi(t)(\frac{1}{V_{\pi'}^{1/2}(x)} - \frac{1}{V_\pi^{1/2}(x)})| \quad (1.12.16)$$

$$= |\frac{\tilde{P}_{\pi'-\pi}(x)'\tilde{P}_{\pi'-\pi}(t)}{V_{\pi'}^{1/2}(x)}| + |\tilde{P}_\pi(x)'\tilde{P}_\pi(t)(\frac{V_{\pi'}(x) - V_\pi(x)}{V_\pi^{1/2}(x)V_{\pi'}^{1/2}(x)(V_\pi^{1/2}(x) + V_{\pi'}^{1/2}(x))})| \quad (1.12.17)$$

$$\leq |\frac{\tilde{P}_{\pi'-\pi}(x)'\tilde{P}_{\pi'-\pi}(t)}{V_{\underline{\pi}}^{1/2}(x)}| + |\frac{\tilde{P}_1(x)'\tilde{P}_1(t)}{V_{\underline{\pi}}^{1/2}(x)}(\frac{V_{\pi'}(x) - V_\pi(x)}{2V_{\underline{\pi}}(x)})| \quad (1.12.18)$$

$$= |\frac{\tilde{P}_1(x)'\tilde{P}_1(t)}{V_{\underline{\pi}}^{1/2}(x)}|(|\frac{\tilde{P}_{\pi'-\pi}(x)'\tilde{P}_{\pi'-\pi}(t)}{\tilde{P}_1(x)'\tilde{P}_1(t)}| + |\frac{V_{\pi'}(x) - V_\pi(x)}{2V_{\underline{\pi}}(x)}|) \quad (1.12.19)$$

$$\leq |\frac{\tilde{P}_1(x)'\tilde{P}_1(t)}{V_{\underline{\pi}}^{1/2}(x)}|(C||(\pi' - \pi)|| + \frac{\kappa||\pi' - \pi||}{2V_{\underline{\pi}}(x)}) \quad (1.12.20)$$

$$\leq |\frac{\tilde{P}_1(x)'\tilde{P}_1(t)}{V_{\underline{\pi}}^{1/2}(x)}| \cdot A||\pi' - \pi|| \quad (1.12.21)$$

where the second inequality uses $V_{\underline{\pi}}(x) \leq V_\pi(x)$ for any $\pi \in \Pi$. The thid inequality uses $||\frac{\tilde{P}_\pi(x)'\tilde{P}_{\pi i}}{V_{\underline{\pi}}^{1/2}}||_\infty \lesssim \zeta_{\bar{K}\pi}$ and the definition of κ. The last inequality uses the $\tilde{P}_\pi(x)'\tilde{P}_\pi(t) \propto \bar{K}\pi$, $\kappa \lesssim V_1(x)$, $V_\pi(x) \asymp V_{\pi'}(x)$ and this holds for some constant A and sufficiently large n. From

this, we have

$$|f_{n,\pi'} - f_{n,\pi}| = |\varepsilon\alpha(x,\pi')'P_{\pi'}(t) - \varepsilon\alpha(x,\pi)'P_{\pi}(t)| \leq |\varepsilon||\frac{\tilde{P}_1(x)'\tilde{P}_1(t)}{V_{\underline{\pi}}^{1/2}(x)}|A||\pi' - \pi|| = |F_n|A||\pi' - \pi||. \tag{1.12.22}$$

Therefore, the class of functions $\mathcal{F}_n = \{f_{n,\pi} : \pi \in \Pi\}$ satisfy Lipschitz conditions, thus it is VC classes, and we have that

$$\sup_Q N(\epsilon||F_n||_{L^2(Q)}, \mathcal{F}_n, L^2(Q)) \leq (A/\epsilon)^V, 0 < \forall\epsilon \leq 1, V > 0 \tag{1.12.23}$$

for some $A > 0$ and for each n with some constant V independent of n. Then, following uniform-entropy condition holds for every $\delta_n \to 0$.

$$J(\delta_n, \mathcal{F}_n, L^2(Q)) = \int_0^{\delta_n} \sqrt{\log \sup_Q N(\epsilon||F_n||_{L^2(Q)}, \mathcal{F}_n, L^2(Q))} \longrightarrow 0. \tag{1.12.24}$$

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 (1.12.5), we have $T_n^*(\pi, \theta_0) \Rightarrow \mathbb{T}(\pi) + \nu(\pi)$ for $\pi \in \Pi$. In addition, if Assumption 1.3 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.

Proof of Theorem 1.2

Proof. We prove the finite dimensional convergence use similar arguments to those used in the proof of Theorem 1.1. We repeat this here, as Assumption 1.4 impose different rates of K compare with the Assumption 1.1. Similar to the proof in Theorem 1.1, 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 $\nu_n(m) = -\sqrt{n} V_{K_m}^{-1/2} r_{K_m}(x)$ from Assumption 1.2. For any $K_{m_1} \lesssim K_{m_2}$, we need to show

$$\begin{pmatrix} t_n(m_1) \\ t_n(m_2) \end{pmatrix} \xrightarrow{d} N(0, I_2).$$

By Cramér-Wold device, it also suffices to show that $\delta_1 t_n(m_1) + \delta_2 t_n(m_2) \xrightarrow{d} N(0, \delta_1^2 + \delta_2^2)$ for $\forall \delta_1, \delta_2 \in \mathbb{R}$. For any δ_1, δ_2, define

$$(\delta_1^2 + \delta_2^2 + 2\delta_1 \delta_2 v_{12,n})^{-1} (\delta_1 t_n(m_1) + \delta_2 t_n(m_1) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \omega_{ni}$$

where $v_{12,n} = \frac{\tilde{P}_{K_{m_1}}(x)' E(\tilde{P}_{K_{m_1} i} \tilde{P}'_{K_{m_2} i} \varepsilon_i^2) \tilde{P}_{K_{m_2}}(x)}{V_{K_{m_1}}^{1/2} V_{K_{m_2}}^{1/2}}$, and $\omega_{ni} = (\delta_1^2 + \delta_2^2 + 2\delta_1 \delta_2 v_{12,n})^{-1/2} (\delta_1 \frac{\tilde{P}_{K_{m_1}}(x)' \tilde{P}_{K_{m_1} i} \varepsilon_i}{V_{K_m}^{1/2}} + \delta_2 \frac{\tilde{P}_{K_{m_1}}(x)' \tilde{P}_{K_{m_1} i} \varepsilon_i}{V_{K_m}^{1/2}}$ are defined similarly as in the proof of Theorem 1.1. Observe that $E\omega_{ni} = 0$, and $\frac{1}{n} \sum_{i=1}^n E[\omega_{ni}^2] = 1$, and

$$|| \frac{\tilde{P}_{K_{\pi_1}}(x)' \tilde{P}_{K_{\pi_1}}}{\sqrt{V_{K_{\pi_1}}}} + \frac{\tilde{P}_{K_{\pi_2}}(x)' \tilde{P}_{K_{\pi_2}}}{\sqrt{V_{K_{\pi_2}}}} ||_\infty \lesssim \zeta_{K_{m_1}} \lambda_{K_{m_1}} + \zeta_{K_{m_2}} \lambda_{K_{m_2}} \lesssim \zeta_{K_{m_2}} \lambda_{K_{m_2}}$$

by Assumptions 1.2 and 1.4. Therefore, Lindberg's condition can be verified similarly as in the proof of Theorem 1.1.

Moreover,

$$v_{12,n} = \frac{\tilde{P}_{K_{m_1}}(x)' E(\tilde{P}_{K_{m_1} i} \tilde{P}'_{K_{m_2} i} \varepsilon_i^2) \tilde{P}_{K_{m_2}}(x)}{V_{K_{m_1}}^{1/2} V_{K_{m_2}}^{1/2}} \leq C \frac{V_{K_{m_1}}^{1/2}}{V_{K_{m_2}}^{1/2}}$$

for some constant $C > 0$ by Assumption 1.2-(2). The latter term converges to 0 as $n \to 0$

by Assumption 1.4, thus $v_{12,n} \to 0$. Therefore, finite dimensional convergence holds by Lindberg-Feller CLT and Slutzky's Theorem. Under $\sup_m |\nu(m)| < \infty$, joint asymptotic distributions of $(T_n(K_1, \theta_0), \cdots, T_n(K_M, \theta_0))'$ also holds with the definition of $\nu = (\nu(1), \cdots, \nu(M))'$.

If $\nu(m) = \infty$ for some $\nu(m)$, then $T_n(K_m, \theta_0) \xrightarrow{p} +\infty$, or if $\nu(m) = -\infty$ then $T_n(K_m, \theta_0) \xrightarrow{p} -\infty$. Let $G(\cdot)$ be a strictly increasing continuous df 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

$$G_{n,m} \xrightarrow{d} G(Z_m + \nu(m)) \tag{1.12.25}$$

by finite dimensional CLT under Assumptions 1.2, 1.4 and the continuous mapping theorem. If $\nu(m) = +\infty$

$$G_{n,m} \xrightarrow{p} 1 \tag{1.12.26}$$

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

$$G_{n,m} \xrightarrow{p} 0 \tag{1.12.27}$$

as $G(x) \to 0$ as $x \to -\infty$. Since (1.12.25), (1.12.26), and (1.12.27) holds jointly, this completes the second part of the proof, as

$$G_n = (G_{n,1}, \cdots, G_{n,M})' \xrightarrow{d} G_\infty \equiv (G(Z_1 + \nu(1)), \cdots, G(Z_M + \nu(M)))' \tag{1.12.28}$$

where $G(\infty) = 1$ if $\nu(m) = +\infty$ and $G(-\infty) = 0$ if $\nu(m) = -\infty$. *Q.E.D.*

Proof of Corollary 1.1

Proof. Under Assumptions 1.1, 1.2 and $\sup_\pi |\nu(\pi)| < \infty$, we have $T_n^*(\pi, \theta_0) \Rightarrow \mathbb{T}(\pi) + \nu(\pi)$ by Theorem 1.1. Then for any continuous function $l(\cdot) : \ell^\infty(\Pi) \to \mathbb{R}$, $l(T_n^*(\pi, \theta_0)) \xrightarrow{d} l(\mathbb{T}(\pi) + \nu(\pi))$ holds by continuous mapping theorem. Thus, $\operatorname{Inf} T_n(\theta_0) = \inf_{K \in \mathcal{K}_n} |T_n(K, \theta_0)| = \inf_{\pi \in \Pi} |T_n^*(\pi, \theta_0)| \xrightarrow{d} \inf_\pi |\mathbb{T}(\pi) + \nu(\pi)|$ holds. In addition, if Assumption 1.3 holds, $\operatorname{Inf} T_n(\theta_0) \xrightarrow{d} \inf_\pi |\mathbb{T}(\pi)|$ by Theorem 1.1. This completes the Corollary 1.1-1.

For the second part in Corollary 1.1, remaining proof use similar approach those of Andrews and Guggenberger (2009) in the moment inequality literature. If some elements of $|\nu_m| = +\infty$ under oversmoothing sequences, joint distribution of $(T_{K_1}, \cdots, T_{K_M})'$ does not converge in distribution to a profer bounded random vector. Thus, continuous mapping theorem cannot be directly applied to obtain asymptotic distribution results in the Corollary.

We first define $S(t) \equiv \inf_m |t_m|$ for $t = (t_1, \cdots, t_M) \in \mathbb{R}^M_{[\pm\infty]} \setminus (\infty^M \cup (-\infty)^M)$ where $\infty^M = (+\infty, \cdots, +\infty), (-\infty)^M = (-\infty, \cdots, -\infty)$ (M copies). Define also $\bar{T}_n(\theta) \equiv (T_n(K_1, \theta), \cdots, T_n(K_M, \theta))'$. Then, under Assumption 1.4

$$\operatorname{Inf} T_n(\theta_0) = \inf_m |T_n(K_m, \theta_0)| = S(\bar{T}_n(\theta_0)). \tag{1.12.29}$$

We define $G^{-1}(\cdot)$ as the inverse of $G(\cdot)$. For $x = (x_1, \cdots, x_M)' \in \mathbb{R}^{M-1}_{[\pm\infty]} \times \mathbb{R}$, define $G_{(M)}(x) \equiv (G(x_1), \cdots, G(x_M))' \in [0,1]^{M-1} \times (0,1)$. 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 $y \in [0,1]^{M-1} \times (0,1)$,

$$S^*(y) \equiv S(G^{-1}_{(M)}(y)). \tag{1.12.30}$$

Note that $S^*(y)$ is continuous at all $y \in [0,1]^{M-1} \times (0,1)$ since $S(x)$ is continuous at all $x \in \mathbb{R}^{M-1}_{[\pm\infty]} \times \mathbb{R}$. We define the continuous function on the extended real space as follows; $S : A \to B$ is continuous at $x \in A$ if $x' \to x$ for $x \in A$ implies $G(x') \to G(x)$ for any set A.

Assumption 1.4 (especially, assumption of at least one $|\nu_m| = O(1)$) excludes the possibility of $\infty^M, (-\infty)^M$ and thus restricts the domain of functions appropriately. Therefore, we can immediately show $S(x)$ is continuous at all $x \in \mathbb{R}^{M-1}_{[\pm\infty]} \times \mathbb{R}$. Then, we have

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

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 Theorem 1.2, and the fourth and fifth equality uses the definition of S^*. If $|\nu_m| = +\infty$, corresponding elements of $|Z_m + \nu_m| = +\infty$ by construction. This completes the proof of Corollary 1.1. *Q.E.D.*

Proof of Corollary 1.2

Proof. We first provide (1.4.3) in Corollary 1.2-1. Under Assumptions 1.1-1.3, we have shown that $\operatorname{Inf} T_n(\theta_0) \xrightarrow{} \xi_{\inf} = \inf_{\pi \in [\underline{\pi}, 1]} |\mathbb{T}(\pi)|$ in Corollary 1.1-1. Therefore,

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

where the first equality holds under subsequence $\{u_n\}$ of $\{n\}$ by the definition of lim sup and the second equality uses the Corollary 1.1-1 and the definition of $c^{\inf}_{1-\alpha}$ in (1.4.2). Moreover,

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

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

Next, we prove Corollary 1.2-2. Under Assumptions 1.1, 1.2 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 $\nu(\pi)$. First, we have

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

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 $\mathbb{T}(\pi) \overset{d}{=} Z \sim N(0,1)$ and the definition of $F(\cdot)$. Finally, the last equality holds since $F(c, |\nu|)$ is monotone increasing function of $|\nu|$. Similarly,

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

Next consider Corollary 1.2-3. We have $\text{Inf}\, T_n(\theta_0) \xrightarrow{d} \inf_{m=1,\cdots,M} |Z_m + \nu(m)|$ by Corollary 1.1-2 under Assumptions 1.2 and 1.4. Then, for any $0 < c < \infty$,

$$\begin{aligned}
\limsup_{n\to\infty} P(\text{Inf}\, T_n(\theta_0) > c) &= P(\inf_{m=1,\cdots,M} |Z_m + \nu(m)| > c) \\
&= \prod_{m=1}^{M} P(|Z_m + \nu(m)| > c) = \prod_{m=M-M_1+1}^{M} F(c, |\nu(m)|)
\end{aligned}$$

by Corollary 1.1-2 and the definition of F and the fact that $F(c, |\nu(m)|) = 1$ for $|\nu(m)| = \infty$. Last equality holds when $|\nu(m)| = \infty$ for $m = 1, \cdots, M - M_1$ since $F(c, \infty) = 1$. By similar derivations we have shown above, $\prod_{m=1}^{M} F(c, |\nu(m)|) \leq F(c, \inf_m |\nu(m)|) = F(c,$ 0) by Assumption 1.4. Moreover, if $c = z_{1-\alpha/2}$ then the asymptotic size is controlled, as

$\limsup_{n\to\infty} P(\text{Inf}\, T_n(\theta_0) > z_{1-\alpha/2}) \leq F(z_{1-\alpha/2}, 0) = \alpha$. This completes the proof.

Q.E.D.

Proof of Corollary 1.3

Proof. Note that Assumption 1.7 holds under Assumptions 1.1, 1.2 with the following additional assumption, and this is essentially by Lemma 3.2 of Chen and Christensen (2015b).

$$|| \sum_{i=1}^{n} \tilde{P}_{Ki}\tilde{P}'_{Ki}\varepsilon_i^2 - E[\tilde{P}_{Ki}\tilde{P}'_{Ki}\varepsilon_i^2]|| = o_p(1)$$

Under Assumptions 1.2, 1.3, 1.5, and 1.6, following finite dimensional convergence holds by Theorem 1.1,

$$\bar{T}_n(\theta) = (T_n(K_1, \theta_0), \cdots, T_n(K_M, \theta_0))' \xrightarrow{d} Z = (Z_1, \cdots, Z_M)', \quad Z \sim N(0, \Sigma) \tag{1.12.31}$$

Under Assumptions 1.2-1.4, above also holds with $\Sigma = I_M$ by Theorem 1.2. 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

$$(T_{n,\widehat{V}}(K_1, \theta_0), \cdots, T_{n,\widehat{V}}(K_M, \theta_0))' = A\bar{T}_n(\theta) \xrightarrow{d} Z \tag{1.12.32}$$

by Assumption 1.7 and Slutzky Theorem, where $A \equiv diag\{\frac{V_{K_1}^{1/2}}{\widehat{V}_{K_1}^{1/2}}, \cdots, \frac{V_{K_M}^{1/2}}{\widehat{V}_{K_M}^{1/2}}\}$, and $A \xrightarrow{p} I_M$

Next consider $\widehat{c}_{1-\alpha}^{\text{inf}}$ which is $(1-\alpha)$ quantile of $\inf_{m=1,\cdots,M} |Z_{m,\widehat{\Sigma}}|$ defined in (1.4.10),

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

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

$j < l$,

$$\widehat{\Sigma}_{jl} = \frac{\widehat{V}_{K_j}^{1/2}}{\widehat{V}_{K_l}^{1/2}} = \frac{\widehat{V}_{K_j}^{1/2}}{V_{K_j}^{1/2}} \frac{V_{K_j}^{1/2}}{V_{K_l}^{1/2}} \frac{V_{K_l}^{1/2}}{\widehat{V}_{K_l}^{1/2}} \xrightarrow{p} \Sigma_{jl} \tag{1.12.33}$$

by Assumption 1.7. 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.*

Proof of Corollary 1.4

Proof. We first show Corollary 1.4-1.

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

where the first equality holds by Corollary 1.3 under Assumptions 1.2, 1.3, 1.5, 1.6, and 1.7, and the second equality holds by Corollary 1.1-1. Similarly, we can show

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{\texttt{inf}}) = P(\inf_m |Z_m| \leq z_{1-\alpha/2}) \geq P(|Z_m| \leq z_{1-\alpha/2}) = 1-\alpha. \tag{1.12.34}$$

Next, consider Corollary 1.4-2.

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{\texttt{inf}}^{\texttt{Robust}}) = 1 - \limsup_{n\to\infty} P(\mathrm{Inf}\, T_n(\theta_0) > \widehat{c}_{1-\alpha}^{inf}) \geq 1 - F(c_{1-\alpha}^{\inf}, \inf_m |\nu(m)|) \tag{1.12.35}$$

by Corollary 1.2-2 and Corollary 1.3. Equation (1.5.5) can be similarly derived from Corollary 1.2-2.

Finally, consider Corollary 1.4-3.

$$\liminf_{n\to\infty} P(\theta_0 \in CI^{\texttt{Robust}}_{\texttt{inf}}) = 1 - \limsup_{n\to\infty} P(\mathrm{Inf}\, T_n(\theta_0) > \widehat{c}^{\,inf}_{1-\alpha}) \tag{1.12.36}$$

$$= 1 - P(\inf_{m=1,\cdots,M} |Z_m + \nu(m)| > c^{\inf}_{1-\alpha}) \tag{1.12.37}$$

$$= 1 - \prod_{m=1}^{M} F(c^{\inf}_{1-\alpha}, |\nu(m)|) \tag{1.12.38}$$

where the second equality uses Corollary 1.1-2 and Corollary 1.3, and the third equality uses asymptotic independence of Z_m by Theorem 1.2. Similarly, we have that $\liminf_{n\to\infty} P(\theta_0 \in CI_{\texttt{inf}}) \geq 1 - F(z_{1-\alpha/2}, 0) = 1 - \alpha$ under Assumption 1.4. *Q.E.D.*

Proof of Corollary 1.5

Proof. Similar to the proof of Corollary 1.3, we can also 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 $\mathrm{Sup}\, T_n(\theta_0) = \sup_m |T_{n,\widehat{V}}(K_m, \theta_0)| \xrightarrow{d} \sup_m |Z_{m,\Sigma}|$ under Assumptions 1.2, 1.3, 1.5, 1.6, and 1.7. Therefore, we have

$$\liminf_{n\to\infty} P(\theta_0 \in CI^{\texttt{Robust}}_{\texttt{pms}}) = \liminf_{n\to\infty} P(|T_{n,\widehat{V}}(\widehat{K}, \theta_0)| \leq \widehat{c}^{\,sup}_{1-\alpha}) \tag{1.12.39}$$

$$\geq \liminf_{n\to\infty} P(\mathrm{Sup}\, T_n(\theta_0) \leq \widehat{c}^{\,sup}_{1-\alpha}) \tag{1.12.40}$$

$$= P(\sup_m |Z_{m,\Sigma}| \leq c^{sup}_{1-\alpha}) = 1 - \alpha \tag{1.12.41}$$

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

Proof of Theorem 1.3

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 $g = [g_1, \cdots, g_n]', g_i = g_0(x_i), g_w = [g_{w1}, \cdots, g_{wn}]', g_{wi} = g_{w0}(x_i) = E[w_i|x_i], v = [v_1, \cdots, v_n]$.

Under Assumption 1.8 and conditional homoskedastic error terms, $E[v_i^2|x_i] = E[v_i^2]$,

$$\widehat{\Gamma}_K = \Gamma_K + o_p(1), \quad \Gamma_K = (1 - K/n)E[v_i^2] \tag{1.12.42}$$

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

$$S_K = \frac{1}{\sqrt{n}} v' M_K \varepsilon + \frac{1}{\sqrt{n}} g_w' M_K g + \frac{1}{\sqrt{n}}(v' M_K g + g_w' M_K \varepsilon) \tag{1.12.43}$$
$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} M_{K,ii} v_i \varepsilon_i - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1, j<i}^{n} P_{K,ij}(v_i \varepsilon_j + v_j \varepsilon_i) + o_p(1) \tag{1.12.44}$$

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 1.8. 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).

To show joint convergence, it suffices to show for any $K_1 < K_2$ in $\mathcal{K}_n$

$$\delta_1 T_n(K_1, \theta_0) + \delta_2 T_n(K_2, \theta_0) \xrightarrow{d} N(0, (\delta_1^2 + \delta_2^2 + 2\delta_1\delta_2 v_{12})) \quad \forall (\delta_1, \delta_2) \in \mathbb{R}^2 \tag{1.12.45}$$

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, Swanson, Hausman, Newey and Woutersen (2012). Define $Y_n, Y_{1,n}$ and $Y_{2,n}$ as follows

$$Y_n = \delta_1 Y_{1,n} + \delta_2 Y_{2,n}, \tag{1.12.46}$$

$$Y_{1,n} = \omega_{1,1n} + \sum_{i=2}^{n} y_{1,in}, \quad y_{1,in} = \omega_{1,in} + \bar{y}_{1,in}, \tag{1.12.47}$$

$$Y_{2,n} = \omega_{2,1n} + \sum_{i=2}^{n} y_{2,in}, \quad y_{2,in} = \omega_{2,in} + \bar{y}_{2,in}, \tag{1.12.48}$$

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, Swanson, Hausman, Newey and Woutersen (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.

$$\sum_{i=2}^{n}(\delta_1 y_{1,in} + \delta_2 y_{2,in}) \xrightarrow{d} N(0, \delta_1^2 + \delta_2^2 + 2\delta_1\delta_2 v_{12}). \tag{1.12.49}$$

It remains to provide Lindeberg-Feller condition.

$$E[(\sum_{i=2}^{n} \delta_1 y_{1,in} + \delta_2 y_{2,in})^2 | X] = \delta_1^2 E[(\sum_{i=2}^{n} y_{1,in})^2 | X] + \delta_2^2 E[(\sum_{i=2}^{n} y_{2,in})^2 | X] + 2\delta_1\delta_2 E[\sum_{i=2}^{n}\sum_{j=2}^{n} y_{1,in} y_{2,in} | X], \tag{1.12.50}$$

where the first and second terms in (1.12.50) goes to δ_1^2, δ_2^2 a.s., respectively, as in the proof of Lemma A.2 in Chao, Swanson, Hausman, Newey and Woutersen (2012). Note that $E[\omega_{1,in}\bar{y}_{2,jn}|X] = 0, E[\omega_{2,in}\bar{y}_{1,jn}|X] = 0$ for all i, j, 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:

$$E[Y_{1,n}Y_{2,n}|X] = \frac{1}{n}V_{K_1}^{-1/2}\Gamma_{K_1}^{-1}E[v'M_{K_1}\varepsilon v'M_{K_2}\varepsilon|X]\Gamma_{K_2}^{-1}V_{K_2}^{-1/2} \tag{1.12.51}$$

$$= \frac{1}{n}V_{K_1}^{-1/2}\Gamma_{K_1}^{-1}\sigma_\varepsilon^2 E[v'M_{K_2}v|X]\Gamma_{K_2}^{-1}V_{K_2}^{-1/2} \tag{1.12.52}$$

$$= V_{K_1}^{-1/2}\Gamma_{K_1}^{-1}\sigma_\varepsilon^2\Gamma_{K_2}\Gamma_{K_2}^{-1}V_{K_2}^{-1/2} \tag{1.12.53}$$

$$= V_{K_1}^{1/2}/V_{K_2}^{1/2} \tag{1.12.54}$$

where the second equality uses conditional homoskedasticity $E[\varepsilon^2|X,Z] = \sigma_\varepsilon^2$ 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 (1.12.50) as follows

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

$$- \sum_{i=2}^{n}E[\omega_{2,1n}y_{1,in}|X] - E[\omega_{1,1n}\omega_{2,1n}|X] \tag{1.12.55}$$

$$= V_{K_1}^{1/2}/V_{K_2}^{1/2} - E[\omega_{1,1n}\omega_{2,1n}|X] \to v_{12} \quad a.s. \tag{1.12.56}$$

Also as in the proof of Lemma A.2 of Chao, Swanson, Hausman, Newey and Woutersen (2012), we have

$$\sum_{i=2}^{n}E[(\delta_1 y_{1,in} + \delta_2 y_{2,in})^4|X] \lesssim \sum_{i=2}^{n}E[(y_{1,in})^4|X] + \sum_{i=2}^{n}E[(y_{2,in})^4|X] \to 0 \quad a.s. \tag{1.12.57}$$

Thus, by similar arguments following the proof of Lemma A.2 in Chao, Swanson, Hausman, Newey and Woutersen (2012), we can apply the martingale central limit theorem. Then, by Slutzky theorem, joint convergence holds with the claimed covariance. We can show the coverage results using similar arguments to those used in the proof of Corollary 1.4. By Theorem 2 in Cattaneo, Jansson and Newey (2015a), Assumption 1.7 holds with the

following variance estimator for V_K

$$\widehat{V}_K = s^2 \widehat{\Gamma}_K^{-1}, \quad s^2 = \frac{1}{n-1-K} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2, \quad \widehat{\varepsilon}_i^2 = \sum_{j=1}^{n} M_{K,ij}(y_j - \widehat{\theta}_K w_j). \tag{1.12.58}$$

Q.E.D.

1.13 Figures and Tables

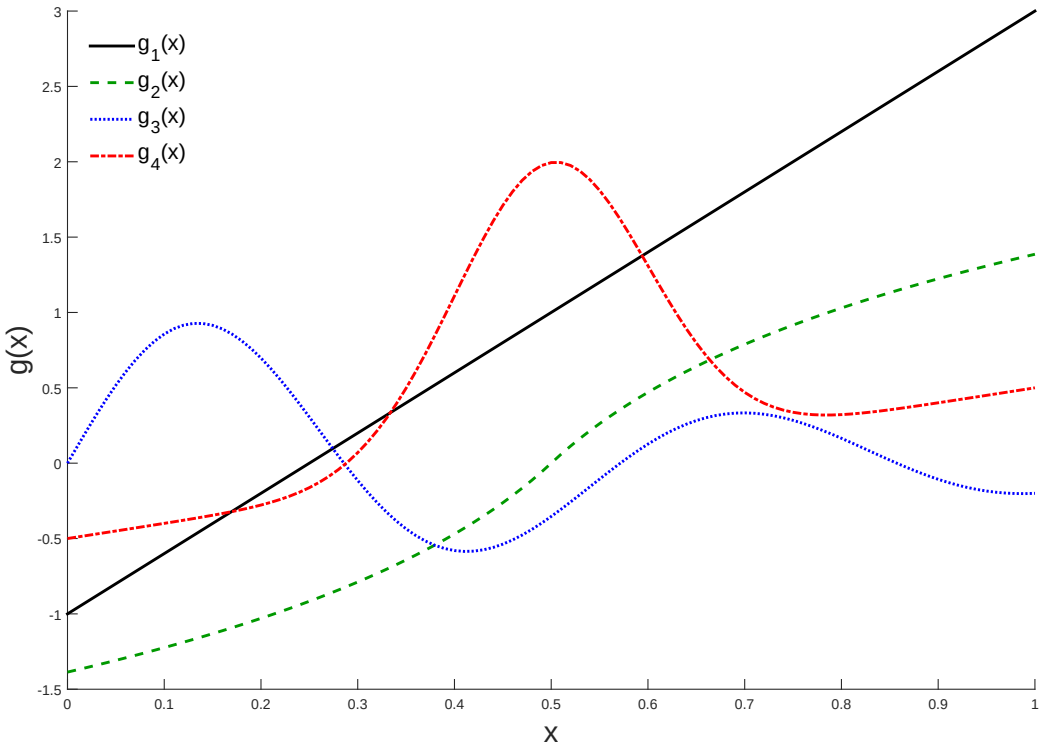


Figure 1.1: Different functions of $g(x)$.
Solid lines (Black) are $g_1(x) = 4x - 1$; Dashed lines (Green) are $g_2(x) = \ln(|6x - 3| + 1)sgn(x - 1/2)$; Dotted lines (Blue) are $g_3(x) = \sin(7\pi x/2)/[1 + 2x^2(sgn(x) + 1)]$; and Dash-dot lines (Red) are $g_4(x) = x - 1/2 + 5\phi(10(x - 1/2))$, where $\phi(\cdot)$ is standard normal pdf.

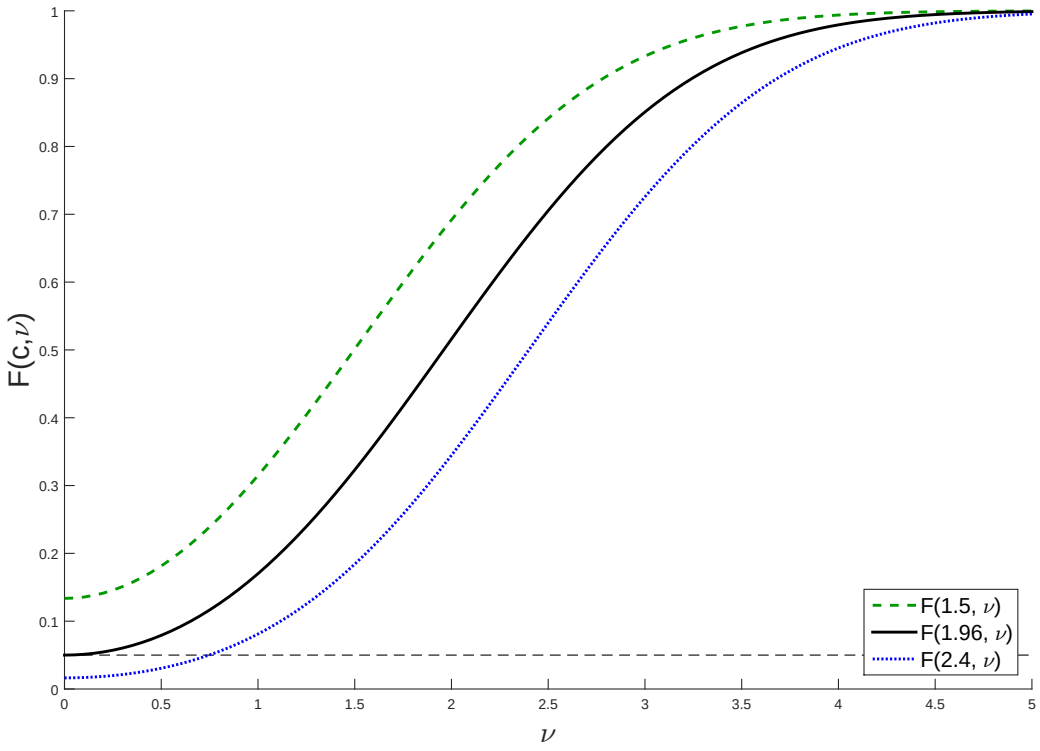
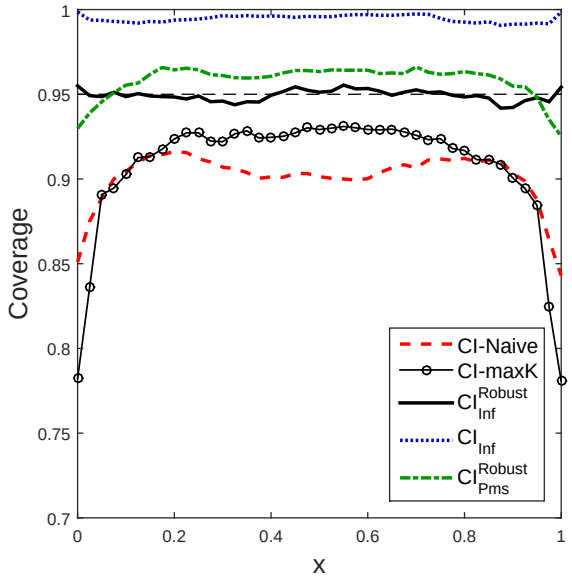


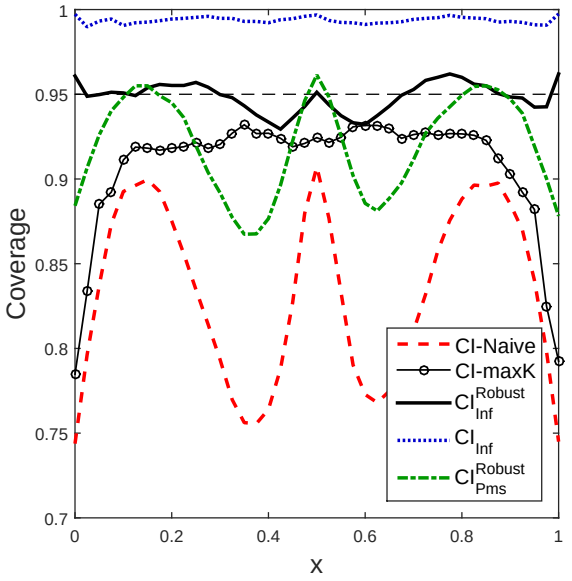
Figure 1.2: Plots of $F(c, \nu)$ as a function of ν for c = 1.5, 1.96, and 2.4.

Figure 1.3: 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 $\widehat{K}_{\texttt{cv}}$.

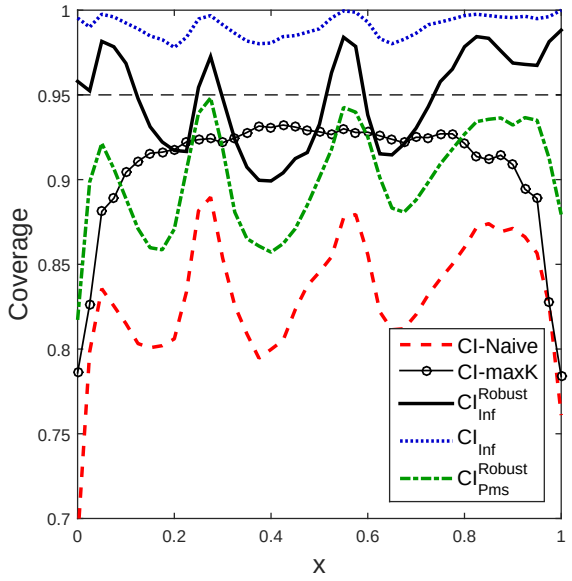
(a) $g_1(x) = 4x - 1$



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

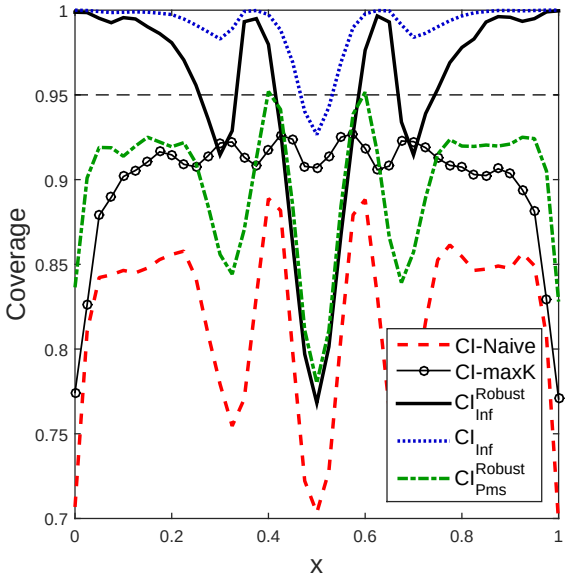
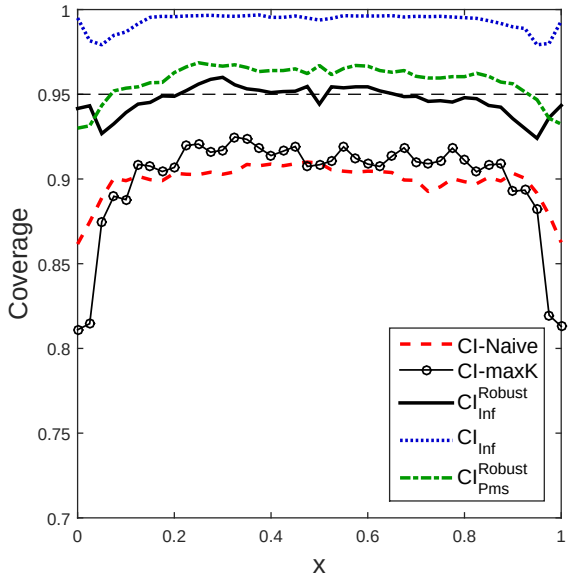
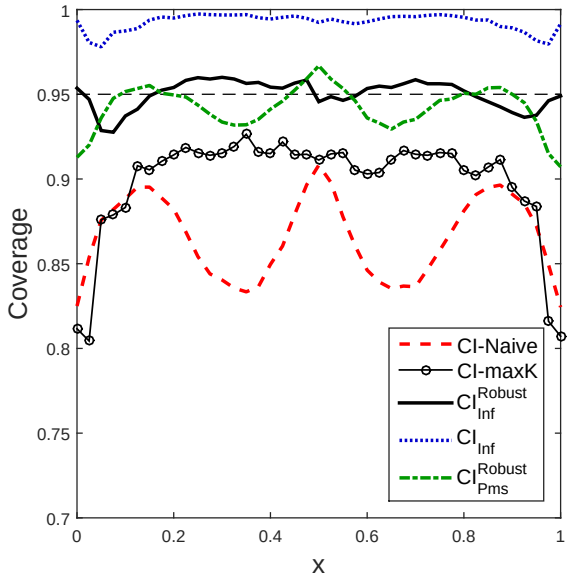


Figure 1.4: Coverage - Splines
Nominal 95% Coverage of Various CIs for $g(x)$:
(1) $CI^{\texttt{Naive}}_{\texttt{pms}}$ with $\widehat{K}_{\texttt{cv}}$ (2) $CI_{\texttt{maxK}}$ with $\bar{K}$ (3) $CI^{\texttt{Robust}}_{\texttt{inf}}$ (4) $CI_{\texttt{inf}}$ (5) $CI^{\texttt{Robust}}_{\texttt{pms}}$ with $\widehat{K}_{\texttt{cv}}$.

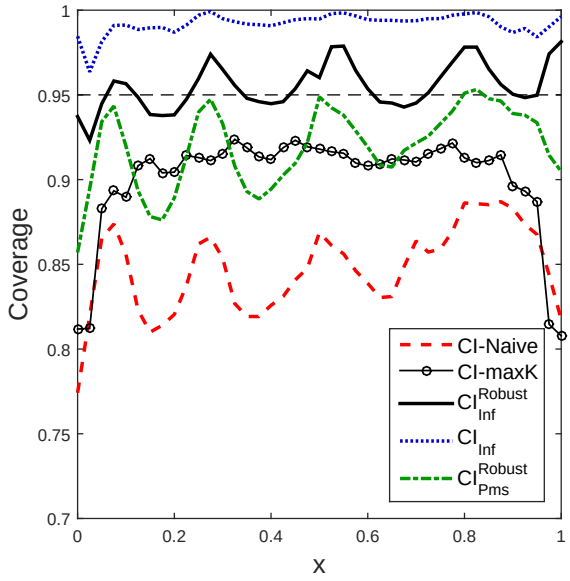
(a) $g_1(x) = 4x - 1$



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

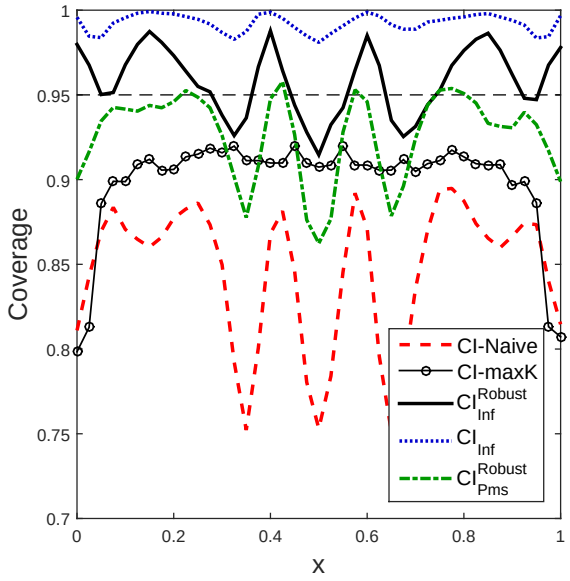
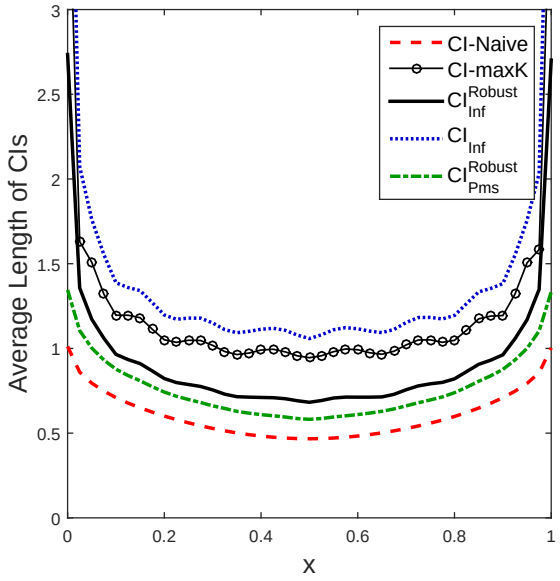
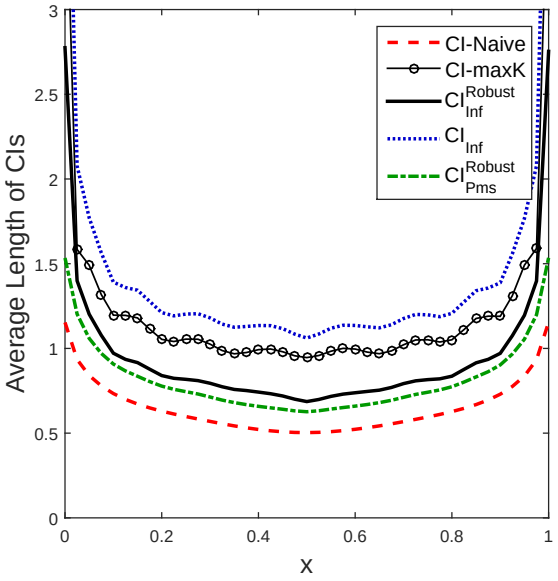


Figure 1.5: Length of CIs - Polynomials
Average lengths of nominal 95% CIs for $g(x)$:
(1) $CI^{\texttt{Naive}}_{\texttt{pms}}$ with $\widehat{K}_{\texttt{cv}}$ (2) $CI_{\texttt{maxK}}$ with $\bar{K}$ (3) $CI^{\texttt{Robust}}_{\texttt{inf}}$ (4) $CI_{\texttt{inf}}$ (5) $CI^{\texttt{Robust}}_{\texttt{pms}}$ with $\widehat{K}_{\texttt{cv}}$.

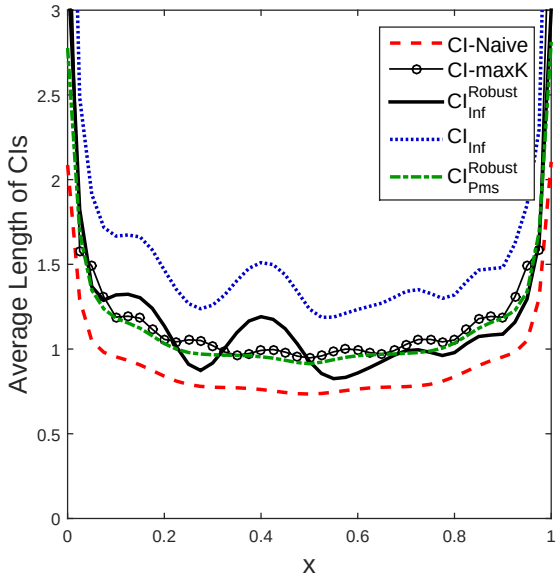
(a) $g_1(x) = 4x - 1$



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

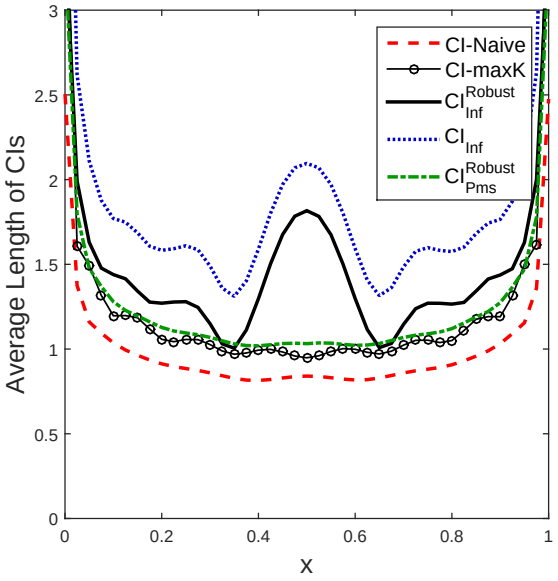
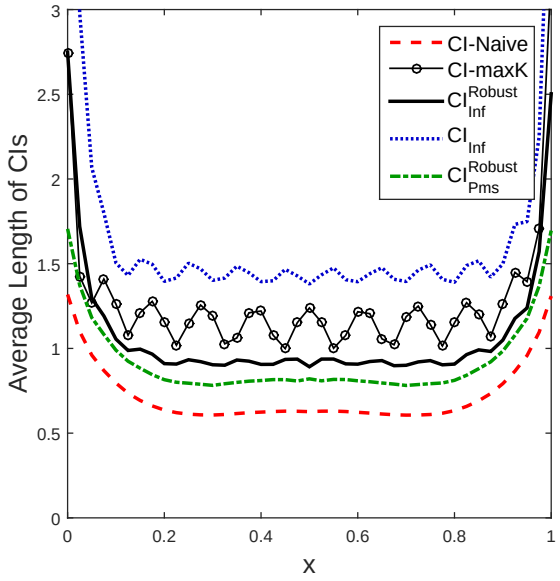
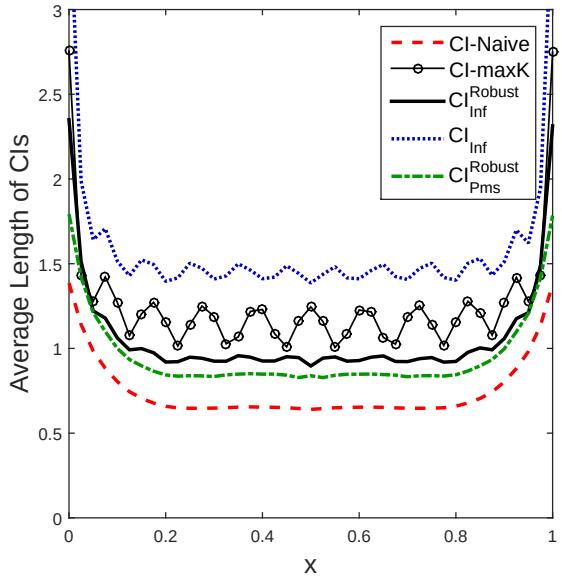


Figure 1.6: Length of CIs - Splines
Average lengths of nominal 95% CIs for $g(x)$:
(1) $CI^{\texttt{Naive}}_{\texttt{pms}}$ with $\widehat{K}_{\texttt{cv}}$ (2) $CI_{\texttt{maxK}}$ with $\bar{K}$ (3) $CI^{\texttt{Robust}}_{\texttt{inf}}$ (4) $CI_{\texttt{inf}}$ (5) $CI^{\texttt{Robust}}_{\texttt{pms}}$ with $\widehat{K}_{\texttt{cv}}$.

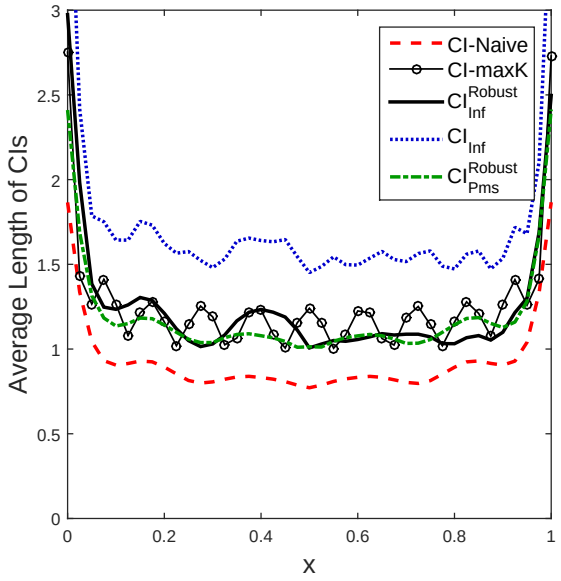
(a) $g_1(x) = 4x - 1$



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

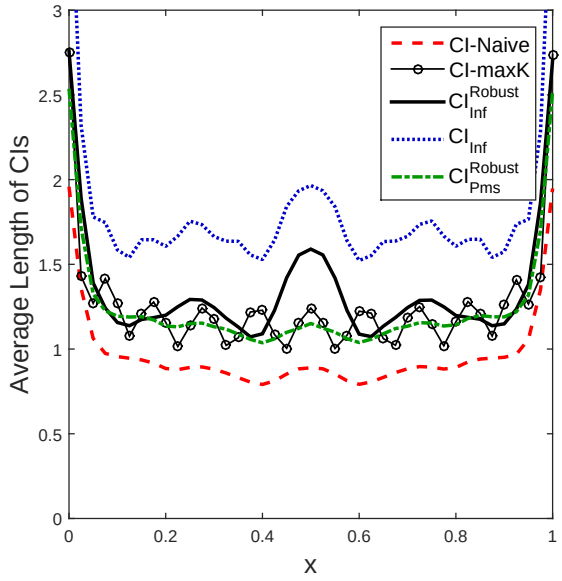
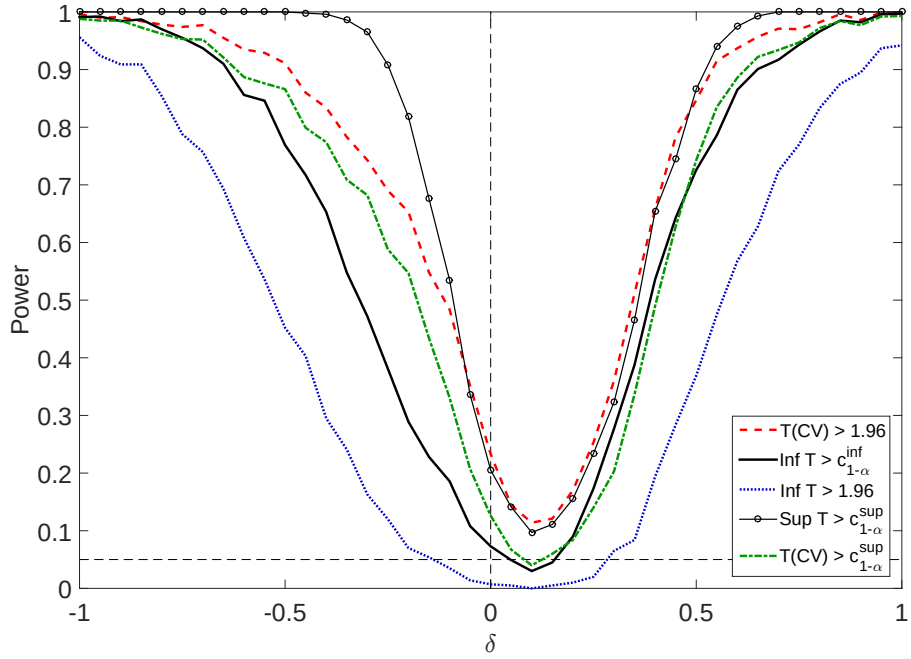


Figure 1.7: Power function against fixed alternatives. Design 2 : $g_2(x) = \ln(|6x-3|+1)sgn(x-1/2)$. $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_0 + \delta$, where $\theta_0 = g_2(x)$ at $x = 0.4$ for figure (a) and $x = 0.5$ for figure (b). Using Polynomials.

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



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

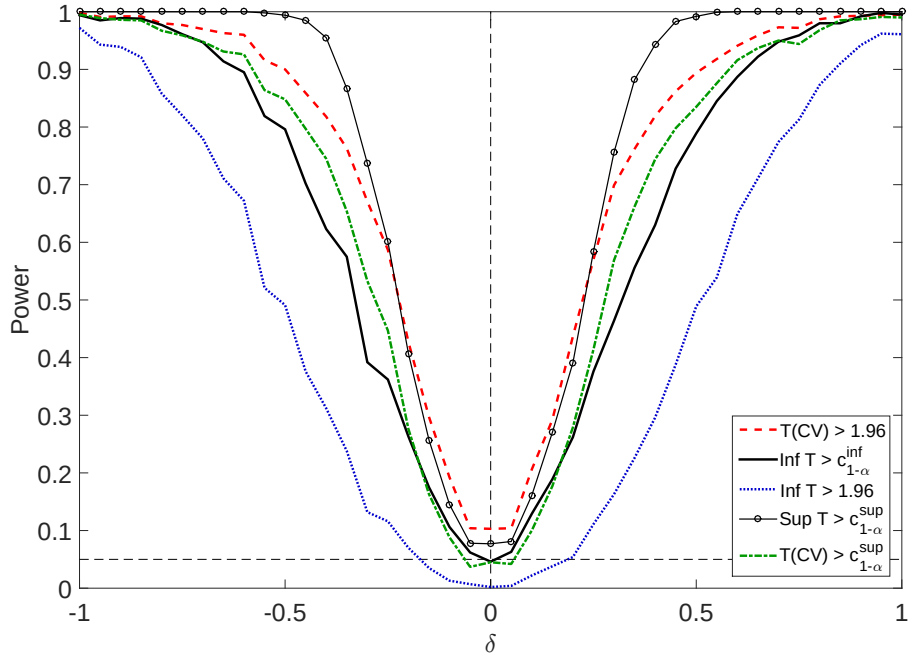
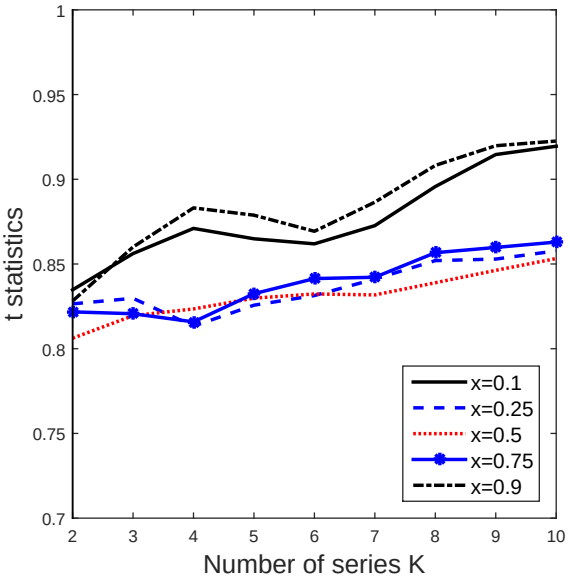
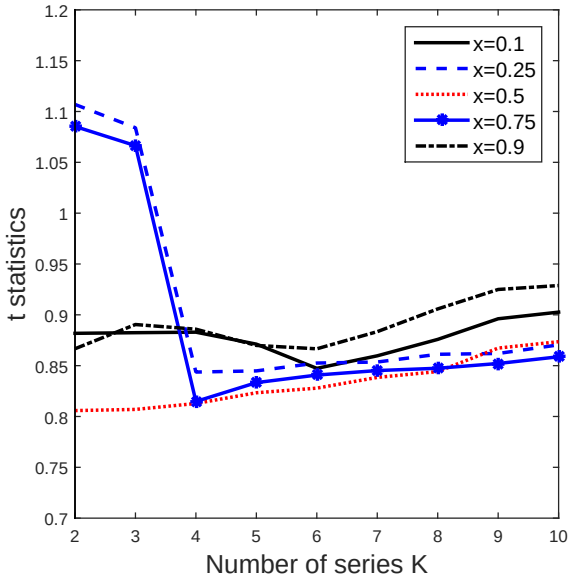


Figure 1.8: Patterns of t-statistics with K - Polynomials.
Plots of $E[T_n(K, \theta_0)]$ as a function of K at different points $x = 0.1, 0.25, 0.5, 0.75, 0.9$.
Vertical Lines are median of selected K by cross validation (The first two graphs coincide with $\underline{K}$).

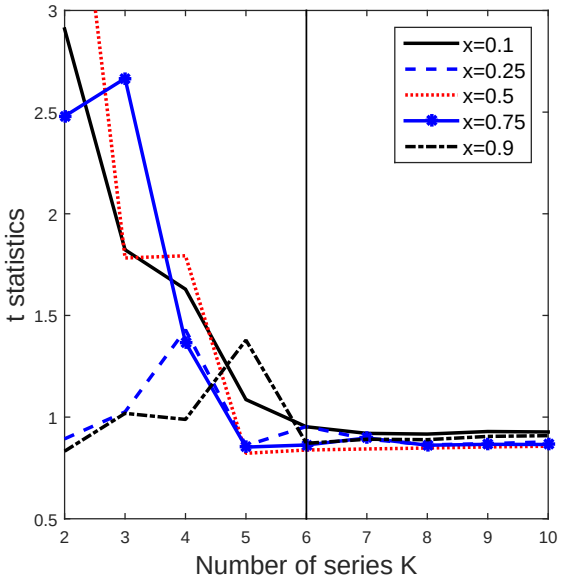
(a) $g_1(x) = 4x - 1$



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

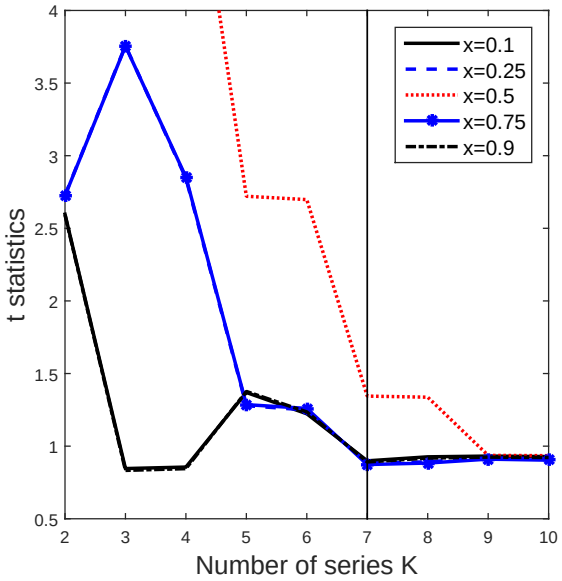


Table 1.1: Nonparametric Wage Elasticity of Hours of Work
Estimates in Blomquist and Newey (2002, Table 1). Wage elasticity evaluated at the mean wage and income.

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.

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

1.14 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. Specification search is also widely used in estimating the parametric model in a less clear way. Nonparametric series estimation gives systematic way of doing specification search by restricting domain of search as $K \in [\underline{K}, \bar{K}]$ (see Ichimura and Todd (2007) for pointing this out). However, even though the specification search are extensively used in nonparametric series estimation, little justification has been done, especially for the inference problems.

Suppose a researcher reports only 'favorable' subset of positive results and hiding large different specifications which shows 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 usual standard normal critical value must be adjusted to control size. The importance of this 'model uncertainty' introduced by specification search (or data mining/ data snooping) has been widely 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 size of the joint test in multiple testing literature.

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. Our 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 literature.

We consider a following 'supremum' t-statistic

$$\operatorname{Sup} T_n(\theta) \equiv \sup_{K \in \mathcal{K}_n} |T_n(K, \theta)|. \tag{1.14.1}$$

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. 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 $K \in \mathcal{K}_n$. Here, controlling FWE corresponds to control following probability asymptotically, $FWE = P(\text{reject at least one hypothesis } H_{K,0}, K \in \mathcal{K}_n) \leq \alpha$.

To derive 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 1.6. *1. Under Assumptions 1.1-1.2 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 1.1. In addition, if Assumption 1.3 holds, then* $\operatorname{Sup} T_n(\theta_0) \xrightarrow{d} \xi_{\sup} = \sup_{\pi \in [\underline{\pi}, 1]} |\mathbb{T}(\pi)|$.

2. Suppose Assumptions 1.2 and 1.4 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* $Z \sim N(0, I_M)$ *and* $\nu = (\nu(1), \cdots, \nu(M))'$ *defined in Theorem 1.2. If*

$\sup_m |\nu(m)| = \infty$, *then* $\operatorname{Sup} T_n(\theta_0) \xrightarrow{p} \infty$.

Corollary 1.6-2 shows that $\operatorname{Sup} T_n(\theta_0)$ converges in probability to infinity under alternative set assumption. This implies that the supremum of the t-statistics may 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 the Corollary 1.2.

Corollary 1.7. *1. Under Assumptions 1.1-1.3, following holds*

$$\limsup_{n\to\infty} P(\operatorname{Sup} T_n(\theta_0) > c^{\sup}_{1-\alpha}) = \alpha. \tag{1.14.2}$$

2. Under Assumptions 1.1-1.2, and $\sup_\pi |\nu_\pi| < \infty$, *following holds*

$$\limsup_{n\to\infty} P(\operatorname{Sup} T_n(\theta_0) > c^{\sup}_{1-\alpha}) \geq F(c^{\sup}_{1-\alpha}, \sup_\pi |\nu(\pi)|) \tag{1.14.3}$$

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

3. Under Assumptions 1.2, 1.4, 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 $\operatorname{Inf} T_n(\theta)$,(1.14.3) in Corollary 1.7-2 shows that the test based on $\operatorname{Sup} T_n(\theta)$ may suffer from the asymptotic bias, thus lead to size distortions. 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 Corollary 1.7-3 under different rate conditions for the $\mathcal{K}_n$ in Assumption 1.4. If $\sup_m |\nu(m)| = \infty$, then the asymptotic size of the test is equal to 1. The asymptotic size of the test based on $\operatorname{Sup} T_n(\theta)$ may be sensitive to the large asymptotic bias, and this leads to the over-rejection of the test.

Next, we define $CI_{\texttt{sup}}$ based on $\operatorname{Sup} T_n(\theta)$ and the critical value $\widehat{c}^{\sup}_{1-\alpha}$ in Section 1.6.

$$\begin{aligned} CI_{\texttt{sup}} &\equiv \{\theta : \sup_{K\in\mathcal{K}_n} |T_{n,\widehat{V}}(K,\theta)| \le \widehat{c}^{\sup}_{1-\alpha}\} \\ &= \bigcap_{K\in\mathcal{K}_n} \{\theta : |T_{n,\widehat{V}}(K,\theta)| \le \widehat{c}^{\sup}_{1-\alpha}\} = [\sup_K(\widehat{\theta}_K - \widehat{c}^{\sup}_{1-\alpha} s(\widehat{\theta}_K)), \inf_K(\widehat{\theta}_K + \widehat{c}^{\sup}_{1-\alpha} s(\widehat{\theta}_K))]. \end{aligned} \tag{1.14.4}$$

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

Corollary 1.8. *1. Under Assumptions 1.2, 1.5, 1.6, and 1.7,*

$$\liminf_{n\to\infty} P(\theta_K \in [\widehat{\theta}_K \pm \widehat{c}^{\sup}_{1-\alpha} s(\widehat{\theta}_K)] \quad \forall K \in \mathcal{K}_n) = 1-\alpha. \tag{1.14.5}$$

In addition, if Assumption 1.3 (undersmoothing) holds,

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{\texttt{sup}}) = \liminf_{n\to\infty} P(\theta_0 \in [\widehat{\theta}_K \pm \widehat{c}^{\sup}_{1-\alpha} s(\widehat{\theta}_K)] \quad \forall K \in \mathcal{K}_n) = 1-\alpha. \tag{1.14.6}$$

2. Under Assumptions 1.2, 1.5, 1.6, 1.7, and $\sup_m |\nu(m)| < \infty$ *where* $\nu(m) \equiv \nu(\pi_m)$ *as in* (1.3.3),

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{\texttt{sup}}) \le 1 - F(c^{\sup}_{1-\alpha}, \sup_m |\nu(m)|). \tag{1.14.7}$$

3. Under Assumptions 1.2, 1.4, 1.7, and $\sup_m |\nu(m)| = \infty$, $\liminf_{n\to\infty} P(\theta_0 \in CI_{\texttt{sup}}) = 0$.

Note that (1.14.5) gives asymptotic coverage of the uniform confidence intervals over $K \in \mathcal{K}_n$ for the pseudo-true value θ_K. (1.14.6) 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 $K \in \mathcal{K}_n$. By using an appropriate critical value from the distribution of $\operatorname{Sup} T_n$, (1.14.5) and (1.14.6) show that joint coverage of CIs, $CI_K = [\widehat{\theta}_K \pm \widehat{c}^{\sup}_{1-\alpha} s(\widehat{\theta}_K)]$,

$K \in \mathcal{K}_n$ for the pseudo-true value θ_K (for true parameter θ_0 with undersmoothing) is equal to $1-\alpha$, asymptotically.

However, Corollary 1.8-2 and 1.8-3 show that the coverage can be sensitive to the asymptotic bias. Especially, uniform coverage results based on $\mathrm{Sup}\, T_n$ in (1.14.6) can be highly sensitive to the finite sample bias when some small $K \in \mathcal{K}$ has large bias (when some K violate undersmoothing assumption), so that the coverage probability can be far below than the nominal level. Recall that $CI_{\texttt{sup}}$ is constructed by intersection of all confidence intervals in $\mathcal{K}_n$ using larger critical value $\widehat{c}^{\,\texttt{sup}}_{1-\alpha}$ than the normal critical value. Intersection can give tighter CI, however, if one of the estimator has a large bias, resulting CI can be too narrow to cover the true parameter. In worst scenario, intersection can be empty sets so that the coverage of uniform CIs can be 0. This was formally stated in 1.8-3. Under Assumption 1.4, if $|\nu(m)| = \infty$ for some m then asymptotic coverage probability of $CI_{\texttt{sup}}$ is exactly 0. For the testing problem in Corollary 1.7, over-rejection property of $\mathrm{Sup}\, T_n$ was also demonstrated.

Proof of the Results in Section 1.14

Proof of Corollary 1.6

Proof. The first part follows from Theorem 1.1 and continuous mapping theorem similar to the proof of Corollary 1.1. For the second part of Corollary 1.6, consider $S^1(t) = \sup_m |t_m|$ for $t = (t_1, \cdots, t_M)$ similarly as in the proof of Corollary 1.1. We have

$$\mathrm{Sup}\, T_n(\theta_0) = \sup_m |T_n(K_m, \theta_0)| = S^1(\bar{T}_n(\theta_0)). \tag{1.14.8}$$

Under the assumpion $\sup_m |\nu(m)| < \infty$, $S^1(t)$ is continuous at all $t \in \mathbb{R}^M$. Therefore, following holds

$$\mathrm{Sup}\, T_n(\theta_0) \xrightarrow{d} S^1(Z+\nu) = \sup_m |Z_m + \nu(m)| \tag{1.14.9}$$

by Theorem 1.2 and continuous mapping theorem. If $|\nu_m| = +\infty$ for some m, then then $|T_n(K_m, \theta_0)| \xrightarrow{p} +\infty$, therefore $\mathrm{Sup}\, T_n(\theta_0) \xrightarrow{p} +\infty$. *Q.E.D.*

Proof of Corollary 1.7

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

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

by Corollary 1.6-1. Next, without assuming Assumption 1.3, we have

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

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 1.7-3 under alternative set assumption. If $\sup_m |\nu(m)| = \infty$, then $\mathrm{Sup}\, T_n(\theta_0) \xrightarrow{p} +\infty$ by Corollary 1.6-2. Thus, for any $0 < c < \infty$, $\limsup_{n\to\infty} P(\mathrm{Sup}\, T_n(\theta_0) > c) = 1$. *Q.E.D.*

Proof of Corollary 1.8

Proof. This follows from Corollary 1.3 and Corollary 1.7 similar to the proof of Corollary 1.5. Recall that the t-statistic in (1.4.11) can be written as

$$T_{n,\widehat{V}}(K,\theta_0) = \frac{\sqrt{n}(\widehat{\theta}_K - \theta_0)}{\widehat{V}_K^{1/2}} = \frac{\sqrt{n}(\widehat{\theta}_K - \theta_K)}{\widehat{V}_K^{1/2}} + \frac{\sqrt{n}r_K}{\widehat{V}_K^{1/2}} \tag{1.14.10}$$

First, consider (1.14.5),

$$\liminf_{n\to\infty} P(\theta_K \in [\widehat{\theta}_K \pm \widehat{c}^{\text{sup}}_{1-\alpha} s(\widehat{\theta}_K)] \quad \forall K \in \mathcal{K}_n) \tag{1.14.11}$$

$$= \liminf_{n\to\infty} P(|\frac{\sqrt{n}(\widehat{\theta}_K - \theta_K)}{\widehat{V}_K^{1/2}}| \leq \widehat{c}^{sup}_{1-\alpha} \quad \forall K \in \mathcal{K}_n) = \liminf_{n\to\infty} P(\sup_K |\frac{\sqrt{n}(\widehat{\theta}_K - \theta_K)}{\widehat{V}_K^{1/2}}| \leq \widehat{c}^{sup}_{1-\alpha}) \tag{1.14.12}$$

$$= P(\sup_m |Z_m| \leq c^{sup}_{1-\alpha}) = 1-\alpha \tag{1.14.13}$$

where the last equality follows from Theorem 1.1 and Corollary 1.3 under Assumptions 1.2, 1.5, 1.6, and 1.7. Under Assumption 1.3, we have that

$$\liminf_{n\to\infty} P(\theta_0 \in CI_{\texttt{sup}}) = \liminf_{n\to\infty} P(\text{Sup}\, T_n(\theta_0) \leq \widehat{c}^{sup}_{1-\alpha}) \tag{1.14.14}$$

$$(= \liminf_{n\to\infty} P(|T_{n,\widehat{V}}(K,\theta_0)| \leq \widehat{c}^{sup}_{1-\alpha} \quad \forall K \in \mathcal{K}_n)) \tag{1.14.15}$$

$$= P(\sup_m |Z_m| \leq c^{sup}_{1-\alpha}) = 1-\alpha. \tag{1.14.16}$$

This completes the first part of Corollary 1.8. The second part can be shown similarly to the proof of Corollary 1.7-2. For the last part, if $\sup_m |\nu(m)| = \infty$, then $\liminf_{n\to\infty} P(\theta_0 \in CI_{\texttt{sup}}) = 0$ by Corollary 1.7-3 and Corollary 1.3.

Q.E.D.

Chapter 2

Higher Order Approximation of IV Estimators with Locally Invalid Instruments

2.1 Introduction

This paper studies the instrument selection problem in an instrumental variable (IV) model with many instruments and their potential invalidity. Many empirical setups in the IV model involve large sets of potential instruments and debates about their validity, which I refer to as the exogeneity condition, i.e., instruments are uncorrelated with the error term in the structural equation.

Finite sample performance of the IV estimator is sensitive to the choice of instruments. The well-known finite sample bias-variance trade-off exists when choosing among valid instruments: using more instruments reduces asymptotic variance and thus achieve efficiency, but it may increase the finite sample bias of the IV estimator. The finite sample bias of the two-stage least squares (2SLS) estimator and the generalized method of moments (GMM) estimator are proportional to the number of instruments; the bias becomes more severe when

the instruments are weak (see Morimune (1983); Bekker (1994); Staiger and Stock (1997); Newey and Smith (2004); Chao and Swanson (2005); and Hansen, Hausman and Newey (2008)).

In addition to the many instruments issue, detecting instruments that are not valid and excluding them is also important for consistent estimation and inferences. In practice, researchers carefully pick their instruments, and firmly believe that their instruments are valid, by institutional features or by the nature of the experimental (or quasi-experimental) design. However, seemingly valid instruments can be correlated with an unobserved error term, and thus are invalid. The validity of instruments is, generally, uncertain; the reasons behind this are at least two-folds. First, instruments may have direct effects on the outcome variables. Second, model misspecification can make instruments invalid. There may also be omitted control variables that are highly correlated with instruments.[1]

The purpose of this paper is to develop an instrument selection criterion that addresses these two issues together. The main contributions of this paper are as follows. 1) I derive a higher-order mean square error (MSE) approximation of the IV estimators including 2SLS estimator, limited information maximum likelihood (LIML), modification of Fuller, and bias-adjusted version of the 2SLS (B2SLS) estimator in linear IV model with many instruments, allowing possible locally invalid instruments. 2) Based on these higher-order approximations, I propose an *Invalidity-Robust Criterion* (IRC) that can be used in empirical practice to choose instruments. The IRC captures two sources of finite sample bias at the same time: bias from using many instruments and bias from using invalid instruments. Thus, the criterion is robust to potentially invalid instruments than the existing literature that assumes an instrument's validity *a priori*. Furthermore, we expect to have better finite sample performance than existing consistent moment selection methods or criteria based on first-order asymptotics, which do not consider finite sample bias from using many instruments.

[1]For questionable IVs with potential invalidity in various empirical applications, see Guggenberger (2012, Section 2.1) and references therein. See also Kolesár et al. (2014) for an interesting empirical application with invalid instruments, even when the instruments are assigned randomly.

Our question was originally motivated from two seminal papers by Donald and Newey (2001) and Andrews (1999). Donald and Newey (2001) developed optimal instrument selection methods among a set of many valid instruments. They derived higher-order MSE of IV estimators, and their criteria are based on these higher-order approximations. However, they assume all instruments are valid. If some set of instruments are invalid, then using these criteria may result in incorrect sets.

Andrews (1999) developed consistent moment selection procedures in GMM setup with valid and invalid moment conditions. Analogous to the widely used model selection criteria, they showed BIC (Bayesian) type criterion consistently select valid moment conditions. They also considered downward and upward testing procedures originating from Sargan-Hansen's over identification test, which is often used in empirical research to choose moment conditions. However, they assume a fixed number of moment conditions, so performance is questionable when the number of moments is large, and the original criteria may need some modifications. Moreover, all criteria based on consistent moment selection and first-order asymptotics will include all valid instruments and locally invalid ones by construction, which may raise the finite sample bias in many instruments setup.

The IRC captures the best of both of worlds by considering the potential invalidity of the instrument in the selection criterion as well as higher-order bias and variance from many instruments. Even if all instruments are truly valid, our method may suggest nearly the same set of instruments as the criterion of Donald and Newey (2001) does, without large additional computational costs.

Implementing our criterion involves some preliminary estimates and reduced-form criteria, such as Mallows' (1973) C_p or cross-validation (CV), which are easy to implement in practice. I also show that optimality of the choice of instruments selected by original Donald and Newey (2001) criteria under certain locally invalid instrument specifications. When we consider drifting sequences faster than $N^{1/2}$ invalid instruments specification, I show that higher-order MSE approximation reduces to the one in Donald and Newey (2001).

2.1.1 Related Literature

There are many influential papers about instrument and/or weight selection in the IV (or GMM) model. Donald, Imbens, and Newey (2009) provided moment selection criteria for the GMM, generalized empirical likelihood estimator (GEL), and continuous updating estimator (CUE) in conditional moment restriction models. Kuersteiner (2012) extended the moment selection problem to linear time series models using a kernel-weighted GMM. Okui (2011) considered shrinkage parameter selection of the first-stage shrinkage IV estimator, and Canay (2010) addressed simultaneous moments and weight selection by using a kernel-weighted GMM with a flat-top kernel. Also, Kuersteiner and Okui (2010) developed an optimal weight selection for the first-stage prediction model averaging IV estimator. Lee and Zhou (2014) considered averaged IV estimators. Another direction was considered by Carrasco (2012). They analyzed the asymptotic properties of regularized IV estimators and smoothing parameter selection while keeping all the instruments. See also the bootstrap approach by Inoue (2006), and the random effects approach by Chamberlain and Imbens (2004). However, all these papers developed selection criteria among valid instruments.

A different and important direction of studies considered moment selection criteria to separate valid moments from the set of invalid moments. Andrews and Lu (2001) extended the ideas of Andrews (1999) to the simultaneous selection of moments and regressors, and Hong, Preston, and Shum (2003) extended to the GEL estimator. Hall and Peixe (2003) proposed selecting valid and relevant moment conditions by a sequential combination of their canonical correlations information criteria with Andrews' (1999) method. Recently, Liao (2013) developed consistent moment selection by using a shrinkage-type GMM estimator. However, all these papers only deal with a fixed number of moment conditions; they do not address higher-order bias from using many moments.

Furthermore, consistent moment selection methods may include all locally invalid moments because these moments are all asymptotically valid. Including slightly invalid moments might increase finite sample bias, but help to reduce variance. An important contribution in

this direction is the recent work of DiTraglia (2014), who developed moment selection criteria based on the first-order asymptotic MSE with possible locally invalid moment conditions in GMM setups. As I mentioned, criteria based on first-order asymptotics include all valid instruments by construction. Our higher-order MSE includes first-order asymptotic MSE as well as higher-order bias and variance terms, which other selection criteria do not have. Another important advantage of higher-order approximations is that under a certain rate of drifting sequences faster than $N^{-1/2}$, only higher-order approximations can capture bias-variance trade-off from many and potentially invalid instruments (see Section 2.4). However, our method should be viewed as complementary to existing literature since they all address different aspects of the problems and/or consider more general setups.

Our results also complement the recent moment selection literature based on high-dimensional estimation and model selection methods (e.g., the Lasso estimator and Dantzig selector). Belloni et al. (2012) proposed the Lasso-based method to construct first-stage optimal instruments, considering only valid instruments. Gautier and Tsybakov (2014) provided an estimation and inference technique allowing invalid instruments based on the Dantzig selector. Some advantages of these methods are that their results do not rely on prior knowledge of the order of the instruments, and can be applied even when the number of instruments is much larger than the sample size. In GMM setups, Caner, Han, and Lee (2013) proposed simultaneous model and moment selection using an adaptive elastic net estimator. Finally, the work of Cheng and Liao (2014) is closely related to our paper. They developed consistent moment selection methods based on adaptive GMM shrinkage estimation, which consistently select valid and relevant moments, even allowing the number of moment conditions to grow with the sample size. However, their method may not be able to avoid finite sample bias from many relevant moment conditions. If there are truly many valid and relevant instruments, the bias of the 2SLS estimator using all those instruments is likely to be large. This situation can be thought of as a violation of the so-called 'sparsity' assumption in the Lasso literature, and requires careful attention (Hansen (2013)). Our method differs

from all these papers as I focus on the higher-order MSE approximations of the IV estimator.

Our paper also contributes to the literature that considered estimation and inference issues coping with invalid instruments. Many papers deal with the size distortion of testing problems (Berkowitz, Caner, and Fang (2008, 2012) and Guggenberger (2012)) and estimation issues with local violation of exogeneity conditions (Hahn and Hausman (2005) and Caner (2014)). Different types of estimation and inference methods were also proposed by Conley, Hansen, and Rossi (2012), Kolesár et al. (2014), Kraay (2012), and Nevo and Rosen (2012). However, they all focused on the estimation and inference rather than instrument selection.

Moreover, our MSE approximation extends the results of Hahn and Hausman (2005). They derived first-order asymptotics of the 2SLS estimator with scalar endogenous variable and normal error terms under similar many and locally invalid instruments setup considered here. Our by-product result can be used as MSE (or bias) comparison with the ordinary least-squares (OLS) estimator and the 2SLS estimator for more general setups, such as vector endogenous variables and non-normal error cases.

The outline of the paper is as follows. Section 2.2 introduces the basic model setup and notation. Section 2.3 describes formal higher-order MSE approximations of the IV estimators. Section 2.4 provides MSE approximations under different local sequence of invalid instruments, and Section 2.5 gives first-order approximation of 2SLS under different rates of the number of instruments. Section 2.6 discusses how to estimate our selection criterion and proposes detailed implementation procedures, and Section 2.7 concludes. All proofs are provided in Section 2.9.

2.2 Linear IV Model With Locally Invalid Instruments

I consider following linear IV model similar to Donald and Newey (2001) allowing potentially invalid instruments,

$$
\begin{aligned}
&y_i = Y_i'\theta_0 + x_{1i}'\beta_0 + \varepsilon_i = W_i'\delta_0 + \varepsilon_i, \\
&W_i = f(x_i) + u_i = \begin{pmatrix} \mathbb{E}(Y_i|x_i) \\ x_{1i} \end{pmatrix} + \begin{pmatrix} \xi_i \\ 0 \end{pmatrix}, \qquad \delta_0 = (\theta_0', \beta_0')', \\
&\varepsilon_i = \mathbb{E}(\varepsilon_i|x_i) + v_i = \frac{g(x_i)}{\sqrt{N}} + v_i, \qquad \mathbb{E}(v_i|x_i) = 0, \qquad \text{for} \quad i = 1, \cdots, N,
\end{aligned}
\tag{2.2.1}
$$

where y_i is a scalar outcome variable and W_i is a $p \times 1$ vector that includes endogenous variables Y_i and $d \times 1$ vector of exogenous variables x_{1i}. $\delta_0 \in \mathbb{R}^p$ is a parameter of interest and x_{1i} are assumed to be a subset of the potential exogenous variables x_i. Here, p and d are finite and fixed, i.e., they don't change with sample size N.

The last line in (2.2.1) indicates a model specification allowing *locally invalid* instruments. In this model, $\mathbb{E}(\varepsilon_i|x_i)$ is not necessarily zero for any finite N unless $g(x_i) = 0$, therefore the unconditional moment condition $\mathbb{E}(\psi(x_i)\varepsilon_i) = 0$ does not necessarily hold for potential instruments $\psi(x_i)$. However, with this definition of local invalidity of x_i, all potential instruments are asymptotically valid, i.e., $\mathbb{E}(\psi(x_i)\varepsilon_i) \to 0$ as $N \to \infty$.

Here, I consider the local-to-zero specification and provides a MSE approximation theory for the IV estimators centered with δ_0, not the pseudo-true parameter that is the probability limit of each IV estimator. Under the local misspecification setup with drifting sequences $g(x_i)/N^{-\gamma}$ for $\gamma \geq 1/2$, all IV estimators considered in this paper are consistent under the standard rate conditions of the number of instruments that increases slower than the sample size.

We may consider *global invalidity* of instruments with the rates $N^{-\gamma}$ with $\gamma = 0$ or other rates with $0 < \gamma < 1/2$, and provide a MSE approximation centered with pseudo-true value (or sequence of pseudo-true value depending on sample size). However, such theory may not

be useful to investigate bias-variance trade-off of IV estimators as there is no "bias" term arising from invalid instruments if we focus on the pseudo-true value. Moreover, different choice of instruments and different IV estimators can lead to a different pseudo-true value, which makes us hard to compare MSE approximations across different number of instruments as well as different estimators. As $g(x_i)$ can be allowed to be large numbers for any finite N, our local misspecification setup may not significantly restrict our approximation theory by providing finite sample behavior of IV estimators with invalid instruments. Particularly with this knife-edge rate, $N^{-1/2}$, the stochastic order of the bias from a locally invalid instrument is $O_p(1)$ that is equal to those of the first-order asymptotic variance. In Section 2.4, I also provide an approximation theory for the sequences $N^{-\gamma}$ with $\gamma > 1/2$.

2.3 Higher-Order MSE Approximation with Locally Invalid Instruments

I first characterize the higher-order MSE formula for the IV estimators under similar regularity conditions that are imposed in Donald and Newey (2001). Recall the model (2.2.1) in previous section with vector forms

$$\begin{aligned} y &= W\delta_0 + \varepsilon = W\delta_0 + g/\sqrt{N} + v, \\ W &= f + u, \end{aligned} \tag{2.3.1}$$

where $y = (y_1, \cdots, y_N)', Y = [Y_1, \cdots, Y_N]', X_1 = [x_{11}, \cdots, x_{1N}]', W = [Y, X_1], f = [f_1, \cdots, f_N]', f_i = f(x_i)$, and $g = [g_1, \cdots, g_N]', g_i = g(x_i)$. I also define $H = f'f/N, H_g = f'g/N$, $X = [x_1, ..., x_N]', \sigma_{uv} = \mathbb{E}(u_i v_i | x_i), \sigma_v^2 = \mathbb{E}(v_i^2 | x_i), \sigma_\varepsilon^2 = \mathbb{E}(\varepsilon_i^2 | x_i)$ and $\Sigma_u = \mathbb{E}(u_i u_i' | x_i)$. Let A^- denotes any generalized inverse of A. Also, $o_p(\cdot)$ and $O_p(\cdot)$ denote the usual stochastic order symbols, convergence in probability, and bounded in probability, respectively.

I define $\psi_i^K \equiv \psi^K(x_i) = (\psi_{1K}(x_i), ..., \psi_{KK}(x_i))'$ as a $K \times 1 (K \geq d)$ vector of instrumental variables (or basis functions). Throughout the paper, I assume that ψ_i^K includes exogenous

variables x_{1i}. For notational simplicity, K indicates both the number of instruments and the index of the instrument sets ψ_i^K.

We first consider the 2SLS estimator,

$$\widehat{\delta}_{2SLS}(K) = (W'P^KW)^{-1}(W'P^Ky),$$

where $P^K = \Psi^K(\Psi^{K'}\Psi^K)^-\Psi^{K'}$ is the projection matrix for the instrument vector $\Psi^K = [\psi_1^K, ..., \psi_N^K]'$. Next, we consider LIML estimator,

$$\widehat{\delta}_{LIML}(K) = (W'P^KW - \hat{\Lambda}(K)W'W)^{-1}(W'P^Ky - \hat{\Lambda}(K)W'y),$$

where

$$\hat{\Lambda}(K) = \min_{\delta} \frac{(y - W\delta)'P^K(y - W\delta)}{(y - W\delta)'(y - W\delta)}.$$

We also consider modified Fuller's (1977) estimator (FULL),

$$\widehat{\delta}_{FULL}(K) = (W'P^KW - \check{\Lambda}(K)W'W)^{-1}(W'P^Ky - \check{\Lambda}(K)W'y),$$

where

$$\check{\Lambda}(K) = \frac{\hat{\Lambda}(K) - C/N(1 - \hat{\Lambda}(K))}{1 - C/N(1 - \hat{\Lambda}(K))}.$$

for some constant C. Popular choices are $C = 1$ or $C = 4$. Finally, we consider B2SLS estimator suggested by Donald and Newey (2001) as a modification of the Nagar (1959) estimator with $\bar{\Lambda}(K) = (K - d - 2)/N$,

$$\widehat{\delta}_{B2SLS}(K) = (W'P^KW - \bar{\Lambda}(K)W'W)^{-1}(W'P^Ky - \bar{\Lambda}(K)W'y).$$

Following Donald and Newey (2001), I derive the Nagar (1959) type higher-order asymptotic MSE for the IV estimators. Specifically, I will find a decomposition for $\hat{\delta}(K)$ with following form:

$$\begin{aligned} &N(\widehat{\delta}(K)-\delta_0)(\widehat{\delta}(K)-\delta_0)' = \widehat{Q}(K)+\widehat{r}(K), \\ &\mathbb{E}(\widehat{Q}(K)|X) = \sigma_v^2 H^{-1} + H^{-1}H_g H_g' H^{-1} + L(K) + T(K), \\ &[\widehat{r}(K)+T(K)]/tr(L(K)) = o_p(1), \quad K\to\infty, N\to\infty. \end{aligned} \tag{2.3.2}$$

The dominating terms in the conditional MSE approximation in (2.3.2) are $\sigma_v^2 H^{-1}$ and $H^{-1}H_g H_g' H^{-1}$ that corresponds to the standard first-order asymptotic variance and the square of the asymptotic bias from locally invalid instrument, respectively. Note that these are the dominating terms that does not depend on K in our large K approximation. Next leading term in the MSE approximation, $L(K)$, includes the higher-order bias and variance terms due to many and invalid instruments, and has different form with each IV estimator. $\widehat{r}(K)$ and $T(K)$ are the remainder terms goes to 0 faster than $S(K)$.[2]

To derive specific terms in $L(K)$, I impose the following assumptions. Define $\|A\| = \sqrt{tr(A'A)}$ as an Euclidean norm.

Assumption 2.1. *$\{y_i, Y_i, x_i\}_{i=1}^N$ are independent and identically distributed (i.i.d.). $\mathbb{E}(v_i^2|x_i) = \sigma_v^2 > 0$, and $\mathbb{E}(\|\xi_i\|^4|x_i), \mathbb{E}(|v_i|^4|x_i)$ are bounded.*

Assumption 2.2. *(i) $\bar{H} = \mathbb{E}(f_i f_i')$ exists and is nonsingular, $\bar{H}_g = \mathbb{E}(f_i g_i)$ exists. (ii) there exists π_K, π_K^g such that $\mathbb{E}(\|f(x) - \pi_K \psi^K(x)\|^2) \to 0$ and $\mathbb{E}(|g(x) - \pi_K^g \psi^K(x)|^2) \to 0$ as $K\to\infty$.*

Assumption 2.3. *(i) $\mathbb{E}((v_i, \xi_i')'(v_i, \xi_i')|x_i)$ is constant. (ii) $\Psi^{K\prime}\Psi^K$ is nonsingular with probability approaching one. (iii) $max_{i\le N} P_{ii}^K \xrightarrow{p} 0$. (iv) f_i and g_i are bounded.*

[2]Under the global misspecification setup ($\varepsilon_i = g(x_i) + v_i$), $\hat{\delta}(K) \xrightarrow{p} \delta_0 + \bar{H}^{-1}\bar{H}_g$ where $\bar{H} = \mathbb{E}(f_i f_i')$, $\bar{H}_g = \mathbb{E}\, f_i g_i$ (assuming expectation exists). In the following results, we may have similar MSE approximations as in (2.3.2) by centering at the pseudo-true value, $\delta_0 + H^{-1}H_g$. We can easily verify that leading term $L(K)$ reduces to the results of Donald and Newey (2001) in this case, thus omitted.

Assumptions 2.1-2.3 are similar to those imposed in Donald and Newey (2001). Assumption 2.1 imposes boundedness of the fourth conditional moments of the error terms. Assumption 2.2(i) is imposed for a usual identification assumption and for the existence of the first-order bias from invalid instruments. Assumption 2.2(ii) requires the mean square approximation error of the $f(x)$ and $g(x)$ by a linear combination of instruments $\psi^K(x)$ goes to 0 as the number of instrument increases. Assumption 2.2 and 2.3 also impose homoskedasticity and restrict the growth rate of K.

Our first result, Proposition 2.1 gives the MSE approximation for the 2SLS estimator. Proposition 2.1 is a generalization of the result in Donald and Newey (2001) allowing possibly invalid instruments.

Proposition 2.1. *If Assumptions 2.1, 2.2, 2.3 are satisfied,* $\sigma_{uv} \neq 0, H_g\sigma'_{uv} \neq 0, H_g \neq 0$, *and* $K^2/N \to 0$, *then the approximate MSE for the 2SLS estimator satisfies decomposition (2.3.2) with the following terms*

$$\begin{aligned} L(K) = H^{-1}\big[&\frac{K}{N^{1/2}}(H_g\sigma'_{uv} + \sigma_{uv}H'_g) + \sigma_{uv}\sigma'_{uv}\frac{K^2}{N} + \sigma_v^2\frac{f'(I-P^K)f}{N} \\ &+ H_gH'_gH^{-1}\frac{f'(I-P^K)f}{N} + \frac{f'(I-P^K)f}{N}H^{-1}H_gH'_g \\ &- \frac{f'(I-P^K)g}{N}H'_g - H_g\frac{g'(I-P^K)f}{N}\big]H^{-1}. \end{aligned} \tag{2.3.3}$$

Moreover, ignoring terms of order $O_p(K^2/N) = o_p(K/\sqrt{N})$, *we have*

$$\begin{aligned} L(K) = H^{-1}\big[&\frac{K}{N^{1/2}}(H_g\sigma'_{uv} + \sigma_{uv}H'_g) + \sigma_v^2\frac{f'(I-P^K)f}{N} + H_gH'_gH^{-1}\frac{f'(I-P^K)f}{N} \\ &+ \frac{f'(I-P^K)f}{N}H^{-1}H_gH'_g - \frac{f'(I-P^K)g}{N}H'_g - H_g\frac{g'(I-P^K)f}{N}\big]H^{-1}. \end{aligned} \tag{2.3.4}$$

We have the following simplifications of the above result if $H_g = 0$. In this case, MSE

approximation results in (2.3.3) reduce to Proposition 1 in Donald and Newey (2001)

$$L(K) = H^{-1}[\sigma_{uv}\sigma'_{uv}\frac{K^2}{N} + \sigma_v^2\frac{f'(I-P^K)f}{N}]H^{-1}.$$

Therefore, $L(K)$ in (2.3.3) includes higher-order terms in Donald and Newey (2001) as well as additional higher-order terms because of invalid instruments. $H_g = 0$ holds if $g(x_i) = 0$ (i.e., no invalid instruments). However, $H_g = 0$ may hold allowing $g(x_i) \neq 0$ if the direct effect of the instruments to the outcome variable (g) are orthogonal to the effect of the instruments on the endogenous variable (f). Suppose $f = \Psi\pi, g = \Psi\gamma$, then $H_g = 0$ holds when $\pi'\Psi'\Psi\gamma/N = 0$ and this is closely related to the identifying assumption in Kolesar et al (2015, Assumption 5).

The second term in (2.3.2) and the first three terms of $L(K)$ in (2.3.3) are approximately the MSE of the following random vectors for large K

$$H^{-1}u'P^Kv/\sqrt{N} + H^{-1}f'g/N.$$

Their conditional expectations are $\mathbb{E}(H^{-1}u'P^Kv/\sqrt{N}|X) = H^{-1}K\sigma_{uv}/\sqrt{N}$, and $H^{-1}H_g$, respectively. Square and cross products of these two terms generate the leading terms in MSE approximation, and this correspond to the two sources of bias we consider: bias from many instruments and bias from invalid instruments. The remaining terms in $L(K)$ regarding $f'(I-P^K)f/N$ represent higher order variance term as it denotes the error of approximation of the reduced form $f(x)$ by a linear combination of instrument sets K, and it decreases as K increases. Additional higher-order variance terms appeared in $L(K)$ because of invalid instruments. Note that the instruments selection criteria without these additional terms may lead to a misleading balance of bias and efficiency. Our MSE approximation is valid under local misspecification and contains higher-order bias and variance from potentially invalid instruments in addition to the many instruments.

It is also interesting to see that the order of the bias from invalid instruments dominates

the bias from many instruments under locally invalid instruments. Moreover, as shown in the equation (2.3.4), the dominating terms in MSE approximation is order of $O_p(K/\sqrt{N})$ which is larger order than those of Donald and Newey (2001), $O_p(K^2/N)$, with all valid instruments. Terms of order $O_p(K/\sqrt{N})$ arise from cross-product of many and invalid instrument bias. However, this is not the case for the other estimators such as LIML and B2SLS because they do not possess, to the order I consider, bias terms depend on K in higher-order approximation (see discussions following Proposition 2.2 and 2.3).

Although it is not clear that increasing number of instrument increase or decrease the higher-order variance terms regarding $f'(I - P^K)g/N$, it is worth mentioning how these terms can help to reduce higher-order MSE. This is easiest to see under linear specification of $f(x)$ and $g(x)$. Suppose $f = Z_1\pi$ and $g = Z_2\gamma$ with *relevant* instrument Z_1 and *invalid* instrument Z_2 with scalar π and γ, then $\frac{f'(I-P^K)g}{N}H_g' = Z_1'(I - P^K)Z_2Z_2'Z_1(\frac{\pi\gamma}{N})^2$. If the choice of instruments ψ_K include invalid instrument Z_2, then $(I - P^K)Z_2$ is zero, thus the above term will be zero. However, if the choice of instruments is independent of the direct effects of instruments, i.e., ψ_K is orthogonal to invalid instrument Z_2, $(I - P^K)Z_2$ term is non-zero, thus may help to decrease $L(K)$ as long as sign is positive.

Next, I give the MSE approximation for the LIML and FULL estimator. Let $\eta_i = u_i - v_i\sigma_{uv}/\sigma_v^2$ and $\Sigma_\eta = \mathbb{E}(\eta_i\eta_i')$. Note that I restrict the growing rate of K that allowed in Donald and Newey (2001) for the simplification, ignoring terms of order $1/\sqrt{N}$ that is $o(K/N)$ under the rate conditions I consider.

Proposition 2.2. *If Assumptions 2.1, 2.2, 2.3 are satisfied,* $\mathbb{E}(v_i^2\eta_i|x_i) = 0$, $K/N \to 0$, $K^2/N \to \infty$, $\Sigma_\eta \neq 0, H_g \neq 0$ *and* $\mathbb{E}(\|\xi_i\|^5|x_i), \mathbb{E}(|v_i|^5|x_i)$ *are bounded, then the approximate MSE for the LIML or FULL estimator satisfies decomposition (2.3.2) with the following*

terms

$$L(K) = H^{-1}\big[\sigma_v^2\Sigma_\eta\frac{K}{N} + \sigma_v^2\frac{f'(I-P^K)f}{N} + H_gH_g'H^{-1}\frac{f'(I-P^K)f}{N} + \frac{f'(I-P^K)f}{N}H^{-1}H_gH_g' - \frac{f'(I-P^K)g}{N}H_g' - H_g\frac{g'(I-P^K)f}{N}\big]H^{-1}. \tag{2.3.5}$$

This is also an extension of Proposition 2 in Donald and Newey (2001). For the LIML or FULL estimator, $L(K)$ does not include higher order bias from many instruments, and the terms in $L(K)$ shows higher-order variance trade-off with many invalid instruments. Similar to Donald and Newey (2001), the third moment condition $\mathbb{E}(v_i^2\eta_i|x_i) = 0$ is imposed for the simplification, which holds when $(v_i, \eta_i')'$ is normally distributed. Without this moment conditions, $L(K)$ will have an additional term that could be estimated. Note also that the LIML and FULL estimator has the same approximate MSE to the order I consider here.

Next result is for B2SLS estimator. Similar to Donald and Newey (2001), MSE approximation for B2SLS is larger than MSE for LIML or FULL, which shows the higher order efficiency of LIML or FULL estimator with locally invalid instruments.

Proposition 2.3. *If Assumptions 2.1, 2.2, 2.3 are satisfied,* $\sigma_{uv} \neq 0, H_g \neq 0, \mathbb{E}(v_i^2u_i|x_i) = 0$*,* $K/N \to 0, K^2/N \to \infty$*, then the approximate MSE for the B2SLS estimator satisfies decomposition (2.3.2) with the following terms*

$$L(K) = H^{-1}\big[(\sigma_v^2\Sigma_\eta + 2\sigma_{uv}\sigma_{uv}')\frac{K}{N} + \sigma_v^2\frac{f'(I-P^K)f}{N} + H_gH_g'H^{-1}\frac{f'(I-P^K)f}{N} + \frac{f'(I-P^K)f}{N}H^{-1}H_gH_g' - \frac{f'(I-P^K)g}{N}H_g' - H_g\frac{g'(I-P^K)f}{N}\big]H^{-1}. \tag{2.3.6}$$

2.4 Higher Order MSE Approximation under Drifting Sequences Faster than $N^{1/2}$

In this section, I consider different drifting sequences of the model considered in (2.2.1). Specifically, I consider

$$\begin{aligned} y_i &= W_i'\delta_0 + \frac{g(x_i)}{N^\gamma} + v_i, \\ W_i &= f(x_i) + u_i, \end{aligned} \tag{2.4.1}$$

where $\gamma > 1/2$. With this specification, bias term $H^{-1}H_g H_g' H^{-1}$ in the decomposition (2.3.2) is now smaller order than the first-order variance, thus move to higher-order term. Higher-order approximation theory is useful to capture changes in the order of the terms in Proposition 2.1-2.3, whereas first-order asymptotic theory can not capture.

Under the model (2.4.1) with fixed γ, I will find a following decomposition for $\hat{\delta}(K)$ in this section,

$$\begin{aligned} &N(\widehat{\delta}(K) - \delta_0)(\widehat{\delta}(K) - \delta_0)' = \widehat{Q}(K) + \widehat{r}(K), \\ &\mathbb{E}(\widehat{Q}(K)|X) = \sigma_v^2 H^{-1} + G + L(K) + T(K), \\ &[\widehat{r}(K) + T(K)]/tr(G + L(K)) = o_p(1), \quad K \to \infty, N \to \infty \end{aligned} \tag{2.4.2}$$

Unlike the $N^{-1/2}$ rates considered in Section 2.3, first-order variance, $\sigma_v^2 H^{-1}$, is the only first-order asymptotic term in conditional MSE approximations (2.4.2). Next leading term in the MSE approximation includes higher-order bias from locally invalid instruments, G, that does not depend on K. Moreover, $L(K)$ is also a leading term that includes the higher-order bias and variance terms due to many instruments. It is important to note that the results generally depend not only on γ, but also on the specific rate of K allowed. For example, terms of order $O_p(1/N^{2\gamma-1})$ is dominated by $O_p(K^2/N)$ under certain rate of K, thus we can set $G = 0$, and $L(K)$ is the only higher-order leading term in the MSE approximation.

For all $\gamma > 1/2$, the following Corollary provides a higher-order MSE approximation result for 2SLS estimator.

Corollary 2.1. *Suppose Assumptions 1, 2 and 3 are satisfied with the model* (2.4.1)*. If* $K^2/N \to 0$*, and* $\sigma_{uv} \neq 0, H_g\sigma'_{uv} \neq 0, H_g \neq 0$*, then the approximate MSE for the 2SLS estimator satisfies decomposition* (2.4.2) *with* $G = \frac{1}{N^{2\gamma-1}} H^{-1}H_g H_g' H^{-1}$*, and the following dominating terms*

$$L(K) = H^{-1}\big[\frac{K}{N^{\gamma}}(H_g\sigma'_{uv} + \sigma_{uv}H_g') + \sigma_{uv}\sigma'_{uv}\frac{K^2}{N} + \sigma_v^2\frac{f'(I-P^K)f}{N}\big]H^{-1}. \tag{2.4.3}$$

We have further simplifications when $\frac{K}{N^{1-\gamma}} \to \infty$*, with* $G = 0$*,*

$$L(K) = H^{-1}\big[\sigma_{uv}\sigma'_{uv}\frac{K^2}{N} + \sigma_v^2\frac{f'(I-P^K)f}{N}\big]H^{-1}. \tag{2.4.4}$$

In the first result of Corollary 2.1 in equation (2.4.3), G and $L(K)$ contain all four higher-order bias terms and higher-order variance from many instruments. Furthermore, Corollary 2.1 shows that Donald and Newey (2001)'s MSE approximation for 2SLS estimator is robust to a very small degree of invalid instruments such that $\gamma \geq 1$. If we consider the case $\gamma \geq 1$ then $K/N^{1-\gamma} \to \infty$ always holds in the assumption of the second result, and $L(K)$ in (2.4.4) has the same form of dominating terms in the MSE approximations of Donald and Newey (2001). Therefore, for any $\gamma \geq 1$ with the same rate conditions they allow ($K^2/N \to 0$), their selection criterion based on MSE approximation still valid without estimating H_g and g. Furthermore, optimality results in Donald and Newey (2001, Proposition 4) also hold without any modifications under $\gamma \geq 1$.

This finding should not be surprising. If we consider smaller degree of invalidity than $N^{-1/2}$ invalidity, higher-order bias and variance terms from many instruments dominate all other bias terms due to the invalid instruments. Nevertheless of these intuitive results, the

relationship between the magnitude of γ and the dominating terms in $L(K)$ is not clear as they depend on the specific rate of K. Here, I quantify the magnitude of the robustness of the MSE approximation of 2SLS estimator in Donald and Newey (2001).

Next result is for LIML and FULL estimator.

Corollary 2.2. *Suppose Assumptions 2.1, 2.2 and 2.3 are satisfied with the model* (2.4.1). *Assume* $K/N \to 0, \Sigma_\eta \neq 0, H_g \neq 0, \mathbb{E}(v_i^2\eta_i|x_i) = 0$, *and* $\mathbb{E}(\|\xi_i\|^5|x_i), \mathbb{E}(|v_i|^5|x_i)$ *are bounded. Then the approximate MSE for the LIML or FULL estimator satisfies decomposition* (2.4.2) *with* $G = \frac{1}{N^{2\gamma-1}} H^{-1} H_g H_g' H^{-1}$, *and the following dominating terms*

$$L(K) = H^{-1}\big[\sigma_v^2 \Sigma_\eta \frac{K}{N} + \sigma_v^2 \frac{f'(I - P^K)f}{N}\big] H^{-1}. \tag{2.4.5}$$

Furthermore, if $\frac{K}{N^{2-2\gamma}} \to \infty$, *then LIML or FULL estimator satisfies decomposition* (2.4.2) *with* $G = 0$ *and the same* $L(K)$ *terms above.*

We have similar results for the B2SLS estimator.

Corollary 2.3. *Suppose Assumptions 2.1, 2.2 and 2.3 are satisfied with the model* (2.4.1). *Assume* $K/N \to 0, \sigma_{uv} \neq 0, H_g \neq 0$, *and* $\mathbb{E}(v_i^2 u_i|x_i) = 0$. *Then the approximate MSE for the B2SLS estimator satisfies decomposition* (2.4.2) *with* $G = \frac{1}{N^{2\gamma-1}} H^{-1} H_g H_g' H^{-1}$, *and the following dominating terms*

$$L(K) = H^{-1}\big[(\sigma_v^2 \Sigma_\eta + 2\sigma_{uv}\sigma_{uv}') \frac{K}{N} + \sigma_v^2 \frac{f'(I - P^K)f}{N}\big] H^{-1} \tag{2.4.6}$$

Furthermore, if $\frac{K}{N^{2-2\gamma}} \to \infty$, *then B2SLS estimator satisfies decomposition* (2.4.2) *with* $G = 0$ *and the same* $L(K)$ *terms above.*

Corollary 2.2 and 2.3 shows that the leading terms (that depends on K) in MSE approximation for LIML, FULL and B2SLS estimator under local invalid instruments is same as the

leading terms in Donald and Newey (2001) when $\gamma > 1/2$. This also shows that robustness of instrument selection criteria based on LIML, FULL and B2SLS in Donald and Newey (2001) under locally invalid instruments when $\gamma > 1/2$.

2.5 MSE Approximation of 2SLS under $K = O(\sqrt{N})$

This section provides MSE approximation of 2SLS under different rates of K with the sample size N. I consider faster growing rate of $K = O(\sqrt{N})$ than the rate conditions imposed in the Proposition 2.1. This rate is considered in existing many instruments literature, such as Morimune (1983) and Hahn and Hausman (2005). Note that assumption in Proposition 2.1 limit the growth rate of the number of instruments to $K = o(\sqrt{N})$, and this guarantees first-order asymptotic properties of the 2SLS estimator, where the bias from many instruments $(O_p(K/\sqrt{N}))$ is dominated by the bias from invalid instruments $(O_p(1))$. However, the bias from many instruments has the same first-order magnitude with the bias from invalid instruments under $K = O(\sqrt{N})$. Therefore, I will find a different decomposition rather than the equation (2.3.2) for this case. Specifically, in the next corollary, I will find the following first-order approximations of the conditional MSE,

$$\begin{aligned} &N(\hat{\delta}(K) - \delta_0)(\hat{\delta}(K) - \delta_0)' = \hat{Q}(K) + o_p(1), \\ &\mathbb{E}(\hat{Q}(K)|X) = \sigma_v^2 H^{-1} + H^{-1} H_g H_g' H^{-1} + L(K) + o_p(1), \qquad K \to \infty, N \to \infty. \end{aligned} \tag{2.5.1}$$

Corollary 2.4. *Suppose Assumptions 2.1-2.3 are satisfied. If $K/\sqrt{N} \to \alpha (0 < \alpha < \infty)$ and $\sigma_{uv} \neq 0$, then the approximate MSE for the 2SLS estimator satisfies decomposition (2.5.1) with following terms*

$$L(K) = H^{-1}\big[\frac{K}{\sqrt{N}}(H_g \sigma_{uv}' + \sigma_{uv} H_g') + \sigma_{uv}\sigma_{uv}' \frac{K^2}{N}\big] H^{-1}. \tag{2.5.2}$$

Corollary 2.4 is an extension of the first-order asymptotic MSE results of the Hahn

and Hausman (2005). For a linear specification of $f = X\pi, g = X\tau$ with all instruments $P^K = X(X'X)^{-1}X'$, a scalar endogenous variable W_i, Corollary 2.4 becomes

$$H^{-1}H_g H_g' H^{-1} + L(K) = (\frac{\Xi + \alpha\sigma_{uv}}{H})^2 \tag{2.5.3}$$

where $\Xi = \pi' X' X\tau/N$. $L(K)$ in equation (2.5.3) corresponds exactly to Theorem 3 of Hahn and Hausman (2005). Under the same rate conditions $K = O(\sqrt{N})$, they derived first-order asymptotic results for the 2SLS estimator when the endogenous variable Y_i is scalar, no included exogenous variables and the error terms are normally distributed. I provide an extension of their results to the general setup. Our MSE results imply that the normality, scalar endogenous variable, and linearity assumption of f and g are not essential for their result, and thus it can be applied in more general cases. This result is new if someone is interested in a MSE (or bias) comparison between the OLS and 2SLS estimators with potential invalid instruments in more general setups, e.g., multivariate endogenous variables, nonlinear reduced-form conditional expectation and non-normal error terms.

2.6 Invalidity-Robust Criteria to Choose Instruments

In this section, I propose the *invalidity-robust instrument selection criterion (IRC)* based on the MSE approximation in the Section 2.3 and 2.4. The selection of the instrument K is based on the approximation to the higher order MSE of $\lambda'\hat{\delta}$ defined in (2.3.2) or (2.4.2) for some fixed $\lambda \in \mathbb{R}^p$. Specifically, we choose K to minimize $\hat{L}_\lambda(K)$ which is an estimate of $L_\lambda(K) = \lambda' L(K)\lambda$, where $L(K)$ is a part of the dominating term in the MSE approximation.

Estimation of $L_\lambda(K)$ requires some preliminary estimates of $g(x_i)$. I assume throughout this paper that we have some known to be valid instrument sets z_i. Note that our derivation of the MSE approximation in the previous Sections does not need this assumption. The assumption of having a small number of valid instruments is also used for identification and similar estimation purposes in recent papers which address similar questions. (e.g., DiTraglia

(2014) and Cheng and Liao (2014)).

Assumption 2.4. *Assume that we have valid instrument sets z_i such that $\mathbb{E}(z_i\varepsilon_i) = 0$ holds, where z_i is $q \times 1$ vector including x_{1i} and $q \geq p$.*[3]

These *conservative sets*, z_i is only for preliminary estimates of H_g and $g(\cdot)$. Our method still allows choosing among z_i when they are also large dimensions. With known to be valid instrument sets z_i, we can have an asymptotically unbiased estimator of $g(x_i)$.

An empirical researcher may want to use only conservative sets if they are already known. Assumption of having small set of valid instruments may be restrictive in some applications, although it would be less restrictive than assuming all instruments' validity *a priori*. We may obtain better instrument sets that have lower MSE of the estimator if we could distinguish valid instruments from *questionable sets* and choose an optimal number of instruments among many valid instruments. By considering the bias-variance trade-off of many and invalid instruments, it is possible that including more instrument from questionable set increase or decrease MSE of the estimator. Our goal is to find the best instrument choices which minimize the MSE of the IV estimator. Even if instrument sets are all (locally) invalid, i.e. $\mathbb{E}(\psi_i^K\varepsilon_i) \neq 0$ for all choice of ψ_i^K, our approximation theory still can provide guidelines to find the best instruments which has the smallest MSE among IV estimators with given (locally) invalid instrument sets.

Similar to the Donald and Newey (2001), estimating the MSE requires preliminary estimates of some parameters of the model and goodness of fit criterion for the first-stage reduced form equation. Let $\tilde{\delta}$ be some preliminary estimator, e.g., the IV estimator using all available instruments, or IV estimator where the instruments $\tilde{K}$ are chosen to minimize the first-stage CV or Mallows' criteria. Let $\tilde{\varepsilon}$ as residuals $\tilde{\varepsilon} = y - W\tilde{\delta}$, and let $\hat{H} = W'P^{\tilde{K}}W/N$

[3]Set of instruments can satisfy moment conditions $\mathbb{E}(z_i\varepsilon_i) = 0$, for example, if the direct effect of instrument $g(x_i)$ is only a function of a subset of x_i, and if z_i are uncorrelated with (or independent of) the invalid instruments. In certain case, we do not need known valid sets. If endogenous variables Y_i are scalar, and explanatory variables x_{1i} are exogenous and do not include intercepts, then we can use $z_i = (1, x_{1i}')'$ as known valid instruments.

as a prelim estimator of $H = f'f/N$. Also, let $\tilde{u} = (I - P^{\tilde{K}})W$ as a preliminary residual vector of the first-stage reduced-form regression. Define $\tilde{u}_\lambda = \tilde{u}\hat{H}^{-1}\lambda$,[4] and

$$\hat{\sigma}_v^2 = \tilde{\varepsilon}'\tilde{\varepsilon}/N, \quad \hat{\sigma}_u^2 = \tilde{u}'\tilde{u}/N, \quad \hat{\sigma}_{uv} = \tilde{u}'\tilde{\varepsilon}/N, \quad \hat{\sigma}_{u_\lambda}^2 = \tilde{u}_\lambda'\tilde{u}_\lambda/N, \quad \hat{\sigma}_{u_\lambda v} = \tilde{u}_\lambda'\tilde{\varepsilon}/N.$$

Define also IV estimator $\hat{\delta}^*$ with known valid instruments z, and residuals as $\hat{\varepsilon}^* = y - W\hat{\delta}^*$. For example, 2SLS estimator $\hat{\delta}^* = (W'P^Z W)^{-1}(W'P^Z y)$ where $Z = [z_1, \cdots, z_n]', P^Z = Z(Z'Z)^{-1}Z'$. Finally, let $\hat{H}_g = W'P^{\tilde{K}}\hat{\varepsilon}^*/\sqrt{N}$ as a prelim estimator of $H_g = f'g/N$, let $\lambda'\hat{H}^{-1}\hat{H}_g = \hat{H}_{g\lambda}$. It is important to note that all of these preliminary estimates remain fixed (does not depend on K) while the criterion is calculated for different sets of instruments. I use corresponding IV estimator for each criterion. Based on Proposition 2.1-2.3, the *invalidity-robust criterion (IRC)* $\hat{L}_\lambda(K)$ is[5]

$$\begin{aligned}
IRC-2SLS &: \hat{L}_\lambda(K) = \hat{H}_{g\lambda}\hat{\sigma}_{u_\lambda v}\frac{2K}{\sqrt{N}} + \hat{\sigma}_{u_\lambda v}^2\frac{K^2}{N} + \hat{\sigma}_v^2(\hat{R}_\lambda(K) - \hat{\sigma}_{u_\lambda}^2\frac{K}{N}) + 2\hat{H}_{g\lambda}(\hat{F}_\lambda(K) - \hat{G}_\lambda(K)) \\
IRC-LIML &: \hat{L}_\lambda(K) = \hat{\sigma}_v^2(\hat{R}_\lambda(K) - \frac{\hat{\sigma}_{u_\lambda v}^2}{\hat{\sigma}_v^2}\frac{K}{N}) + 2\hat{H}_{g\lambda}(\hat{F}_\lambda(K) - \hat{G}_\lambda(K)) \\
IRC-B2SLS &: \hat{L}_\lambda(K) = \hat{\sigma}_v^2(\hat{R}_\lambda(K) + \frac{\hat{\sigma}_{u_\lambda v}^2}{\hat{\sigma}_v^2}\frac{K}{N}) + 2\hat{H}_{g\lambda}(\hat{F}_\lambda(K) - \hat{G}_\lambda(K)).
\end{aligned} \tag{2.6.1}$$

where $\hat{u}^K = (I - P^K)W$, $\hat{u}_\lambda^K = \hat{u}^K\hat{H}^{-1}\lambda$ denote residual vectors, $\hat{F}_\lambda(K) = \lambda'\hat{H}^{-1}(\hat{R}(K) - \hat{\sigma}_u^2\frac{K}{N})\hat{H}^{-1}\hat{H}_g$, and $\hat{G}_\lambda(K) = \lambda'\hat{H}^{-1}W'(I - P^K)\hat{\varepsilon}^*/\sqrt{N}$. For the $\hat{R}(K)$ and $\hat{R}_\lambda(K)$, I use the following Mallows' criterion in (2.6.1),

$$\hat{R}_\lambda(K) = \frac{\hat{u}_\lambda^{K\prime}\hat{u}_\lambda^K}{N} + 2\hat{\sigma}_{u_\lambda}^2\frac{K}{N} \qquad \hat{R}(K) = \frac{\hat{u}^{K\prime}\hat{u}^K}{N} + 2\hat{\sigma}_u^2\frac{K}{N}.$$

[4]Note that $\tilde{u}_\lambda$ are preliminary residuals of the regression $WH^{-1}\lambda = fH^{-1}\lambda + uH^{-1}\lambda \Leftrightarrow W_\lambda = f_\lambda + u_\lambda$, which is obtained by multiplying $H^{-1}\lambda$ with first-stage regression.

[5]Since higher-order MSE approximation for FULL estimator is same with those of LIML as in Proposition 2.2, one can choose K by using IRC-LIML for the FULL estimator, thus omitted here.

CV criterion can also be used

$$\hat{R}_\lambda(K) = \frac{1}{N}\sum_{i=1}^{N}\frac{(\hat{u}_{\lambda i}^K)^2}{(1-P_{ii}^K)^2} \qquad \hat{R}(K) = \frac{1}{N}\sum_{i=1}^{N}\frac{(\hat{u}_{i}^K)^2}{(1-P_{ii}^K)^2}.$$

Following Donald and Newey (2001) (see also Li (1987)), the goodness of fit criterion for the first-stage reduced form is used for the estimation of $f'(I-P^K)f/N$. For practical implementation, Mallows' criterion is simpler to use, however, the difference between Mallows' criterion and the CV criterion are not significant in the preliminary simulation results.

Next, I also consider Donald and Newey (2001)'s original criterion (DN, hereafter) which coincides to IRC criterion based on the MSE approximation under $N^{-\gamma}$ locally invalid instrument specification as in Corollary 2.1 ($\gamma \geq 1$), and Corollary 2.2-2.3 ($\gamma > 1/2$).

$$\begin{aligned}
&DN-2SLS: \hat{L}_\lambda(K) = \hat{\sigma}_{u_\lambda v}^2\frac{K^2}{N} + \hat{\sigma}_v^2(\hat{R}_\lambda(K) - \hat{\sigma}_{u_\lambda}^2\frac{K}{N}),\\
&DN-LIML: \hat{L}_\lambda(K) = \hat{\sigma}_v^2(\hat{R}_\lambda(K) - \frac{\hat{\sigma}_{u_\lambda v}^2}{\hat{\sigma}_v^2}\frac{K}{N}),\\
&DN-B2SLS: \hat{L}_\lambda(K) = \hat{\sigma}_v^2(\hat{R}_\lambda(K) + \frac{\hat{\sigma}_{u_\lambda v}^2}{\hat{\sigma}_v^2}\frac{K}{N}).
\end{aligned} \tag{2.6.2}$$

Now we define $\hat{K}$ as the instrument set which minimizes the criterion $\hat{L}_\lambda(K)$ over some user-specified consideration set $\mathcal{K}$,

$$\hat{K} = \arg\min_{K\in\mathcal{K}} \hat{L}_\lambda(K).$$

This requires calculating $\hat{L}_\lambda(K)$ over different set of instruments $K \in \mathcal{K}$. The IRC selected IV estimator of δ can be defined as $\hat{\delta}(\hat{K})$ for each IV estimator.

In the important special case when there is only one endogenous variable, i.e., Y_i is a

scalar, approximate MSE for 2SLS with Proposition 2.1 can be reduced to

$$L_\lambda(K) = (\lambda' H^{-1} e_1)^2 (\sigma_{\xi v}^2 \frac{K^2}{N} + \sigma_v^2 \frac{\bar{Y}'(I - P^K)\bar{Y}}{N}) + (\lambda' H^{-1} H_g)(\lambda' H^{-1} e_1)\big(\sigma_{\xi v} \frac{2K}{\sqrt{N}} + 2(e_1' H^{-1} H_g) \frac{\bar{Y}'(I - P^K)\bar{Y}}{N} - 2\frac{\bar{Y}'(I - P^K)g}{N}\big) \tag{2.6.3}$$

where $\bar{Y} = (\mathbb{E}[Y_1|x_1], \cdots, \mathbb{E}[Y_N|x_N])'$, $\sigma_{uv} = \sigma_{\xi v} e_1$, $\sigma_{\xi v} = \mathbb{E}(\xi_i v_i)$ and $e_1 = (1, 0, \cdots, 0)'$ is the $p \times 1$ first unit vector. The choice of λ such that $H^{-1}\lambda = e_1$ makes further simplifications as follows, which can be used for implementation,

$$2SLS : L(K) = \sigma_{\xi v}^2 \frac{K^2}{N} + (\sigma_v^2 + 2(e_1' H^{-1} H_g) H_{g1}) \frac{\bar{Y}'(I - P^K)\bar{Y}}{N} + 2H_{g1}(\sigma_{\xi v} \frac{K}{\sqrt{N}} - \frac{\bar{Y}'(I - P^K)g}{N})$$

where $H_{g1} = \bar{Y}'g/N$. We can get similar results for LIML and B2SLS in Proposition 2.2 and 2.3,

$$LIML : L(K) = (\sigma_\xi^2 \sigma_v^2 - \sigma_{\xi v}^2) \frac{K}{N} + (\sigma_v^2 + 2(e_1' H^{-1} H_g) H_{g1}) \frac{\bar{Y}'(I - P^K)\bar{Y}}{N} - 2H_{g1} \frac{\bar{Y}'(I - P^K)g}{N},$$
$$B2SLS : L(K) = (\sigma_\xi^2 \sigma_v^2 + \sigma_{\xi v}^2) \frac{K}{N} + (\sigma_v^2 + 2(e_1' H^{-1} H_g) H_{g1}) \frac{\bar{Y}'(I - P^K)\bar{Y}}{N} - 2H_{g1} \frac{\bar{Y}'(I - P^K)g}{N}$$

with $\sigma_\xi^2 = \mathbb{E}(\xi_i^2)$. In this case, the IRC criterion are also simplified with preliminary estimates $\hat{H}, \hat{H}_g, \hat{\sigma}_v^2, \hat{\sigma}_u^2, \hat{\sigma}_{uv}$ and estimate of each term in $L(K)$ using Mallows or Cross-validation that are defined same as in general vector cases. With a conservative set of valid instruments z_i, we have an asymptotically unbiased estimator for the terms involving $\bar{Y}'(I - P^K)g/N$. As pointed out in Donald and Newey, we can easily see that DN criterion is a multiplication of $(\lambda' H^{-1} e_1)^2$ from (2.6.3) in the scalar endogenous case, so that choice of K does not depend on λ.[6]

[6]In a simple case with scalar endogenous regressor and no additional exogenous regressors (i.e., W_i is scalar), then K that minimizes IRC criterion $L_\lambda(K)$ does not depend on λ as the $L_\lambda(K)$ is the scaled by $(\lambda H^{-1})^2$. We can think of this case as the covariates have already been partialled out. Specifically, from the original data, $(\tilde{y}, \tilde{Y}, \tilde{X})$, $y = M_{X_1}\tilde{y}, Y = M_{X_1}\tilde{Y}, X = M_{X_1}\tilde{X}$ where $M_{X_1} = I - X_1(X_1'X_1)^- X_1'$ is the orthogonal projection matrix of exogenous covariates x_{1i}.

To provide optimality properties of DN criterion under locally invalid instruments, I impose following assumption similar to those of Donald and Newey (2001) including consistency of the preliminary estimators.

Assumption 2.5. *Y_i is scalar, $\hat{\sigma}_v^2 - \sigma_v^2 = o_p(1), \hat{\sigma}_{u_\lambda v} - \sigma_{u_\lambda v} = o_p(1), \hat{\sigma}_{u_\lambda}^2 - \sigma_{u_\lambda}^2 = o_p(1)$, $\hat{\sigma}_u^2 - \sigma_u^2 = o_p(1), \hat{\sigma}_{uv} - \sigma_{uv} = o_p(1)$, $\hat{H} - H = o_p(1)$, $\lambda' \bar{H}^{-1} \sigma_{uv} \neq 0$, and $var(\lambda' H^{-1} \eta_i) > 0$. Also assume, $\sup_K \sup_i P_{ii}^K \xrightarrow{p} 0$, $\mathbb{E}(u_i^8 | x_i) < \infty$, $\inf_K NR(K) \to \infty$ where $R(K) = \sigma_{u_\lambda}^2 (K/N) + \lambda' H^{-1} [f'(I - P^K) f / N] H^{-1} \lambda$.*

Corollary 2.5. *Suppose that the same assumptions as in Corollary 2.1 hold, and that, in addition, Assumption 2.5 hold. For 2SLS estimator with $\hat{K} = \arg\min_{K \in \mathcal{K}} \hat{L}_\lambda(K)$, where $\hat{L}_\lambda(K)$ is defined in* (2.6.2)*, following holds for all $\gamma \geq 1$,*

$$\frac{L_\lambda(\hat{K})}{\inf_K L_\lambda(K)} \xrightarrow{p} 1. \tag{2.6.4}$$

For LIML (and FULL) or B2SLS estimator, with $\hat{L}_\lambda(K)$ is defined in (2.6.2)*,* (2.6.4) *holds for all $\gamma > 1/2$ under the same assumptions as in Corollary 2.2 or Corollary 2.3.*

2.7 Conclusions

In this paper, we develop an instrument selection criteria that are robust to the potential invalidity of instrument in many instruments setup. We derive higher-order MSE approximations of the IV estimator allowing locally invalid instruments. Based on these higher-order approximations, we propose an instrument selection criteria. By considering the bias-variance trade-off from using many instruments and using invalid instruments at the same time, our robust instrument selection criteria can be useful in practice when researchers have potentially large sets of instruments without assuming perfectly valid instruments.

This paper has some limitations. First, here, we only focused on the linear IV model, which we believe will be the most useful to many empirical researchers. We believe that our

locally invalid instrument specification can be generalized to GMM settings by combining the ideas in this paper with existing work of Donald, Imbens, and Newey (2009). Assessing the cost of not imposing nonlinear model and heteroskedasticity assumption will be interesting future research. Second, though our model considers some weak instruments, one of our assumptions explicitly rules out the case where all instruments are weak. Third, the assumption of having small set of valid instruments may be restrictive in some applications, although it would be less restrictive than assuming all instruments' validity *a priori*. It would be desirable to extend the analysis without these conservative sets. Finally, our paper does not discuss post-model selection inference issues. This is another possible direction for future research.

2.8 References

Andrews, D. W. K. (1999): "Consistent Moment Selection Procedures for Generalized Method of Moments Estimation," *Econometrica*, 67(3), 543-563.

Andrews, D. W. K. and B. Lu (2001): "Consistent Model and Moment Selection Procedures for GMM Estimation with Application to Dynamic Panel Data Models," *Journal of Econometrics*, 101(1), 123-164.

Angrist, J. and A. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?," *Quarterly Journal of Economics*, 106(4), 979-1014.

Bekker, P. A. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica*, 62(3), 657-681.

Belloni, A., D. Chen, V. Chernozhukov and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, 80(6), 2369-2430.

Berkowitz, D., M. Caner and Y. Fang (2008): "Are Nearly Exogenous Instruments Reliable?," *Economics Letters*, 101(1), 20-23.

Berkowitz, D., M. Caner and Y. Fang (2012): "The Validity of Instruments Revisited," *Journal of Econometrics*, 166(2), 255-266.

Bound, J., D. Jaeger, and R. Baker (1995): "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variables is Weak," *Journal of American Statistical Association*, 90, 443-450.

Cameron, S. V. and C. Taber (2004): "Estimation of Educational Borrowing Constraints Using Returns to Schooling," *Journal of Political Economy*, 112(1), 132-182.

Canay, I. (2010):"Simultaneous Selection and Weighting of Moments in GMM Using a Trapezoidal Kernel," *Journal of Econometrics*, 156(2), 284-303.

Caner, M. (2014): "Near Exogeneity and Weak Identification in Generalized Empirical Likelihood Estimators: Many Moment Asymptotics," *Journal of Econometrics*, 18(2), 247-268.

Caner, M., X. Han and Y. Lee (2013): "Adaptive Elastic Net GMM Estimator with Many Invalid Moment Conditions: A Simultaneous Model and Moment Selection," *Working paper*.

Card, D. (1995): "Using geographic variation in college proximity to estimate the return to schooling," *Aspects of Labor Market Behavior: Essays in Honour of John Vanderkamp*, L.N. Christofides, E.K. Grant, and R. Swidinsky, eds., Toronto: University of Toronto Press, 201-222.

Card, D. (1999): "The Causal Effect of Education on Earnings", *Handbook of Labor Economics*, vol. 3A, O. Ashenfelter, and D. Card, eds., Amsterdam: Elsevier Science/North-Holland, 1801-1863.

Carrasco, M. (2012): "A Regularization Approach to the Many Instruments Problem," *Journal of Econometrics*, 170(2), 383-398.

Chamberlain, G., and G. W. Imbens (2004): "Random Effects Estimators with Many Instrumental Variables," *Econometrica*, 72(1), 295-306.

Chao, J. and N. Swanson (2005): "Consistent Estimation With a Large Number of Weak Instruments," *Econometrica*, 73(5), 1673-1692.

Cheng, X. and Z. Liao (2014): "Select the Valid and Relevant Moments: An Information-Based LASSO for GMM with Many Moments," *Journal of Econometrics*, forthcoming.

Conley T. G., C. B. Hansen, and P. E. Rossi (2012): "Plausibly Exogenous," *Review of Economics and Statistics*, 94(1), 260-272.

DiTraglia, F. (2014): "Using Invalid Instruments on Purpose: Focused Moment Selection and Averaging for GMM," *Working paper*, University of Pennsylvania

Donald, S. G., G. W. Imbens, and W. K. Newey (2009): "Choosing Instrumental Variables in Conditional Moment Restriction Models," *Journal of Econometrics*, 152(1), 28-36.

Donald, S. G. and W. K. Newey (2001): "Choosing the Number of Instruments," *Econometrica*, 69(5), 1161-1191.

Gautier, E., and A. B. Tsybakov (2014): "High-dimensional Instrumental Variables Regression and Confidence Sets," *Preprint*, arXiv: 1105.2454v2.

Griliches Z. (1976): "Wages of Very Young Men," *Journal of Political Economy*, 84, 69-86.

Guggenberger, P. (2012): "On the Asymptotic Size Distortion of Tests When Instruments Locally Violate the Exogeneity Assumption," *Econometric Theory*, 28(2), 387-421.

Hahn, J. and J. Hausman (2002): "A New Specification Test for the Validity of Instrumental Variables," *Econometrica*, 70(1), 163-189.

Hahn, J. and J. Hausman (2005): "Estimation with Valid and Invalid Instruments," *Annales d'Économie et de Statistique*, 25-57.

Hall, A. R. and A. Inoue (2003): "The Large Sample Behaviour of the Generalized Method of Moments Estimator in Misspecified Models," *Journal of Econometrics*, 114(2), 361-394.

Hall, A. R., A. Inoue, K. Jana, and C. Shin (2007): "Information in Generalized Method of Moments Estimation and Entropy-Based Moment Selection," *Journal of Econometrics*, 138(2), 488-512.

Hall, A. R. and F. P. M. Peixe (2003): "A Consistent Method for the Selection of Relevant Instruments," *Econometric Reviews*, 22(3), 269-287.

Hansen, B. E. (2013): "The Risk of James-Stein and Lasso Shrinkage," *Econometric Reviews*, forthcoming.

Hansen, C. and J. Hausman, W. K. Newey (2008): "Estimation with Many Instrumental Variables?, *Journal of Business and Economic Statistics*, 26(4), 398-422.

Hong, H., B. Preston, and M. Shum (2003): "Generalized Empirical Likelihood-based Model Selection Criteria for Moment Condition Models," *Econometric Theory*, 19(6), 923-943.

Inoue, A. (2006): "A Bootstrap Approach to Moment Selection," *The Econometrics Journal*, 9, 48-75.

Kabaila, P. (1995): "The Effect of Model Selection on Confidence Regions and Prediction Regions," *Econometric Theory*, 11, 537-549.

Kolesár, M., R. Chetty, J. Friedman, E. Glaeser, and G. W. Imbens (2014): "Identification and Inference with Many Invalid Instruments," *Journal of Business and Economic Statistics*, forthcoming.

Kraay, A. (2012): "Instrumental Variables Regressions with Uncertain Exclusion Restrictions: A Bayesian Approach," *Journal of Applied Econometrics*, 27(1), 108-128.

Kuersteiner, G. (2012): "Kernel Weighted GMM Estimators for Linear Time Series Models," *Journal of Econometrics*, 170(2), 399-421.

Kuersteiner, G. and R. Okui (2010): "Constructing Optimal Instruments by First-Stage Prediction Averaging," *Econometrica*, 78(2), 697-718.

Lee, Y. and R. Okui (2012): "Hahn-Hausman test as a specification test," *Journal of Econometrics*, 167(1), 133-139.

Lee, Y. and Y. Zhou (2014): "Averaged Instrumental Variables Estimators," *Working Paper*.

Leeb, H. and B. M. Pötscher (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21(1), 21-59.

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

Liao, Z. (2013): "Adaptive GMM Shrinkage Estimation with Consistent Moment Selection," *Econometric Theory*, 29(5), 857-904.

Mallows, C. L. (1973): "Some comments on C_p," *Technometrics*, 15, 661-75.

McCloskey, A. (2014): "Bonferroni-Based Size-Correction for Nonstandard Testing Problems," *Working Paper*, Brown University.

Morimune, K. (1983): "Approximate Distributions of k-Class Estimators when the Degree of Overidentifiability is Large Compared with the Sample Size," *Econometrica*, 51, 821-841.

Nagar, A. L. (1959): "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations," *Econometrica*, 27, 575-595.

Nevo, A. and A. Rosen (2012): "Identification with Imperfect Instruments," *Review of Economics and Statistics*, 94(3), 659-671.

Newey, W. K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79(1), 147-168.

Newey, W. K. and R.J. Smith (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72(1), 219-255.

Okui, R. (2011): "Instrumental Variable Estimation in the Presence of Many Moment Conditions," *Journal of Econometrics*, 165(1), 70-86.

Staiger, D. and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65(3), 557-586.

2.9 Proofs

This section contains the proofs of the lemma, proposition, and corollary. We closely follows the steps of MSE derivations to those of Donald and Newey (2001) allowing possibly invalid instruments. This requires some modifications of lemmas.

In the following proofs of proposition and corollary, each IV estimator has a representation $\sqrt{N}(\hat{\delta}(K)-\delta_0) = \hat{H}^{-1}\hat{h} + \hat{H}^{-1}\hat{h}_g$. Define $h = f'v/\sqrt{N}, H = f'f/N$, and $H_g = f'g/N$. Also, define $\rho_{K,N} = tr(L(K))$. Throughout the section, we will denote $\sum_i = \sum_{i=1}^N, \sum_{i\neq j} = \sum_{i=1}^N \sum_{j\neq i}$. LLN denotes (weak) law of large numbers, and CLT denotes Lindberg-Levy central limit theorem.

Lemma 2.1. *If there is a decomposition* $\hat{h} = h + T^h + Z^h, \hat{H} = H + T^H + Z^H, \hat{h}_g = H_g + T^g + Z^g$, *and*

$$
\begin{aligned}
&(h+T^h)(h+T^h)' - hh'H^{-1}T^{H'} - T^H H^{-1}hh' = \hat{A}_1(K) + Z^{A_1}(K),\\
&(H_g+T^g)(H_g+T^g)' - H_gH_g'H^{-1}T^{H'} - T^H H^{-1}H_gH_g' = \hat{A}_2(K) + Z^{A_2}(K),\\
&(h+T^h)(H_g+T^g)' - hH_g'H^{-1}T^{H'} - T^H H^{-1}hH_g' = \hat{A}_3(K) + Z^{A_3}(K),
\end{aligned}
$$

such that $T^h = o_p(1), h = O_p(1), H_g = O_p(1), T^g = o_p(1)$, *and* $H = O_p(1)$, *the determinant of* H *is bounded away from zero with probability 1,* $\rho_{K,N} = tr(L(K))$, *and* $\rho_{K,N} = o_p(1)$,

$$
\begin{aligned}
&\|T^H\|^2 = o_p(\rho_{K,N}), \|T^h\|\|T^H\| = o_p(\rho_{K,N}), \|Z^h\| = o_p(\rho_{K,N}), \|Z^H\| = o_p(\rho_{K,N}),\\
&\|Z^g\| = o_p(\rho_{K,N}), \|T^g\|\|T^H\| = o_p(\rho_{K,N}), \quad Z^{A_i}(K) = o_p(\rho_{K,N}) \quad \textit{for all } i = 1,2,3,\\
&\mathbb{E}(\hat{A}_1(K) + \hat{A}_2(K) + \hat{A}_3(K) + \hat{A}_3(K)'|X) = H\Phi H + HL(K)H + o_p(\rho_{K,N}),
\end{aligned}
$$

then

$$N(\hat{\delta}(K)-\delta_0)(\hat{\delta}(K)-\delta_0)' = \hat{Q}(K)+\hat{r}(K),$$
$$\mathbb{E}(\hat{Q}(K)|X) = \Phi + L(K) + T(K),$$
$$[\hat{r}(K)+T(K)]/tr(L(K)) = o_p(1), \quad as \quad K\to\infty, N\to\infty.$$

Proof of Lemma 2.1 The proof closely follows the steps in Donald and Newey (2001) with modifications due to the invalid instruments specification. First, we observe that

$$\hat{H}^{-1}\hat{h}+\hat{H}^{-1}\hat{h}_g = H^{-1}(\hat{h}-(\hat{H}-H)H^{-1}h)+\hat{Z}+H^{-1}(\hat{h}_g-(\hat{H}-H)H^{-1}H_g)+\hat{Z}_g,$$
$$\hat{Z} = H^{-1}(H-\hat{H})\hat{H}^{-1}(H-\hat{H})H^{-1}h+\hat{H}^{-1}(H-\hat{H})H^{-1}(\hat{h}-h),$$
$$\hat{Z}_g = H^{-1}(H-\hat{H})\hat{H}^{-1}(H-\hat{H})H^{-1}H_g+\hat{H}^{-1}(H-\hat{H})H^{-1}(\hat{h}_g-H_g).$$

Also note that, $\hat{H}-H = T^H+Z^H, \|T^H\|^2 = o_p(\rho_{K,N}), \|Z^H\|^2 = o_p(\rho_{K,N}), \|T^g\|\|T^H\| = o_p(\rho_{K,N}), \|Z^g\| = o_p(\rho_{K,N})$, and $\hat{h}_g = H_g+T^g+Z^g = O_p(1)$. Thus, $\|\hat{H}-H\|^2 \le 2(||T^H||^2 + ||Z^H||^2) = o_p(\rho_{K,N})$, and $\|\hat{h}_g-H_g\|\|H-\hat{H}\| \le \|T^g\|\|T^H\| + \|Z^g\|\|T^H\| + \|T^g\|\|Z^H\| + \|Z^g\|\|Z^H\| = o_p(\rho_{K,N})$. Also, H is nonsingular wpa 1 by assumption in the lemma, so that $H^{-1} = O_p(1)$. Moreover, $\hat{H} = H+o_p(1)$ and $\hat{H}^{-1} = H^{-1}+o_p(1) = O_p(1)$. Thus,

$$\|\hat{Z}_g\| \le \|H^{-1}\|\|H-\hat{H}\|^2\|\hat{H}^{-1}\|\|H^{-1}H_g\| + \|\hat{h}_g-H_g\|\|H-\hat{H}\|\|H^{-1}\|\|\hat{H}^{-1}\| = o_p(\rho_{K,N}).$$

Similarly, we can show that $\|\hat{Z}\| = o_p(\rho_{K,N})$. Next, define $\tilde{\tau}_g = H_g+T^g-T^HH^{-1}H_g$. Then, we obtain

$$\hat{H}^{-1}\hat{h}_g = H^{-1}\tilde{\tau}_g + o_p(\rho_{K,N})$$

by using $\|\hat{Z}_g\| = o_p(\rho_{K,N}), \|Z^g\| = o_p(\rho_{K,N}), \|H^{-1}Z^HH^{-1}H_g\| = o_p(\rho_{K,N})$. Similarly, for $\hat{h} = h+T^h+o_p(\rho_{K,N}) = O_p(1)$, we obtain $\hat{H}^{-1}\hat{h} = H^{-1}\tilde{\tau}+o_p(\rho_{K,N})$ with $\tilde{\tau} = h+T^h-T^HH^{-1}h$

by using $\|\hat{Z}\| = o_p(\rho_{K,N})$, $\|Z^h\| = o_p(\rho_{K,N})$, and $\|Z^H\| = o_p(\rho_{K,N})$.

Then,

$$\begin{aligned}\tilde{\tau}\tilde{\tau}' &= \hat{A}_1(K) + Z^{A_1}(K) - T^h h' H^{-1} T^{H'} - T^H H^{-1} h T^{h'} + T^H H^{-1} h h' H^{-1} T^{H'} \\ &= \hat{A}_1(K) + o_p(\rho_{K,N})\end{aligned}$$

by $\|T^h\|\|T^H\| = o_p(\rho_{K,N})$, $\|T^H\|^2 = o_p(\rho_{K,N})$, and $\|Z^{A_1}(K)\| = o_p(\rho_{K,N})$. Also,

$$\begin{aligned}\tilde{\tau}_g\tilde{\tau}_g' &= \hat{A}_2(K) + Z^{A_2}(K) - T^g H_g' H^{-1} T^{H'} - T^H H^{-1} H_g T^{g'} + T^H H^{-1} H_g H_g' H^{-1} T^{H'} \\ &= \hat{A}_2(K) + o_p(\rho_{K,N})\end{aligned}$$

by using $\|T^g\|\|T^H\| = o_p(\rho_{K,N})$, $\|T^H\|^2 = o_p(\rho_{K,N})$, and $\|Z^{A_2}(K)\| = o_p(\rho_{K,N})$.

For the cross term, we obtain

$$\begin{aligned}\tilde{\tau}\tilde{\tau}_g' &= \hat{A}_3(K) + Z^{A_3}(K) - T^h H_g' H^{-1} T^{H'} - T^H H^{-1} h T^{g'} + T^H H^{-1} h H_g' H^{-1} T^{H'} \\ &= \hat{A}_3(K) + o_p(\rho_{K,N})\end{aligned}$$

by $\|T^h\|\|T^H\| = o_p(\rho_{K,N})$, $\|T^g\|\|T^H\| = o_p(\rho_{K,N})$, $\|T^H\|^2 = o_p(\rho_{K,N})$ and $\|Z^{A_3}(K)\| = o_p(\rho_{K,N})$.

Since $\sqrt{N}(\hat{\delta}(K) - \delta_0) = H^{-1}\tilde{\tau} + H^{-1}\tilde{\tau}_g + o_p(\rho_{K,N})$, it follows that

$$N(\hat{\delta}(K) - \delta_0)(\hat{\delta}(K) - \delta_0)' = H^{-1}(\hat{A}_1(K) + \hat{A}_2(K) + \hat{A}_3(K) + \hat{A}_3(K)')H^{-1} + o_p(\rho_{K,N}).$$

Then, desired conclusion directly follows from the assumption in the Lemma. *Q.E.D.*

Next we provide useful lemmas for the proof of Proposition 2.1-2.3. We do not provide proofs of Lemma 2.2, as they are available in Donald and Newey (2001). Define $e_f(K) = f'(I - P^K)f/N$, $\Delta(K) = tr(e_f(K))$, $e_g(K) = g'(I - P^K)g/N$, and $\Delta_g(K) = tr(e_g(K))$.

Lemma 2.2. *(Donald and Newey (2001) Lemma A.2, A.3) If Assumptions 2.1-2.3 are satisfied, then we have*

(i) $tr(P^K) = K$, (ii) $\sum_i (P^K_{ii})^2 = o_p(K)$, (iii) $\sum_{i\neq j} P^K_{ii} P^K_{jj} = K^2 + o_p(K)$, (iv) $\sum_{i\neq j} P^K_{ij} P^K_{ij} = K + o_p(K)$,

(v) $h = f'v/\sqrt{N} = O_p(1), H = f'f/N = O_p(1)$,

(vi) $\Delta(K) = o_p(1)$,

(vii) $f'(I - P^K)v/\sqrt{N} = O_p(\Delta(K)^{1/2})$,

(viii) $u'P^K v = O_p(K)$,

(ix) $\mathbb{E}(u'P^K vv'P^K u|X) = \sigma_{uv}\sigma'_{uv}K^2 + (\sigma_v^2\Sigma_u + \sigma_{uv}\sigma'_{uv})K + o_p(K) = \sigma_{uv}\sigma'_{uv}K^2 + o_p(K^2)$,

(x) $\mathbb{E}(f'vv'P^K u) = \sum_i f_i P^K_{ii}\, \mathbb{E}(v_i^2 u_i'|x_i) = O_p(K)$,

(xi) $\Delta(K)^{1/2}/\sqrt{N} = o_p(K/N + \Delta(K))$,

(xii) $\mathbb{E}(hh'H^{-1}u'f/N|X) = \sum_i f_i f_i' H^{-1}\, \mathbb{E}(v_i^2 u_i|x_i) f_i'/N^2 = O_p(1/N)$,

(xiii) $\mathbb{E}(f'(I - P^K)vv'P^K u/N|X) = o_p(\Delta(K)^{1/2}\sqrt{K}/\sqrt{N})$.

Next, lemma gives useful calculations that will appear in the MSE approximation due to the invalid instruments.

Lemma 2.3. *If Assumptions 2.1-2.3 are satisfied, then we have*

(i) $H_g = f'g/N = O_p(1)$,

(ii) $\Delta_g(K) = o_p(1)$,

(iii) $g'(I - P^K)v/\sqrt{N} = O_p(\Delta_g(K)^{1/2})$, $f'(I - P^K)g/N = O_p((\Delta(K)\Delta_g(K))^{1/2})$,

(iv) $\mathbb{E}(f'vg'u/N|X) = H_g\sigma'_{uv}$,

(v) $\mathbb{E}(u'P^K v|X) = K\sigma_{uv}$,

(vi) $\mathbb{E}((u'f + f'u)/NH^{-1}f'v\sqrt{N}|X) = (\sum_i \sigma_{uv} f_i' H^{-1} f_i + \sum_i f_i \sigma'_{uv} H^{-1} f_i)/N^{3/2} = O_p(1/\sqrt{N})$,

(vii) $\mathbb{E}(f'v(f'g)'H^{-1}(u'f + f'u)|X) = \Sigma_i f_i (f'g)' H^{-1}\sigma_{uv} f_i' + \Sigma_i f_i (f'g)' H^{-1} f_i \sigma'_{uv}$.

Proof of Lemma 2.3 (i) holds by LLN. Observer that $(I - P^K)$ is idempotent, and

$$\mathbb{E}(\Delta_g(K)) \leq \mathbb{E}[tr(g - \Psi\pi_K^{g'})'(g - \Psi\pi_K^{g'})]/N = \mathbb{E}(|g(x) - \pi_K^g \psi^K(x)|^2) \to 0,$$

by Assumption 2.2(ii). Thus, $\Delta_g(K) = o_p(1)$ by Markov inequality. Next, observe that $\mathbb{E}(g'(I - P^K)v/\sqrt{N}|X) = 0$, and

$$\mathbb{E}(g'(I - P^K)vv'(I - P^K)g/N|X) = \sigma_v^2 e_g(K).$$

Therefore, $g'(I - P^K)v/\sqrt{N} = O_p(\Delta_g(K)^{1/2})$ by Chebyshev inequality. Moreover,

$$\|f'(I - P^K)g/N\| \leq \sqrt{f'(I - P^K)f/N}\sqrt{g'(I - P^K)g/N} = O_p(\Delta(K)^{1/2}\Delta_g(K)^{1/2}),$$

by Cauchy-Schwarz inequality, $(I - P^K)$ is idempotent.

Also, $\mathbb{E}(f'vg'u/N|X) = \sum_{i,j}\mathbb{E}(f_i v_i g_j u_j'|X)/N = \sum_i f_i g_i \mathbb{E}(v_i u_i'|x_i)/N = H_g\sigma_{uv}'$, and this gives (iv). Moreover, $\mathbb{E}(u'P^K v|X) = \sum_i \mathbb{E}(u_i P_{ii}^K v_i|X) + \sum_{i,j}\mathbb{E}(u_i P_{ij}^K v_j|X) = \sum_i P_{ii}^K \mathbb{E}(u_i v_i|X) = K\sigma_{uv}$, so that (v) holds. Next, $\mathbb{E}(u'f/NH^{-1}f'v/\sqrt{N}|X) = \sum_i \mathbb{E}(u_i f_i' H^{-1} f_i v_i|x_i)/N^{3/2} = \sigma_{uv}(\sum_i f_i' H^{-1} f_i/N^{3/2}) = O_p(1/\sqrt{N})$, and similarly, $\mathbb{E}(f'u/NH^{-1}f'v/\sqrt{N}|X) = (\sum_i f_i\sigma_{uv}' H^{-1} f_i)/N^{3/2} = O_p(1/\sqrt{N})$. This gives (vi). Similarly, $\mathbb{E}(f'v(f'g)'H^{-1}u'f|X) = \Sigma_i \mathbb{E}(f_i v_i (f'g)'H^{-1}u_i f_i'|X) = \Sigma_i f_i (f'g)'H^{-1}\sigma_{uv}f_i'$, and $\mathbb{E}(f'v(f'g)'H^{-1}f'u|X) = \Sigma_i f_i(f'g)'H^{-1}f_i\sigma_{uv}'$ gives (vii). *Q.E.D.*

Proof of Proposition 2.1

The 2SLS estimator, $\hat{\delta}(K) = (W'P^K W)^{-1}(W'P^K y)$ has the following decomposition with locally invalid instruments specification

$$\sqrt{N}(\hat{\delta}(K) - \delta_0) = \hat{H}^{-1}\hat{h} + \hat{H}^{-1}\hat{h}_g,$$

where,

$$\hat{H} = \frac{W'P^K W}{N}, \hat{h} = \frac{W'P^K v}{\sqrt{N}}, \hat{h}_g = \frac{W'P^K g}{N}.$$

Also, $\hat{h}, \hat{H}$ and $\hat{h}_g$ can be decomposed as

$$\hat{h} = h + T_1^h + T_2^h,$$
$$T_1^h = -f'(I - P^K)v/\sqrt{N} = O_p(\Delta(K)^{1/2}), \quad T_2^h = u'P^K v/\sqrt{N} = O_P(K/\sqrt{N}),$$
$$\hat{H} = H + T_1^H + T_2^H + Z^H,$$
$$T_1^H = -f'(I - P^K)f/N = -e_f(K) = O_p(\Delta(K)), \quad T_2^H = (u'f + f'u)/N = O_p(1/\sqrt{N}),$$
$$Z^H = (u'P^K u - u'(I - P^K)f - f'(I - P^K)u)/N = O_p(K/N + \Delta(K)^{1/2}/\sqrt{N}),$$
$$\hat{h}_g = H_g + T_1^g + Z^g,$$
$$T_1^g = -f'(I - P^K)g/N = O_p(\Delta(K)^{1/2}\Delta_g(K)^{1/2}),$$
$$Z^g = u'g/N - u'(I - P^K)g/N = O_p(1/\sqrt{N} + \Delta_g(K)^{1/2}/\sqrt{N}).$$

We show that the conditions of Lemma 2.1 are satisfied, and $L(K)$ has the representations given in the proposition. Note that $L(K)$ contains the terms of order $K/\sqrt{N}$ and K^2/N. Thus to show the term is $o_p(\rho_{K,N})$, it is enough to show $o_p(K/\sqrt{N} + K^2/N + \Delta(K))$.

Note that $h = O_p(1), H = O_p(1)$ by Lemma 2.2 (v). Also, $T^h = -f'(I - P)v/\sqrt{N} + u'Pv/\sqrt{N} = O_p(\Delta(K)^{1/2}) + O_p(K/\sqrt{N}) = o_p(1)$ by Lemma 2.2(vii), (viii), and using $\Delta(K) = o_p(1), K/\sqrt{N} = o(1)$. Moreover, $T_1^H = -f'(I - P)f/N = O_p(\Delta_K)$ by the definition of Δ_K, and $T_2^H = (u'f + f'u)/N = O_p(1/\sqrt{N})$ by the CLT. Thus $\|T^H\|^2 \leq \|T_1^H\|^2 + \|T_2^H\|^2 + 2\|T_1^H\|\|T_2^H\| = O_p(\Delta(K)^2) + O_p(1/N) + O_p(\Delta(K)/\sqrt{N}) = o_p(\rho_{K,N})$. Also,

$$\|T^h\|\|T^H\| = O_p(\Delta(K)^{3/2}) + O_p(\Delta(K)^{1/2}/\sqrt{N}) + O_p(\Delta(K)K/\sqrt{N}) + O_p(K/N) = o_p(\rho_{K,N}),$$

since $\Delta(K)^{1/2}/\sqrt{N} = o_p(\rho_{K,N})$ by Lemma 2.2 (xi). In addition, $Z^h = 0$ in this case, thus $\|Z^h\| = o_p(\rho_{K,N})$. Next, $Z^H = (u'Pu - u'(I - P)f - f'(I - P)u)/N = O_p(K/N) + O_p(\Delta(K)^{1/2}/\sqrt{N}) = o_p(\rho_{K,N})$ by Lemma 2.2 (vii), (viii), (xi).

Next, $H_g = O_p(1)$ by Lemma 2.3 (i), and $T_1^g = O_p(\Delta(K)^{1/2}\Delta_g(K)^{1/2}) = o_p(1)$ by Lemma 2.2 (vi), 2.3 (ii), (iii). Morover, $u'g/N = O_p(1/\sqrt{N}) = o_p(\rho_{K,N})$ by CLT and $1/\sqrt{N} =$

$o(K/\sqrt{N})$. Also, $u'(I - P^K)g/N = O_p(\Delta_g(K)^{1/2}/\sqrt{N}) = o_p(\rho_{K,N})$ by Lemma 2.3 (iii) (replacing v with u) and this gives $\|Z^g\| = o_p(\rho_{K,N})$.

Also,

$$\|T^g\|\|T^H\| = O_p(\Delta(K)^{3/2}\Delta_g(K)^{1/2}) + O_p(\Delta(K)^{1/2}\Delta_g(K)^{1/2}/\sqrt{N}) = o_p(\rho_{K,N}),$$

by $\Delta(K)^{3/2} = o_p(\rho_{K,N})$, $\Delta(K)^{1/2}/\sqrt{N} = o_p(\rho_{K,N})$ using Lemma 2.2 (xi).

Next, we calculate the expectation of each term $\hat{A}_1(K), \hat{A}_2(K)$, and $\hat{A}_3(K)$ defined in Lemma 2.1. For $Z^{A_1}(K) = 0, \hat{A}_1(K) = (h + T_1^h + T_2^h)(h + T_1^h + T_2^h)' - hh'H^{-1}(T_1^H + T_2^H)' - (T_1^H + T_2^H)H^{-1}hh'$, by the proof of the Proposition 1 in Donald and Newey (2001), $\mathbb{E}(\hat{A}_1(K)|X) = \sigma_v^2 H + \sigma_v^2 e_f(K) + \sigma_{uv}\sigma_{uv}' K^2/N + o_p(\rho_{K,N})$.

Next, for $Z^{A_2}(K) = 0$, we analyze expectation of $\hat{A}_2(K) = (H_g + T_1^g)(H_g + T_1^g)' - H_g H_g' H^{-1}(T_1^H + T_2^H)' - (T_1^H + T_2^H)H^{-1}H_g H_g'$. First of all, $\mathbb{E}(H_g T_1^{g\prime}|X) = -H_g g'(I - P^K)f/N$ and $\mathbb{E}(T_1^g H_g'|X) = -f'(I - P^K)g/N H_g'$. Second, $\mathbb{E}(T_1^g T_1^{g\prime}|X) = O_p(\Delta_K \Delta_{g,K}) = o_p(\rho_{K,N})$ by Lemma 2.3 (ii), (iii). Next,

$$\mathbb{E}(H_g H_g' H^{-1} T_1^{H\prime}|X) = -H_g H_g' H^{-1} e_f(K).$$

Lastly,

$$\mathbb{E}(H_g H_g' H^{-1} T_2^{H\prime}|X) = H_g H_g' H^{-1}\,\mathbb{E}(\frac{u'f + f'u}{N}|X) = 0.$$

Thus,

$$\begin{aligned}\mathbb{E}(\hat{A}_2(K)|X) = {} & H_g H_g' + H_g H_g' H^{-1} e_f(K) + e_f(K) H^{-1} H_g H_g' - H_g g'(I - P^K)f/N \\ & - f'(I - P^K)g/N H_g' + o_p(\rho_{K,N}).\end{aligned}$$

For $Z^{A_3}(K) = (\sum_{j=1}^2 T_j^h)T_1^{g\prime}$, we investigate expectation of $\hat{A}_3(K) = h(H_g + T_1^g)' +$

$(T_1^h + T_2^h)H_g' - hH_g'H^{-1}(T_1^H + T_2^H)' - (T_1^H + T_2^H)H^{-1}hH_g'$. First, observe that $\mathbb{E}(hH_g'|X) = \mathbb{E}(f'v/\sqrt{N}(f'g/N)'|X) = 0$, $\mathbb{E}(hT_1^{g\prime}|X) = -\mathbb{E}(f'v/\sqrt{N}(f'(I - P^K)g/N)'|X) = 0$, and $\mathbb{E}(T_1^h H_g'|X) = -\mathbb{E}(f'(I - P^K)v/\sqrt{N}H_g'|X) = 0$. Second,

$$\mathbb{E}(T_2^h H_g'|X) = \mathbb{E}(u'P^K v\sqrt{N}H_g'|X) = \frac{K}{\sqrt{N}}\sigma_{uv}H_g'$$

by Lemma 2.3 (v). Second, $\mathbb{E}(hH_g'H^{-1}T_1^{H'}|X) = \mathbb{E}(f'v/\sqrt{N}|X)H_g'H^{-1}(f'(I - P)f/N) = 0$, and $\mathbb{E}(T_1^H H^{-1}hH_g'|X) = 0$. Third, by Lemma 2.3 (vii)

$$\begin{aligned}\mathbb{E}(hH_g'H^{-1}T_2^{H'}|X) &= \mathbb{E}[\frac{f'v}{\sqrt{N}}H_g'H^{-1}\frac{u'f + f'u}{N}|X] \\ &= (\Sigma_i f_i H_g'H^{-1}\sigma_{uv}f_i' + \Sigma_i f_i H_g'H^{-1}f_i\sigma_{uv}')/N^{3/2} = O_p(\frac{1}{\sqrt{N}})\end{aligned}$$

Fourth, by Lemma 2.3 (vi),

$$\begin{aligned}\mathbb{E}(T_2^H H^{-1}hH_g'|X) &= \mathbb{E}[\frac{u'f + f'u}{N}H^{-1}\frac{f'v}{\sqrt{N}}H_g'|X] \\ &= \frac{\sum_i \sigma_{uv}f_i'H^{-1}f_i + \sum_i f_i\sigma_{uv}'H^{-1}f_i}{N^{3/2}}H_g' = O_p(\frac{1}{\sqrt{N}})\end{aligned}$$

Note that $\|T_1^h\|\|T_1^g\| = O_p(\Delta(K)\Delta_g(K)^{1/2}) = o_p(\rho_{K,N})$ by $\Delta(K) = O_p(\rho_{K,N})$ and $\|T_2^h\|\|T_1^g\| = O_p(K/\sqrt{N}\Delta(K)^{1/2}\Delta_g(K)^{1/2})$ by $\Delta(K)^{1/2}K/\sqrt{N} \leq K^2/N + \Delta_K$, thus $Z^{A_3}(K) = (T_1^h + T_2^h)T_1^{g\prime} = o_p(\rho_{K,N})$. Thus, we have

$$\mathbb{E}(\hat{A}_3(K)|X) = \frac{K}{\sqrt{N}}\sigma_{uv}H_g' + o_p(\rho_{K,N}),$$

by $1/\sqrt{N} = o(K/\sqrt{N})$.

In sum,

$$\begin{aligned}
&\mathbb{E}(\hat{A}_1(K) + \hat{A}_2(K) + \hat{A}_3(K) + \hat{A}_3(K)'|X) = \sigma_v^2 H + \sigma_v^2 e_f(K) + \sigma_{uv}\sigma_{uv}' K^2/N + \\
&+ H_g H_g' + H_g H_g' H^{-1} e_f(K) + e_f(K) H^{-1} H_g H_g' + \frac{K}{\sqrt{N}}(H_g \sigma_{uv}' + \sigma_{uv} H_g') \\
&- H_g g'(I - P^K) f/N - f'(I - P^K) g/N H_g + o_p(\rho_{K,N}) \\
&= H\Phi H + HL(K)H + o_p(\rho_{K,N})
\end{aligned}$$

with $\Phi = \sigma_v^2 H^{-1} + H^{-1} H_g H_g' H^{-1}$. We have further simplification, since $K^2/N = o(K/\sqrt{N})$ and this complete the proof. *Q.E.D.*

We also use following Lemma from Donald and Newey (2001). Let $\tilde{\sigma}_v^2 = v'v/N$, and $\tilde{\Lambda}(K) = v' P^K v / N\sigma_v^2$. For the next lemma and the proof of Proposition 2, we denote $\rho_{K,N} = tr(L(K))$ for $L(K)$ from Proposition 2.2 .

Lemma 2.4. *If the hypotheses of Proposition 2.2 are satisfied, then we have*
(i) $\hat{\Lambda}(K) = \tilde{\Lambda}(K) - (\tilde{\sigma}_v^2/\sigma_v^2 - 1)\tilde{\Lambda}(K) - v' f (f'f)^{-1} f' v / 2N\sigma_v^2 + \hat{R}_\Lambda = \tilde{\Lambda}(K) + o_p(K/N)$, $\sqrt{N}\hat{R}_\Lambda = o_p(\rho_{K,N})$
(ii) $u' P^K u/N - \tilde{\Lambda}(K)\Sigma_\eta = o_p(K/N)$,
(iii) $\mathbb{E}(h\tilde{\Lambda}(K) v'\eta/\sqrt{N}|X) = (K/N)\Sigma_i f_i \, \mathbb{E}(v_i^2 \eta_i'|x_i)/N + O_p(K/N^2)$,
(iv) $\mathbb{E}(hh' H^{-1} h/\sqrt{N}|X) = O_p(1/N)$,
(v) $\mathbb{E}(v' P^K v \eta' v/\sqrt{N}|X) = \sum_i P_{ii} \, \mathbb{E}(v_i^3 \eta_i|x_i)/\sqrt{N} = O(K/\sqrt{N})$

Proof of Lemma 2.4 Proof of Lemma 2.4 (i)-(iv) immediately follows from the proof of

Lemma A.7 and A.8. of Donald and Newey (2001). For (v), observe that

$$\begin{aligned}\mathbb{E}(v'P^K v\eta' v/\sqrt{N}|X) &= \sum_{i,j,k}\mathbb{E}(v_i P^K_{ij} v_j \eta_k v_k|X)/\sqrt{N} \\ &= \sum_i P^K_{ii}\,\mathbb{E}(v_i^3\eta_i|x_i)/\sqrt{N} + \sum_i \mathbb{E}(v_i^2|x_i)P^K_{ii}\,\mathbb{E}(\eta_j v_j) \\ &= O(K/\sqrt{N})\end{aligned}$$

because $\mathbb{E}(v_i^3\eta_i|x_i)$ is bounded by Assumption 2.2 and $\mathbb{E}(\eta_j v_j) = 0$ by construction. *Q.E.D.*

Proof of Proposition 2.2

LIML estimator, $\hat{\delta}(K) = (W'P^K W - \hat{\Lambda}(K)W'W)^{-1}(W'P^K y - \hat{\Lambda}(K)W'y)$ has the following form with locally invalid instruments specification;

$$\sqrt{N}(\hat{\delta}(K) - \delta_0) = \hat{H}^{-1}\hat{h} + \hat{H}^{-1}\hat{h}_g,$$

where,

$$\hat{H} = \frac{W'P^K W}{N} - \hat{\Lambda}(K)\frac{W'W}{N}, \quad \hat{h} = \frac{W'P^K v}{\sqrt{N}} - \hat{\Lambda}(K)\frac{W'v}{\sqrt{N}}, \quad \hat{h}_g = \frac{W'P^K g}{N} - \hat{\Lambda}(K)\frac{W'g}{N}.$$

Similar to Donald and Newey (2001), we have following decomposition for $\hat{h}$ and $\hat{H}$

$$\hat{h} = h + \sum_{j=1}^{5} T_j^h + Z^h,$$

$$T_1^h = -f'(I - P^K)v/\sqrt{N} = O_p(\Delta(K)^{1/2}), \quad T_2^h = \eta' P^K v/\sqrt{N} = O_P(\sqrt{K}/\sqrt{N}),$$

$$T_3^h = -\tilde{\Lambda}(K)h = O_p(K/N), \quad T_4^h = -\tilde{\Lambda}(K)\eta' v/\sqrt{N} = O_p(K/N),$$

$$T_5^h = -h' H^{-1} h \sigma_{uv}/2\sqrt{N}\sigma_v^2 = O_p(1/\sqrt{N}),$$

$$Z^h = (\tilde{\Lambda}(K) - \hat{\Lambda}(K) + \hat{R}_\Lambda)\sqrt{N}(\frac{W'v}{N} - \sigma_{uv}) - \hat{R}_\Lambda \frac{W'v}{\sqrt{N}}$$

$$\hat{H} = H + \sum_{j=1}^{3} T_j^H + Z^H,$$

$$T_1^H = -f'(I - P^K)f/N = -e_f(K) = O_p(\Delta(K)), \quad T_2^H = (u'f + f'u)/N = O_p(1/\sqrt{N}),$$

$$T_3^H = -\tilde{\Lambda}(K)H = O_p(K/N)$$

$$Z^H = \frac{u' P^K u}{N} - \tilde{\Lambda}(K)\Sigma_u - \hat{\Lambda}(K)\frac{W'W}{N} + \tilde{\Lambda}(K)(H + \Sigma_u) - u'(I - P^K)f/N - f'(I - P^K)u/N$$

where the O_p results, $T^h = o_p(1)$, $\|T^H\|^2 = o_p(\rho_{K,N})$, $\|T^h\|\|T^H\| = o_p(\rho_{K,N})$, $\|Z^h\| = o_p(\rho_{K,N})$, and $\|Z^H\| = o_p(\rho_{K,N})$ follows similarly to the proof of Proposition 2 in Donald and Newey (2001).

Also, $\hat{h}_g$ is decomposed as

$$\hat{h}_g = H_g + \sum_{j=1}^{4} T_j^g + Z^g,$$

$$T_1^g = -f'(I - P^K)g/N = O_p(\Delta(K)^{1/2}\Delta_g(K)^{1/2}),$$

$$T_2^g = u'g/N = O_p(1/\sqrt{N}), \quad T_3^g = -u'(I - P^K)g/N = O_p(\Delta_g(K)^{1/2}/\sqrt{N}),$$

$$T_4^g = -\tilde{\Lambda}(K)H_g = O_p(K/N)$$

$$Z^g = -\hat{\Lambda}(K)\frac{W'g}{N} + \tilde{\Lambda}(K)H_g$$

where the O_p results follows from the proof of Proposition 2.1 and $\tilde{\Lambda}(K) = O_p(K/N)$. Also,

$$\begin{aligned}\hat{\Lambda}(K)\frac{W'g}{N} - \tilde{\Lambda}(K)H_g &= (\hat{\Lambda}(K) - \tilde{\Lambda}(K))\frac{W'g}{N} + \tilde{\Lambda}(K)(\frac{W'g}{N} - H_g) \\ &= o_p(K/N) + O_p(K/N)o_p(1) = o_p(\rho_{K,N})\end{aligned}$$

by Lemma 2.4 (i), and by the LLN, $W'g/N = H_g + o_p(1)$, which implies $\|Z^g\| = o_p(\rho_{K,N})$. Note that $\|T_1^H\| = O_p(\Delta(K)), \|T_3^H\| = O_p(K/N)$. Since $o_p(\rho_{K,N}) = o_p(K/n + \Delta(K))$, $\|T^g\|\|T_1^H\| = o_p(\rho_{K,N}), \|T^g\|\|T_3^H\| = o_p(\rho_{K,N})$. Moreover,

$$\begin{aligned}\|T^g\|\|T_2^H\| &= O_p(\Delta(K)^{1/2}\Delta_g(K)^{1/2}/\sqrt{N}) + O_p(1/N) + O_p(\Delta_g(K)^{1/2}/N) + O_p(K/N^{3/2}) \\ &= o_p(\rho_{K,N})\end{aligned}$$

by Lemma 2.2 (xi), $1/N = o(K/N)$, and $K/N^{3/2} = o(K/N)$. Thus, $\|T^g\|\|T^H\| = o_p(\rho_{K,N})$ holds by Cauchy-Schwarz inequality.

Next, we calculate expectation of each term $\hat{A}_1(K), \hat{A}_2(K)$, and $\hat{A}_3(K)$ to apply Lemma 2.1. First, for $\hat{A}_1(K) = hh' + h(\sum_{j=1}^5 T_j^h)' + (\sum_{j=1}^5 T_j^h)h' + (\sum_{j=1}^2 T_j^h)(\sum_{j=1}^2 T_j^h)' - hh'H^{-1}T^{H'} - T^H H^{-1}hh'$,

$$\mathbb{E}(\hat{A}_1(K)|X) = \sigma_v^2 H + \sigma_v^2 e_f(K) + \sigma_v^2 \Sigma_\eta \frac{K}{N} + \hat{\zeta} + \hat{\zeta}' + o_p(\rho_{K,N}),$$

and $\|Z^{A_1}(K)\| = o_p(\rho_{K,N})$ by the proof of Proposition 2 in Donald and Newey (2001), where

$$\hat{\zeta} = \sum_i f_i P_{ii}^K \, \mathbb{E}(v_i^2 \eta_i' | X)/N - \frac{K}{N}\sum_i f_i \, \mathbb{E}(v_i^2 \eta_i'|X)/N.$$

Next, for $Z^{A_2}(K) = (\sum_{j=2}^4 T_j^g)(\sum_{j=2}^4 T_j^g)'$, we analyze expectation of $\hat{A}_2(K) = (H_g + T_1^g)(H_g + T_1^g)' + (H_g + T_1^g)(\sum_{j=2}^4 T_j^g)' + (\sum_{j=2}^4 T_j^g)(H_g + T_1^g)' - H_g H_g' H^{-1}T^{H'} - T^H H^{-1} H_g H_g'$. First, note that $H_g T_4^{g'} - H_g H_g' H^{-1} T_3^{H'} = 0$. Second, $\mathbb{E}(H_g T_2^{g'}|X) = \mathbb{E}(H_g g' u/N|X) = H_g \sum_i \mathbb{E}(u_i'|X) g_i/N = 0$. Similarly, we have $\mathbb{E}(H_g T_3^{g'}|X) = 0$, $\mathbb{E}(T_1^g T_2^{g'}|X) = 0$, and

$\mathbb{E}(T_1^g T_3^{g\prime}|X) = 0$. Also,

$$\mathbb{E}(T_1^g T_4^{g\prime}|X) = \mathbb{E}(\frac{f'(I-P^K)g}{N}\tilde{\Lambda}(K)H_g'|X) = \frac{K}{N}\frac{f'(I-P^K)g}{N}H_g' = o_p(\rho_{K,N})$$

by $\mathbb{E}(\tilde{\Lambda}(K)|X) = \mathbb{E}(v'P^K v/N\sigma_v^2|X) = K/N$ and using $K/N = O(\rho_{K,N})$. Lastly, $1/N = o(\rho_{K,N})$, $\Delta_g(K) = o_p(1)$, and $K/N^{3/2} = o(K/N)$ so that $\|T_2^g\|\|T_j^g\|$ for each $j \geq 2$. It also follows similarly that $\|T_3^g\|\|T_3^g\|$, $\|T_3^g\|\|T_4^g\|$ and $\|T_4^g\|\|T_4^g\|$ are $o_p(\rho_{K,N})$. Thus, $Z^{A_2}(K) = o_p(\rho_{K,N})$.

With the calculations in the proof of Proposition 2.1, we have

$$\begin{aligned}\mathbb{E}(\hat{A}_2(K)|X) = {} & H_g H_g' + H_g H_g' H^{-1} e_f(K) + e_f(K) H^{-1} H_g H_g' - H_g g'(I-P^K)f/N \\ & - f'(I-P^K)g/N H_g' + o(\rho_{K,N}).\end{aligned}$$

Next, for $Z^{A_3}(K) = (\sum_{j=3}^5 T_j^h)(\sum_{j=1}^4 T_j^g)' + (\sum_{j=1}^2 T_j^h)(\sum_{j=2}^4 T_j^g)'$, we investigate expectation of $\hat{A}_3(K)$. From the proof of Proposition 2.1, we have calculations of $\mathbb{E}(h(H_g + T_1^{g\prime}) + (T_1^h + T_2^h)H_g' - hH_g' H^{-1}(T_1^H + T_2^H)' - (T_1^H + T_2^H)H^{-1}hH_g'|X)$, except the term

$$\mathbb{E}(T_2^h H_g'|X) = \mathbb{E}(\eta' P^K v/\sqrt{N} H_g'|X) = 1/\sqrt{N}\sum_i P_{ii}^K \mathbb{E}(\eta_i v_i) H_g' = 0.$$

First, note that $hT_4^{g\prime} - hH_g{}' H^{-1} T_3^{H\prime} = 0$ and $T_3^h H_g' - T_3^H H^{-1} h H_g' = 0$. Second, by Lemma 2.3(iv),

$$\mathbb{E}(hT_2^{g\prime}|X) = \mathbb{E}(f'v\sqrt{N}g'u/N|X) = \frac{1}{\sqrt{N}}H_g\sigma_{uv}' = O_p(1/\sqrt{N})$$

Third,

$$\mathbb{E}(hT_3^{g\prime}|X) = -\mathbb{E}(f'v\sqrt{N}g'(I-P^K)u/N|X) = -\frac{1}{\sqrt{N}}\frac{f'(I-P^K)g}{N}\sigma_{uv}' = o_p(\rho_{K,N})$$

Fourth, $\mathbb{E}(T_1^h T_1^{g\prime}|X) = \mathbb{E}(f'(I-P^K)v\sqrt{N}g'(I-P^K)f/N|X) = 0$. Fifth, $\mathbb{E}(T_2^h T_1^{g\prime}|X) =$

$-\mathbb{E}(\eta' P^K v\sqrt{N}|X)g'(I-P^K)f/N = 0$ as $\mathbb{E}\,\eta_i v_i = 0$.

Sixth,

$$\mathbb{E}(T_4^h H_g'|X) = -\mathbb{E}(\frac{v'P^K v}{N\sigma_v^2}\eta' v/\sqrt{N}H_g'|X) = O_p(\sqrt{K}/N^{3/2}) = o_p(\rho_{K,N})$$

by Lemma 2.4(v). Seventh,

$$\mathbb{E}(T_5^h H_g'|X) = -\mathbb{E}(h'H^{-1}h|X)\frac{\sigma_{uv}}{2\sqrt{N}\sigma_v^2}H_g' = \sum_i \mathbb{E}(v_i^2|x_i)f_i'H^{-1}f_i/N\frac{\sigma_{uv}}{2\sqrt{N}\sigma_v^2}H_g' = O_p(1/\sqrt{N}).$$

Lastly, $K/N = o(\sqrt{K}/\sqrt{N})$, $1/\sqrt{N} = o(\sqrt{K}/\sqrt{N})$, and $o_p(\sqrt{K}/\sqrt{N}(\Delta(K)^{1/2}\Delta_g(K)^{1/2} + 1/\sqrt{N} + \Delta_g(K)^{1/2}/\sqrt{N}) + K/N)) = o_p(\rho_{K,N})$, thus $\|T_j^h\|\|T_k^g\| = o_p(\rho_{K,N})$ for $j \geq 3$ and each k. It also follows similarly that $\|T_1^h\|\|T_j^g\| = o_p(\rho_{K,N})$ and $\|T_2^h\|\|T_j^g\| = o_p(\rho_{K,N})$ for $j \geq 2$, and this gives $Z^{A_3}(K) = o_p(\rho_{K,N})$. Thus,

$$\mathbb{E}(\hat{A}_3(K)|X) = O_p(1/\sqrt{N}).$$

In sum, we have,

$$\begin{aligned}
&\mathbb{E}(\hat{A}_1(K) + \hat{A}_2(K) + \hat{A}_3(K) + \hat{A}_3(K)'|X) = \sigma_v^2 H + \sigma_v^2 e_f(K) + \sigma_v^2\Sigma_\eta K/N + \hat{\zeta} + \hat{\zeta}' \\
&\quad + H_g H_g' + H_g H_g' H^{-1} e_f(K) + e_f(K)H^{-1}H_g H_g' + O_p(1/\sqrt{N}) \\
&\quad - H_g g'(I-P^K)f/N - f'(I-P^K)g/N H_g + o_p(\rho_{K,N}) \\
&\quad = H\Phi H + HL(K)H + o_p(\rho_{K,N}),
\end{aligned}$$

where $\Phi = \sigma_v^2 H^{-1} + H^{-1}H_g H_g' H^{-1}$ and $1/\sqrt{N} = o(K/N)$ under the assumption $K^2/N \to \infty$. If we assume $\mathbb{E}(v_i^2\eta_i'|X) = 0$, then $\hat{\zeta} = 0$ and we get the desired results.

For the FULL estimator, observe that

$$\begin{aligned}\check{\Lambda}(K) &= \hat{\Lambda}(K) - \frac{C(1-\hat{\Lambda}(K))^2}{N - C(1-\hat{\Lambda}(K))} \\ &= \hat{\Lambda}(K) + O_p(1/N)\end{aligned}$$

by $0 \leq 1 - \hat{\Lambda}(K) \leq 1$. Therefore we have

$$\begin{aligned}\frac{W'P^KW}{N} - \check{\Lambda}(K)\frac{W'W}{N} &= \frac{W'P^KW}{N} - \hat{\Lambda}(K)\frac{W'W}{N} + o_p(\rho_{K,N}), \\ \frac{W'P^Kv}{\sqrt{N}} - \check{\Lambda}(K)\frac{W'v}{\sqrt{N}} &= \frac{W'P^Kv}{\sqrt{N}} - \hat{\Lambda}(K)\frac{W'v}{\sqrt{N}} + o_p(\rho_{K,N}), \\ \frac{W'P^Kg}{N} - \check{\Lambda}(K)\frac{W'g}{N} &= \frac{W'P^Kg}{N} - \hat{\Lambda}(K)\frac{W'g}{N} + o_p(\rho_{K,N}),\end{aligned}$$

by using $W'W/N = O_p(1), W'v/\sqrt{N} = O_p(1), W'g/N = O_p(1)$ and $1/N = o_p(\rho_{K,N})$. Thus, FULL estimator $\hat{\delta}(K) = (W'P^KW - \check{\Lambda}(K)W'W)^{-1}(W'P^Ky - \check{\Lambda}(K)W'y)$ has the same higher-order MSE decomposition with LIML estimator. *Q.E.D.*

Proof of Proposotion 2.3

For B2SLS estimator, $\hat{\delta}(K) = (W'P^KW - \bar{\Lambda}(K)W'W)^{-1}(W'P^Ky - \bar{\Lambda}(K)W'y)$ with $\bar{\Lambda}(K) = (K-d-2)/N$ has the following decomposition

$$\sqrt{N}(\hat{\delta}(K) - \delta_0) = \hat{H}^{-1}\hat{h} + \hat{H}^{-1}\hat{h}_g,$$

where,

$$\hat{H} = \frac{W'P^KW}{N} - \bar{\Lambda}(K)\frac{W'W}{N}, \quad \hat{h} = \frac{W'P^Kv}{\sqrt{N}} - \hat{\Lambda}(K)\frac{W'v}{\sqrt{N}}, \quad \hat{h}_g = \frac{W'P^Kg}{N} - \bar{\Lambda}(K)\frac{W'g}{N}.$$

We have following decomposition for $\hat{h}, \hat{H}$ and $\hat{h}_g$,

$$\hat{h} = h + \sum_{j=1}^{4} T_j^h,$$

$$T_1^h = -f'(I - P^K)v/\sqrt{N} = O_p(\Delta(K)^{1/2}), \quad T_2^h = u'P^K v/\sqrt{N} - \sqrt{N}\bar{\Lambda}(K)\sigma_{uv} = O_P(\sqrt{K}/\sqrt{N}),$$

$$T_3^h = -\bar{\Lambda}(K)h = O_p(K/N), \quad T_4^h = -\bar{\Lambda}(K)(u'v/N - \sigma_{uv}) = O_p(K/N),$$

$$\hat{H} = H + \sum_{j=1}^{3} T_j^H + Z^H,$$

$$T_1^H = -f'(I - P^K)f/N = -e_f(K) = O_p(\Delta(K)), \quad T_2^H = (u'f + f'u)/N = O_p(1/\sqrt{N}),$$

$$T_3^H = -\bar{\Lambda}(K)H = O_p(K/N)$$

$$Z^H = \frac{u'P^K u}{N} - \bar{\Lambda}(K)\Sigma_u - \bar{\Lambda}(K)(\frac{W'W}{N} - H - \Sigma_u) - u'(I - P^K)f/N - f'(I - P^K)u/N$$

$$\hat{h}_g = H_g + T_1^g + T_2^g + T_3^g,$$

$$T_1^g = -f'(I - P^K)g/N = O_p(\Delta(K)^{1/2}\Delta_g(K)^{1/2}),$$

$$T_2^g = u'g/N = O_p(1/\sqrt{N}), \quad T_3^g = -u'(I - P^K)g/N = O_p(\Delta_g(K)^{1/2}/\sqrt{N}),$$

$$T_4^g = -\bar{\Lambda}(K)H_g = O_p(K/N), \quad Z^g = -\bar{\Lambda}(K)(\frac{W'g}{N} - H_g),$$

where the O_p results, $T^h = o_p(1)$, $\|T^H\|^2 = o_p(\rho_{K,N})$, $\|T^h\|\|T^H\| = o_p(\rho_{K,N})$, $\|Z^h\| = o_p(\rho_{K,N})$, and $\|Z^H\| = o_p(\rho_{K,N})$ follows immediately from the proof of Proposition 3 in Donald and Newey (2001), and $\|Z^g\| = o_p(\rho_{K,N})$, $\|T^g\|\|T^H\| = o_p(\rho_{K,N})$ similarly to the proof of Proposition 2.2 using $\bar{\Lambda}(K) = O(K/N)$.

Next, we calculate expectation of each term $\hat{A}_1(K), \hat{A}_2(K)$, and $\hat{A}_3(K)$ to apply Lemma 2.1. First, for $\hat{A}_1(K) = hh' + h(\sum_{j=1}^4 T_j^h)' + (\sum_{j=1}^4 T_j^h)h' + (\sum_{j=1}^2 T_j^h)(\sum_{j=1}^2 T_j^h)' - hh'H^{-1}T^{H'} - T^H H^{-1}hh'$

$$\mathbb{E}(\hat{A}_1(K)|X) = \sigma_v^2 H + \sigma_v^2 e_f(K) + (\sigma_v^2\Sigma_u + \sigma_{uv}\sigma_{uv}')\frac{K}{N} + \hat{\zeta} + \hat{\zeta}' + o(\rho K, N)$$

by the proof of Proposition 2 in Donald and Newey (2001), where

$$\hat{\zeta} = \sum_i f_i P_{ii}^K \, \mathbb{E}(v_i^2 u_i' | X)/N - \frac{K}{N} \sum_i f_i \, \mathbb{E}(v_i^2 u_i' | X)/N.$$

Next, for $Z^{A_2}(K) = (\sum_{j=2}^4 T_j^g)(\sum_{j=2}^4 T_j^g)'$, we analyze expectation of $\hat{A}_2(K) = (H_g + T_1^g)(H_g + T_1^g)' + (H_g + T_1^g)(\sum_{j=2}^4 T_j^g)' + (\sum_{j=2}^4 T_j^g)(H_g + T_1^g)' - H_g H_g' H^{-1} T^{H\prime} - T^H H^{-1} H_g H_g'$. Observe that $H_g T_4^{g\prime} - H_g H_g' H^{-1} T_3^{H\prime} = 0$ and $\mathbb{E}(T_1^g T_4^{g\prime} | X) = \bar{\Lambda}(K) \frac{f'(I-P^K)g}{N} H_g' = o_p(\rho_{K,N})$. Similar to the proof of Proposition 2.2 by replacing $\bar{\Lambda}(K)$ with $\tilde{\Lambda}(K)$, we have

$$\begin{aligned}\mathbb{E}(\hat{A}_2(K)|X) = {} & H_g H_g' + H_g H_g' H^{-1} e_f(K) + e_f(K) H^{-1} H_g H_g' - H_g g'(I - P^K) f/N \\ & - f'(I - P^K) g/N H_g' + o(\rho_{K,N}).\end{aligned}$$

and $Z^{A_2}(K) = o_p(\rho_{K,N})$. Lastly, for $Z^{A_3}(K) = (\sum_{j=3}^4 T_j^h)(\sum_{j=1}^4 T_j^g)' + (\sum_{j=1}^2 T_j^h)(\sum_{j=2}^4 T_j^g)'$, we investigate expectation of $\hat{A}_3(K)$. From the proof of Proposition 2.2, we have

$$\mathbb{E}(h(H_g + T_1^g + T_2^g + T_3^g)' + T_1^h (H_g + T_1^g)' - h H_g' H^{-1} (T_1^H + T_2^H)' - (T_1^H + T_2^H) H^{-1} h H_g' | X) = O_p(\frac{1}{\sqrt{N}})$$

and $Z^{A_3}(K) = o_p(\rho_{K,N})$. Also observe that $h T_4^{g\prime} - h H_g{}' H^{-1} T_3^{H\prime} = 0$ and $T_3^h H_g' - T_3^H H^{-1} h H_g' = 0$. Next,

$$\mathbb{E}(T_2^h H_g' | X) = \mathbb{E}((u' P^K v/\sqrt{N} - \sqrt{N}\bar{\Lambda}(K)\sigma_{uv}) H_g' | X) = (K/\sqrt{N} - \sqrt{N}\bar{\Lambda}(K))\sigma_{uv} H_g' = O_p(1/\sqrt{N})$$

by $K/\sqrt{N} - \sqrt{N}\bar{\Lambda}(K) \leq C/\sqrt{N}$ for some large C. Similarly,

$$\mathbb{E}(T_2^h T_1^{g\prime} | X) = O_p(\frac{1}{\sqrt{N}} f'(I - P^K) g/N) = o_p(\rho_{K,N}).$$

Lastly, $\mathbb{E}(T_4^h H_g'|X) = -\mathbb{E}(\bar{\Lambda}(K)(u'v/N - \sigma_{uv})H_g'|X) = 0$, and thus

$$\mathbb{E}(\hat{A}_3(K)|X) = O_p(1/\sqrt{N}) + o(\rho_{K,N}).$$

Thus, we have,

$$\begin{aligned}
\mathbb{E}(\hat{A}_1(K) + \hat{A}_2(K) + \hat{A}_3(K) + \hat{A}_3(K)'|X) &= \sigma_v^2 H + \sigma_v^2 e_f(K) + (\sigma_v^2\Sigma_u + \sigma_{uv}\sigma_{uv}')K/N + \hat{\zeta} + \hat{\zeta}' \\
&+ H_g H_g' + H_g H_g' H^{-1} e_f(K) + e_f(K) H^{-1} H_g H_g' - H_g g'(I - P^K) f/N \\
&- f'(I - P^K) g/N H_g + O_p(1/\sqrt{N}) + o_p(\rho_{K,N}) \\
&= H\Phi H + HL(K)H + o_p(\rho_{K,N}).
\end{aligned}$$

where $\Phi = \sigma_v^2 H^{-1} + H^{-1} H_g H_g' H^{-1}$ and $1/\sqrt{N} = o(K/N)$ under the assumption $K^2/N \to \infty$. Using $\sigma_v^2\Sigma_u = \sigma_v^2\Sigma_\eta + \sigma_{uv}\sigma_{uv}'$ and assuming $\mathbb{E}(v_i^2 u_i'|X) = 0$, we get the desired conclusion. *Q.E.D.*

For the proof of Corollary 2.1-2.3, we use the following Lemma to handle higher-order terms due to the $N^{-\gamma}$ ($\gamma > 1/2$) locally invalid instruments specification. This is a slight modification of Lemma 2.1 without $O_p(1)$ term in the decomposition of $\hat{h}_g$. We define $\rho_{K,N} = tr(G + L(K))$.

Lemma 2.5. *If there is a decomposition* $\hat{h} = h + T^h + Z^h, \hat{H} = H + T^H + Z^H, \hat{h}_g = T^g + Z^g$, *and*

$$\begin{aligned}
&(h + T^h)(h + T^h)' - hh'H^{-1}T^{H'} - T^H H^{-1} hh' = \hat{A}_1(K) + Z^{A_1}(K), \\
&T^g T^{g\prime} = \hat{A}_2(K) + Z^{A_2}(K), \quad (h + T^h)T^{g\prime} = \hat{A}_3(K) + Z^{A_3}(K),
\end{aligned}$$

such that $T^h = o_p(1), h = O_p(1), T^g = o_p(1)$, *and* $H = O_p(1)$, *the determinant of* H *is*

bounded away from zero with probability 1, $\rho_{K,N} = tr(G + L(K))$*, and* $\rho_{K,N} = o_p(1)$*,*

$$\|T^H\|^2 = o_p(\rho_{K,N}), \|T^h\|\|T^H\| = o_p(\rho_{K,N}), \|Z^h\| = o_p(\rho_{K,N}), \|Z^H\| = o_p(\rho_{K,N}),$$
$$\|Z^g\| = o_p(\rho_{K,N}), \|T^g\|\|T^H\| = o_p(\rho_{K,N}), \quad Z^{A_i}(K) = o_p(\rho_{K,N}) \quad \textit{for all } i = 1, 2, 3,$$
$$\mathbb{E}(\hat{A}_1(K) + \hat{A}_2(K) + \hat{A}_3(K) + \hat{A}_3(K)'|X) = H\Phi H + H(G + L(K))H + o_p(\rho_{K,N}),$$

then

$$N(\hat{\delta}(K) - \delta_0)(\hat{\delta}(K) - \delta_0)' = \hat{Q}(K) + \hat{r}(K),$$
$$\mathbb{E}(\hat{Q}(K)|X) = \Phi + G + L(K) + T(K),$$
$$[\hat{r}(K) + T(K)]/tr(G + L(K)) = o_p(1), \quad as \quad K \to \infty, N \to \infty.$$

Proof of Lemma 2.5 Follows by Lemma 2.1 replacing $H_g = 0$ with $\rho_{K,N} = tr(G + L(K))$. *Q.E.D.*

Proof of Corollary 2.1

The 2SLS estimator, $\hat{\delta}(K) = (W'P^K W)^{-1}(W'P^K y)$ has the following decomposition with $N^{-\gamma}$ locally invalid instruments specification in model (2.4.1)

$$\sqrt{N}(\hat{\delta}(K) - \delta_0) = \hat{H}^{-1}\hat{h} + \hat{H}^{-1}\hat{h}_g,$$

where $\hat{h}, \hat{H}$ is defined and decomposed as in the proof of Proposition 2.1, but $\hat{h}_g$ can be decomposed as follows,

$$\hat{h}_g = \frac{W'P^K g}{N^{1/2+\gamma}} = T_0^g + T_1^g + Z^g,$$
$$T_0^g = \frac{H_g}{N^{\gamma-1/2}} = O_p(\frac{1}{N^{\gamma-1/2}}), \quad T_1^g = \frac{-f'(I - P^K)g}{N}\frac{1}{N^{\gamma-1/2}} = O_p(\Delta(K)^{1/2}\Delta_g(K)^{1/2}/N^{\gamma-1/2}),$$
$$Z^g = (\frac{u'g}{N} - \frac{u'(I - P^K)g}{N})\frac{1}{N^{\gamma-1/2}} = O_p(1/N^{\gamma} + \Delta_g(K)^{1/2}/N^{\gamma}),$$

where O_p results immediately follows from the proof of Proposition 2.1. We show that the conditions of Lemma 2.5 are satisfied. Note that $G + L(K)$ in the Corollary 2.1 contains the terms of order $1/N^{2\gamma-1}, K/N^{\gamma}$, and K^2/N. It is important to note that these terms may have same order, thus can be the dominant terms in MSE approximation. For example, if $K = O(N^{1-\gamma})$ for $1/2 < \gamma < 1$. Thus to show the term is $o_p(\rho_{K,N})$, it is enough to show $o_p(K/N^{2\gamma-1} + K/N^{\gamma} + K^2/N + \Delta(K))$.

By similar arguments as in the proof of Proposition 2.1, $T^g = o_p(1)$, $\|Z^g\| = o_p(\rho_{K,N})$ by $1/N^{\gamma} = o(K/N^{\gamma})$ and $\|T^g\|\|T^H\| = o(K/N^{\gamma})$. Calculation of the expectation of $\hat{A}_1(K)$ and $Z^{A_1} = o_p(\rho_{K,N})$ defined in Lemma 2.5 follows similarly to the proof of Proposition 2.1. For $\hat{A}_2(K) = (\sum_{j=0}^{1} T_j^g)(\sum_{j=0}^{1} T_j^g)'$ and $Z^{A_2}(K) = 0$, we have

$$\begin{aligned}
\mathbb{E}(\hat{A}_2(K)|X) &= \mathbb{E}(T_0^g T_0^{g\prime}|X) + \mathbb{E}(T_0^g T_1^{g\prime}|X) + \mathbb{E}(T_1^g T_0^{g\prime}|X) + \mathbb{E}(T_1^g T_1^{g\prime}|X) \\
&= \frac{H_g H_g'}{N^{2\gamma-1}} - H_g \frac{g'(I-P^K)f}{N}\frac{1}{N^{2\gamma-1}} - \frac{f'(I-P^K)g}{N} H_g' \frac{1}{N^{2\gamma-1}} + o_p(1/N^{2\gamma-1}) \\
&= \frac{H_g H_g'}{N^{2\gamma-1}} + o_p(\rho_{K,N}),
\end{aligned}$$

by Lemma 2.3 (iii) and $1/N^{2\gamma-1} = O_p(\rho_{K,N})$.

For $Z^{A_3}(K) = (T_1^h + T_2^h)T_1^{g\prime}$, we can easily show $\mathbb{E}(hT_0^{g\prime}|X) = 0, \mathbb{E}(hT_1^{g\prime}|X) = 0$, $\mathbb{E}(T_1^h T_0^{g\prime}|X) = 0$, and

$$\mathbb{E}(T_2^h T_0^{g\prime}|X) = \mathbb{E}(\frac{u'P^K v}{\sqrt{N}}\frac{H_g'}{N^{\gamma-1/2}}|X) = \frac{K}{N^{\gamma}}\sigma_{uv}H_g'$$

Moreover, $Z^{A_3}(K) = o_p(\rho_{K,N})$ by inspection. Thus, $\mathbb{E}(\hat{A}_3(K)|X) = \frac{K}{N^{\gamma}}\sigma_{uv}H_g'$. In sum,

$$\begin{aligned}
&\mathbb{E}(\hat{A}_1(K) + \hat{A}_2(K) + \hat{A}_3(K) + \hat{A}_3(K)'|X) = \sigma_v^2 H + \sigma_v^2 e_f(K) + \sigma_{uv}\sigma_{uv}' K^2/N + \\
&+ \frac{H_g H_g'}{N^{2\gamma-1}} + \frac{K}{N^{\gamma}}(H_g\sigma_{uv}' + \sigma_{uv}H_g') + o_p(\rho_{K,N}) \\
&= H\Phi H + H(G + L(K))H + o_p(\rho_{K,N})
\end{aligned}$$

with $\Phi = \sigma_v^2 H^{-1}$, $G = H^{-1} H_g H_g' H^{-1}/N^{2\gamma-1}$. Second result holds because $O_p(1/N^{2\gamma-1}) = O_p((\frac{N^{1-\gamma}}{K})^2 \frac{K^2}{N}) = o_p(K^2/N)$ under $\frac{K}{N^{1-\gamma}} \to \infty$, and this complete the proof. *Q.E.D.*

Proof of Corollary 2.2

For LIML estimator, we have similar decomposition as in the proof of Proposition 2.2 with $\hat{h}_g$,

$$\hat{h}_g = \sum_{j=0}^{2} T_j^g + Z^g,$$
$$T_0^g = \frac{H_g}{N^{\gamma-1/2}} = O_p(\frac{1}{N^{\gamma-1/2}}), \quad T_1^g = -\frac{f'(I-P^K)g}{N}\frac{1}{N^{\gamma-1/2}} = O_p(\Delta(K)^{1/2}\Delta_g(K)^{1/2}/N^{\gamma-1/2}),$$
$$T_2^g = \frac{u'g}{N}\frac{1}{N^{\gamma-1/2}} = O_p(\frac{1}{N^{\gamma}})$$
$$Z^g = -\frac{u'(I-P^K)g}{N}\frac{1}{N^{\gamma-1/2}} - \tilde{\Lambda}(K)\frac{H_g}{N^{\gamma-1/2}} - \hat{\Lambda}(K)\frac{W'g}{N^{\gamma-1/2}} + \tilde{\Lambda}(K)\frac{H_g}{N^{\gamma-1/2}}$$

where the O_p results and $T^g = o_p(1), Z^g = o_p(\rho_{K,N}), \|T^g\|\|T^H\| = o_p(\rho_{K,N})$ follows from the proof of Proposition 2.2, and using $1/N^\gamma = o_p(\frac{1}{N^{2\gamma-1}} + \frac{K}{N})$. To see $1/N^\gamma = o_p(\frac{1}{N^{2\gamma-1}} + \frac{K}{N})$, consider the function $(K/a)+a$ which is convex, and has a global minimum at $a = \sqrt{K}$ which gives function value $2\sqrt{K}$. Therefore, for $a = N^{1-\gamma}$, $(1/N^\gamma)/(1/N^{2\gamma-1} + K/N) = 1/(a + K/a) \leq 1/(2\sqrt{K}) \to 0$. To show the term is $o_p(\rho_{K,N})$ it is enough to show $o_p(K/N^{2\gamma-1} + K/N + \Delta(K))$.

Calculation of the expectation of $\hat{A}_1(K)$ and $Z^{A_1} = o_p(\rho_{K,N})$ follows from the proof of Proposition 2.2. Next, for $Z^{A_2}(K) = 0$, we have $\mathbb{E}(T_0^g T_2^{g\prime}|X) = 0, \mathbb{E}(T_1^g T_2^{g\prime}|X) = 0$, and

$$\mathbb{E}(T_2^g T_2^{g\prime}|X) = \frac{1}{N^{2\gamma-1}}\frac{1}{N^2}\mathbb{E}(u'gg'u|X) = \frac{1}{N^{2\gamma-1}}\frac{g'g}{N^2}\Sigma_u = o_p(\rho_{K,N})$$

by $1/N^{2\gamma-1} = O_p(\rho_{K,N}), g'g/N = O_p(1)$. Therefore, with calculations in the proof of Corol-

lary 2.1, we have

$$\mathbb{E}(\hat{A}_2(K)|X) = \frac{H_g H_g'}{N^{2\gamma-1}} + o_p(\rho_{K,N}).$$

Next, for $Z^{A_3}(K) = (\sum_{j=3}^{5} T_j^h)(\sum_{j=0}^{2} T_j^g)' + (T_1^h + T_2^h)T_2^{g\prime}$, observe that $\mathbb{E}(hT_0^{g\prime}|X) = 0$, $\mathbb{E}(hT_1^{g\prime}|X) = 0, \mathbb{E}(T_1^h T_0^{g\prime}|X) = 0, \mathbb{E}(T_1^h T_1^{g\prime}|X) = 0, \mathbb{E}(T_1^h T_2^{g\prime}|X) = 0$, and $\mathbb{E}(hT_2^{g\prime}|X) = \frac{1}{N^\gamma} H_g \sigma_{uv}'$. Similar to the proof of Proposition 2.2, $\mathbb{E}(T_2^h T_0^{g\prime}|X) = \mathbb{E}(\eta' P^K v/\sqrt{N} H_g'/N^{\gamma-1/2}|X) = 0$, $\mathbb{E}(T_2^h T_1^{g\prime}|X) = 0$ by using $\sigma_{\eta v} = 0$. Thus,

$$\mathbb{E}(\hat{A}_3(K)|X) = \frac{1}{N^\gamma} H_g \sigma_{uv}' = o_p(\rho_{K,N})$$

by using $O_p(1/N^\gamma) = o_p(\rho_{K,N})$. We can also verify $Z^{A_3}(K) = o_p(\rho_{K,N})$ from the proof of Proposition 2.2 and by inspections. In sum,

$$\mathbb{E}(\hat{A}_1(K) + \hat{A}_2(K) + \hat{A}_3(K) + \hat{A}_3(K)'|X) = \sigma_v^2 H + \sigma_v^2 e_f(K) + \sigma_v^2 \Sigma_\eta K/N + \hat{\zeta} + \hat{\zeta}' + \frac{H_g H_g'}{N^{2\gamma-1}} + o_p(\rho_{K,N})$$
$$= H\Phi H + H(G + L(K))H + o_p(\rho_{K,N})$$

with $\Phi = \sigma_v^2 H^{-1}$, $G = H^{-1} H_g H_g' H^{-1}/N^{2\gamma-1}$. Second result holds because $O_p(1/N^{2\gamma-1}) = O_p((\frac{N^{2-2\gamma}}{K})\frac{K}{N}) = o_p(K/N)$ under $\frac{K}{N^{2-2\gamma}} \to \infty$. The results for FULL estimator follows similarly to the proof of Proposition 2.2, and this complete the proof. *Q.E.D.*

Proof of Corollary 2.3

This follows similarly as in the proof of Corollary 2.2 with the results in the proof of Proposition 2.3. *Q.E.D.*

Proof of Corollary 2.4

Similar to the Proof of Proposition 2.1, the 2SLS estimator, $\hat{\delta}(K) = (W' P^K W)^{-1}(W' P^K y)$

has the following form with locally invalid instruments specification;

$$\sqrt{N}(\hat{\delta}(K) - \delta_0) = \hat{H}^{-1}\hat{h} + \hat{H}^{-1}\hat{h}_g,$$

where,

$$\hat{H} = \frac{W'P^K W}{N}, \hat{h} = \frac{W'P^K v}{\sqrt{N}}, \hat{h}_g = \frac{W'P^K g}{N}.$$

Also, $\hat{h}, \hat{H}$ and $\hat{h}_g$ are decomposed as

$\hat{h} = h + T^h,$

$h = \dfrac{f'v}{\sqrt{N}} + u'P^K v/\sqrt{N} = O_p(1), \quad T^h = -f'(I - P^K)v/\sqrt{N} = o_p(1)$

$\hat{H} = H + T^H,$

$T^H = -f'(I - P^K)f/N + (u'f + f'u)/N + (u'P^K u - u'(I - P^K)f - f'(I - P^K)u)/N = o_p(1),$

$\hat{h}_g = H_g + T^g,$

$T^g = -f'(I - P^K)g/N + u'g/N - u'(I - P^K)g/N = o_p(1)$

It is important to note that h includes term $u'P^K v/\sqrt{N}$ which is $O_p(K/\sqrt{N}) = O_p(1)$ by $K/\sqrt{N} = O(1)$, where $o_p(1)$ results immediately follows from the proof of Proposition 2.1.

By using similar arguments as in the proof of Lemma 2.1, we have

$$\sqrt{N}(\hat{\delta}(K) - \delta_0) = H^{-1}\tilde{\tau} + H^{-1}\tilde{\tau}_g + o_p(1)$$

where $\tilde{\tau} = h + T^h - T^H H^{-1}h = h + o_p(1), \tilde{\tau}_g = H_g + T^g - T^H H^{-1}H_g = H_g + o_p(1)$ by using $T^h = o_p(1), T^H = o_p(1), h = O_p(1), H_g = O_p(1), T^g = o_p(1)$, and $H^{-1} = O_p(1), \hat{H}^{-1} = O_p(1)$.

Since $\sqrt{N}(\hat{\delta}(K) - \delta_0) = H^{-1}h + H^{-1}H_g + o_p(1)$, it follows that

$$N(\hat{\delta}(K) - \delta_0)(\hat{\delta}(K) - \delta_0)' = H^{-1}(hh' + hH_g' + H_g h' + H_g H_g')H^{-1} + o_p(1).$$

First, we have

$$\begin{aligned}
\mathbb{E}(hh'|X) &= \mathbb{E}\left(\frac{f'v}{\sqrt{N}} + \frac{u'P^K v}{\sqrt{N}}\right)\left(\frac{f'v}{\sqrt{N}} + \frac{u'P^K v}{\sqrt{N}}\right)' \\
&= \mathbb{E}\left(\frac{f'vv'f}{N}\right) + \mathbb{E}\left(\frac{f'vv'P^K u}{N}\right) + \mathbb{E}\left(\frac{u'P^K vv'f}{N}\right) + \mathbb{E}\left(\frac{u'P^K vv'P^K u}{N}\right) \\
&= \sigma_v^2 H + \sigma_{uv}\sigma_{uv}'\frac{K^2}{N} + \frac{1}{N}\sum f_i P_{ii}^K \,\mathbb{E}(v_i^2 u_i'|x_i) + \frac{1}{N}\sum P_{ii}^K \,\mathbb{E}(u_i v_i^2|x_i) f_i' + o_p(\frac{K^2}{N}) \\
&= \sigma_v^2 H + \sigma_{uv}\sigma_{uv}'\frac{K^2}{N} + o_p(1).
\end{aligned}$$

by Lemma 2.2 (ix), (x), and $K/N = o(K/\sqrt{N}), K^2/N = O(1)$ Second,

$$\mathbb{E}(hH_g'|X) = \mathbb{E}\left[(\frac{f'v}{\sqrt{N}} + \frac{u'P^K v}{\sqrt{N}})H_g'|X\right] = \frac{K}{\sqrt{N}}\sigma_{uv}H_g'.$$

Therefore,

$$\mathbb{E}(hh' + hH_g' + H_g h' + H_g H_g') = \sigma_v^2 H + \sigma_{uv}\sigma_{uv}'\frac{K^2}{N} + \frac{K}{\sqrt{N}}(\sigma_{uv}H_g' + H_g\sigma_{uv}') + H_g H_g' + o_p(1),$$

so that we get the desired conclusion. *Q.E.D.*

Proof of Corollary 2.5

This immediately follows by Corollary 2.1-2.3 and the Proposition 4 of Donald and Newey (2001) with the Assumption 2.5. *Q.E.D.*