UNIVERSITY OF CALIFORNIA SAN DIEGO

Essays on Non-parametric and High-dimensional Econometrics

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy

in

Economics

by

Zhenting Sun

Committee in charge:

Professor Brendan K. Beare, Co-Chair
Professor Yixiao Sun, Co-Chair
Professor Jelena Bradic
Professor Dimitris Politis
Professor Andres Santos

2018

The Dissertation of Zhenting Sun is approved and is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Co-Chair

University of California San Diego

2018

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Brendan K. Beare and Professor Yixiao Sun for their constant support as the co-chairs of my committee. They do not only help build my knowledge system, but also establish my interests and enthusiasm in econometrics. Their guidance has proved to be invaluable.

I would like to acknowledge Professor Andres Santos from whom I learned a lot of fundamental knowledge in econometrics which is extremely helpful in my study. His valuable comments and suggestions have helped me make significant progress on my research.

I would also like to acknowledge Professor Jelena Bradic and Professor Dimitris Politis for their generous help and insightful comments on my dissertation.

Chapter 1, in part is currently being prepared for submission for publication of the material. Sun, Zhenting. The dissertation author was the primary investigator and author of this material.

Chapter 2, in part is currently being prepared for submission for publication of the material. Sun, Zhenting; Beare, Brendan K. The dissertation author was the primary investigator and author of this material.

Chapter 3, in part is currently being prepared for submission for publication of the material. Sun, Zhenting. The dissertation author was the primary investigator and author of this material.

VITA

2009	Bachelor of Arts, Nankai University, China
2012	Master of Arts, Tsinghua University, China
2016	Master of Arts, University of California San Diego, US
2018	Doctor of Philosophy, University of California San Diego, US

FIELDS OF STUDY

Major Field: Econometric Theory and Applied Econometrics

Research Interests: Micro-econometrics, Non-parametrics and Semi-parametrics, and High-dimensional Big Data

ABSTRACT OF THE DISSERTATION

Essays on Non-parametric and High-dimensional Econometrics

by

Zhenting Sun

Doctor of Philosophy in Economics

University of California San Diego, 2018

Professor Brendan K. Beare, Co-Chair
Professor Yixiao Sun, Co-Chair

Chapter 1 studies the instrument validity for local average treatment effects. we provide a testable implication for instrument validity in the local average treatment effect (LATE) framework with multivalued treatments. Based on this testable implication, we construct a nonparametric test of instrument validity in the multivalued treatment LATE framework. The test is asymptotically consistent. The size of the test can be promoted to the nominal significance level over much of the null, indicating a good power property. Simulation evidence is provided to show the good performance of the test in finite samples. Chapter 2 constructs improved nonparametric bootstrap tests of Lorenz dominance based on preliminary estimation of a contact

set. Our tests achieve the nominal rejection rate asymptotically on the boundary of the null; that is, when Lorenz dominance is satisfied, and the Lorenz curves coincide on some interval. Numerical simulations indicate that our tests enjoy substantially improved power compared to existing procedures at relevant sample sizes. Chapter 3 proposes a sieve focused GMM (SFGMM) estimator for general high-dimensional semiparametric conditional moment models in the presence of endogeneity. Under certain conditions, the SFGMM estimator has oracle consistency properties and converges at a desirable rate. We then establish the asymptotic normality of the plug-in SFGMM estimator for possibly irregular functionals. Simulation evidence illustrates the performance of the proposed estimator.

Chapter 1

Instrument Validity for Local Average Treatment Effects

Abstract

This paper provides a testable implication for instrument validity in the local average treatment effect (LATE) framework with multivalued treatments, generalizing the one obtained by Balke & Pearl (1997), Imbens & Rubin (1997), and Heckman & Vytlacil (2005) for the LATE framework with binary treatments. Based on this testable implication, we construct a nonparametric test of instrument validity in the multivalued treatment LATE framework. Specifically, we transform the testable implication into an inequality involving the value of the supremum of a continuous map over a particular function space. A modified variance-weighted Kolmogorov-Smirnov test statistic is employed in our test. We extend the delta method and establish the asymptotic distribution of the test statistic, which takes a non-standard Kolmogorov-Smirnov form. We then construct the critical value for this asymptotic distribution using the bootstrap method developed by Fang & Santos (2014) and show that the test is asymptotically consistent. The size of the test can be promoted to the nominal significance level over much of the null, indicating a good power property. We also show that with a minor modification the proposed test can easily be applied when there are conditioning covariates with finitely many possible values. Simulation evidence is provided to show the good performance of the test in finite samples. Finally, we use Vietnam-era draft lottery data to illustrate application of the test

in practice.

1.1 Introduction

The local average treatment effect (LATE) framework, introduced by the seminal works Imbens & Angrist (1994) and Angrist *et al.* (1996), is a commonly used approach to study instrumental variables (IV) models with treatment effect heterogeneity. The LATE framework relies on several strong and often controversial assumptions of instrument validity: 1) the instrument should not affect the outcome directly; 2) it should be as good as randomly assigned; and 3) it affects the treatment in a monotone way. Violations of these conditions will generally lead to inconsistent treatment effect estimates. Since the plausibility of the analysis of LATE depends on IV validity, economics research has focused attention on examining these assumptions based on testable implications.

The present paper proposes a testable implication of IV validity in the LATE framework with multivalued treatments, generalizing the testable implication obtained by Balke & Pearl (1997), Imbens & Rubin (1997) and Heckman & Vytlacil (2005) for the LATE framework with binary treatments.[1] To the best of our knowledge, the proposed testable implication is new in the literature. It is stronger than the first-order stochastic dominance condition discussed in Angrist & Imbens (1995). Based on this testable implication, we propose a nonparametric test for IV validity in the LATE framework with binary or multivalued treatments, and with binary or multivalued instruments. Also, we show that with a minor modification, the proposed test can easily be applied when there are discrete conditioning covariates with finitely many possible values, such as gender and age.

Kitagawa (2015) provides a test of IV validity in the LATE framework with binary treatments based on the testable implication in Balke & Pearl (1997), Imbens & Rubin (1997), and Heckman & Vytlacil (2005). This paper uses a variance-weighted Kolmogorov-Smirnov

[1] Studies of LATE with binary treatments can be found in Angrist (1990), Angrist & Krueger (1991), and Vytlacil (2002). Those of LATE with multivalued treatments can be found in Angrist & Imbens (1995), Angrist & Krueger (1995), and Vytlacil (2006).

test statistic and constructs the critical value by a bootstrap method. The test is shown to be uniformly size-controlled and asymptotically consistent, but conservative as the bootstrap critical value converges to a number larger than the $1-\alpha$ quantile of the true asymptotic distribution of the test statistic. Mourifié & Wan (2017) reformulate the testable implication used in Kitagawa (2015) as conditional inequalities, and show that they can be tested in the intersection bounds framework of Chernozhukov *et al.* (2013).[2] Compared to Kitagawa (2015), this test is more convenient to implement. However, it is also conservative and it restricts the support of the outcome variables to be compact, ruling out the case where outcomes are unbounded. Huber & Mellace (2015) derive a testable implication for a weaker LATE identifying condition, that is, the potential outcomes are mean independent of instruments, conditional on each selection type. However, the condition of potential outcomes being mean independent of instruments is not sufficient if we are concerned about distributional features of a complier's potential outcomes, for example, the quantile treatment effects for compliers; see Abadie *et al.* (2002) for details. Our focus in this paper will be on full statistical independence of potential outcomes and instruments.

Since the tests in both Kitagawa (2015) and Mourifié & Wan (2017) are conservative, an important contribution of the present paper is that the proposed test is more powerful when applied in the LATE framework with binary treatments. As shown in Kitagawa (2015) and Mourifié & Wan (2017), the IV validity assumption is refutable but nonverifiable. The testable implication is a necessary but insufficient condition for IV validity; therefore, failing to reject the hypothesis of the testable implication doesn't allow us to confirm IV validity. In this sense, it is always important to improve the power of the test in order to rule out any invalid instruments. The test will be constructed in a framework similar to that in Kitagawa (2015). The key difference is that the proposed test allows multivalued treatments.

We transform the proposed testable implication into an inequality involving the value of the supremum of a continuous map over a particular function space. A modified variance-

[2]It is also worth noting that the test designed by Mourifié & Wan (2017) can easily be implemented using the Stata package of Chernozhukov *et al.* (2014).

weighted Kolmogorov-Smirnov (KS) test statistic is employed in our test.[3] There are two major complications in deriving and approximating the asymptotic distribution of the test statistic. First, the continuous map becomes random after being weighted by an estimated standard deviation using data. As a consequence, the standard delta method cannot be applied for establishing the asymptotic distribution. To overcome this difficulty, we provide an extended delta method that works even when the map is random. This might be of independent interest. By showing that the particular function space is a VC class and applying the extended delta method, we establish the asymptotic distribution of the test statistic and show that it takes the form of a supremum over a smaller function space. Second, since the supremum map is not linear, the standard bootstrap method may fail to approximate this asymptotic distribution consistently.[4] To achieve a consistent approximation, we employ the bootstrap method proposed by Fang & Santos (2014).[5] A consistent estimator of the supremum map in the asymptotic distribution is provided and a series of conditions for the bootstrap method to work are verified. Then we show that the size of this bootstrap-based test can be elevated to the nominal significance level over much of the null, which suggests a good power property. We also show theoretically that the finite sample power of the proposed test is higher than that in Kitagawa (2015) when the test is applied in the binary treatment LATE. This is because this paper's test statistic is equivalent to that used in Kitagawa (2015), but its bootstrap critical value is always smaller.

To implement the test, we propose an empirical approach for choosing the tuning parameter. We find that under certain data generating processes (DGP), the test statistic asymptotic distribution is equivalent to the supremum of a Gaussian process over the whole function space. We then exploit this relationship and choose the tuning parameter from a set of candidates such

[3] As mentioned in Kitagawa (2015), variance-weighted KS statistics have been widely applied in the literature on conditional moment inequalities, such as Andrews & Shi (2013), Armstrong (2014), Armstrong & Chan (2016), and Chetverikov (2018). More general KS statistics can be found in the stochastic dominance testing literature, such as Abadie (2002), Barrett & Donald (2003), Horváth *et al.* (2006), Linton *et al.* (2010), Barrett *et al.* (2014), and Donald & Hsu (2016).

[4] Discussions of this can be found in Hirano & Porter (2012), Fang & Santos (2014), Hong & Li (2016) and Hansen (2017).

[5] Other applications of this bootstrap method can be found in Beare & Moon (2015), Beare & Shi (2018), Seo (Forthcoming), Beare & Fang (2017), and Sun & Beare (2018).

that the critical value constructed by using this tuning parameter is close to an equivalent critical value under a certain DGP. Simulation evidence is provided and shows that the finite sample power of the proposed test is indeed higher than that in Kitagawa (2015) and the empirical size of the test is close to or below the nominal significance level. Finally, we use Vietnam-era draft lottery data to illustrate application of the proposed test in practice.

The remainder of the paper is organized as follows: Section 1.2 introduces the general setup of the LATE framework and the assumptions of IV validity. Based on these assumptions, we provide a testable implication for IV validity in the multivalued treatment LATE framework. Section 1.3 introduces the proposed hypothesis test in the binary treatment LATE framework. We establish the asymptotic distribution of the test statistic and show the improvement in the power of the test. Section 1.4 shows that the test proposed in the previous section can be extended to cases where the treatment and the instrument are multivalued. The case with a continuous instrument is briefly discussed in Section 1.4.2. When one additional condition holds, the proposed test can be applied to continuous instruments. Section 1.5 shows that when the conditioning covariates are discrete variables with finitely many possible values, the test proposed in Section 1.4 can easily be applied in a slightly different framework. Section 1.6 provides an empirical approach of choosing the tuning parameter. Section 1.7 reports the simulation results and compares them with those of Kitagawa (2015). Section 1.8 provides an empirical example of how to examine instrument validity using Vietnam-era draft lottery data. All proofs are contained in the appendix.

1.2 Setup and Testable Implication

To formally introduce the issue of interest, we first briefly introduce the setup of the heterogeneous causal effect model considered in Imbens & Angrist (1994). Let $D \in \{0,1\}$ be the observable treatment variable, where $D = 1$ indicates that an individual receives treatment and $D = 0$ indicates the opposite. Let $Z \in \{0,1\}$ be a binary instrumental variable. Let $Y_{dz} \in \mathscr{Y} \subset \mathbb{R}$, with $d \in \{0,1\}$ and $z \in \{0,1\}$ be the potential outcome variable when $D = d$ and $Z = z$, and let

Y be the observable outcome variable. Similarly, let D_z be the potential treatment variable when $Z = z$.

The instrument validity in the binary treatment LATE framework is formalized by the following assumption.

Assumption 1.2.1 *IV Validity for Binary Z:*

(i) Instrument Exclusion: With probability 1, $Y_{d1} = Y_{d0}$ *for* $d = 0, 1$.

(ii) Random Assignment: The variable Z is jointly independent of $(Y_{11}, Y_{10}, Y_{01}, Y_{00}, D_1, D_0)$.

(iii) Instrument Monotonicity (No defier): The potential treatment response indicators satisfy $D_1 \geq D_0$ *with probability* 1.

Assumption 1.2.1 is almost the same as that in Imbens & Rubin (1997), except that our version of Assumption 1.2.1 does not require strict monotonicity, that is, we don't require the strict inequality in Assumption 1.2.1(iii) to hold with positive probability. The strict monotonicity assumption is also referred to as the instrument relevance assumption, but we do not include it in Assumption 1.2.1. Let $\mathscr{B}_{\mathbb{R}}$ denote the Borel σ-algebra on $\mathbb{R}$. Let $\mathscr{P}$ denote the set of probability measures defined on the Borel σ-algebra of $\mathbb{R}^2$. For every Borel set $B \subset \mathbb{R}$ and $d = 0, 1$, define probability measures as follows:

$$P(B, d) = \mathbb{P}(Y \in B, D = d | Z = 1),$$
$$Q(B, d) = \mathbb{P}(Y \in B, D = d | Z = 0).$$

Clearly $P, Q \in \mathscr{P}$. Under Assumption 1.2.1, we can define $Y_d = Y_{d0} = Y_{d1}$. Imbens & Rubin

(1997) showed that for every Borel subset B of $\mathbb{R}$,

$$\begin{aligned} P(B,1) - Q(B,1) &= \mathbb{P}(Y_1 \in B, D_1 > D_0), \\ Q(B,0) - P(B,0) &= \mathbb{P}(Y_0 \in B, D_1 > D_0). \end{aligned} \tag{1.1}$$

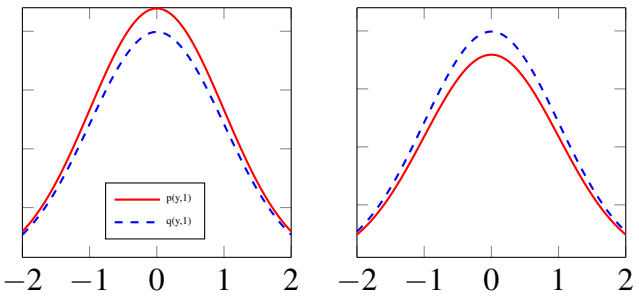
To see why (1.1) is true, we can write

$$\begin{aligned} P(B,1) - Q(B,1) &= \mathbb{P}(Y_{11} \in B, D_1 = 1 | Z = 1) - \mathbb{P}(Y_{10} \in B, D_0 = 1 | Z = 0) \\ &= \mathbb{P}(Y_1 \in B, D_1 = 1) - \mathbb{P}(Y_1 \in B, D_0 = 1) = \mathbb{P}(Y_1 \in B, D_1 = 1, D_0 = 0), \end{aligned}$$

where the second equality follows from Assumption 1.2.1(i) and 1.2.1(ii) and the third equality follows from Assumption 1.2.1(iii). Similar reasoning gives the second equation in (1.1). Since the probabilities in (1.1) are nonnegative, we obtain the testable implication of Assumption 1.2.1 in Balke & Pearl (1997) and Heckman & Vytlacil (2005): For all $B \in \mathscr{B}_{\mathbb{R}}$,

$$\begin{aligned} P(B,1) - Q(B,1) &\geq 0, \\ Q(B,0) - P(B,0) &\geq 0. \end{aligned} \tag{1.2}$$

Proposition 1.1. in Kitagawa (2015) shows an optimality of the testable implication (1.2), namely, that any other feature of the data distribution cannot make a greater contribution to the screening out of invalid instruments than (1.2) can. To understand (1.2) graphically, suppose that Y is a continuous variable and that $p(y,1)$, $p(y,0)$, $q(y,1)$, $q(y,0)$ are density functions or derivatives of the functions $P((-\infty,y],1)$, $P((-\infty,y],0)$, $Q((-\infty,y],1)$ and $Q((-\infty,y],0)$ with respect to y. As functions of y, the later are not probability density functions, because the integral of each of them over the entire real line is not equal to 1. The following graphs show one possible case where (1.2) holds. The first inequality in (1.2) is shown in Figure 1.1a, where the density $p(y,1)$ is greater than $q(y,1)$. The second inequality in (1.2) is shown in Figure 1.1b, where the



(a) $P(B,1) > Q(B,1)$ **(b)** $Q(B,0) > P(B,0)$

Figure 1.1. Graphs of the Testable Implication.

density $q(y,0)$ is greater than $p(y,0)$. Additional graphical examples can be found in Kitagawa (2015).

The LATE framework shown above involves a simple binary treatment and a binary instrument. In many applications, however, D and Z may be multivalued. See, for example, Angrist & Imbens (1995), where the treatment variable is the number of years of schooling completed by a student and can take more than two values.

Suppose, more generally, that $D \in \mathscr{D}_J = \{d_1, d_2, \ldots\}$ and $Z \in \mathscr{Z}_K = \{z_1, z_2, \ldots, z_K\}$. We let $d_{\max}$ be the maximum value of D if it exists, and $d_{\min}$ the minimum value of D if it exists. Suppose the existence of potential variables $Y_{dz} \in \mathscr{Y}$ for $d \in \mathscr{D}_J$ and $z \in \mathscr{Z}_K$ and the existence of D_z for $z \in \mathscr{Z}_K$. The IV validity for the multivalued treatment D and the multivalued instrument Z is formalized by the following assumption.

Assumption 1.2.2 *IV Validity for Multivalued D and Z:*

(i) Instrument Exclusion: With probability 1, $Y_{dz_1} = Y_{dz_2} = \cdots = Y_{dz_K}$ *for all* $d \in \mathscr{D}_J$.

(ii) Random Assignment: The variable Z is jointly independent of $(\tilde{Y}, \tilde{D})$*, where*

$$\tilde{Y} = (Y_{d_1 z_1}, \ldots, Y_{d_1 z_K}, Y_{d_2 z_1}, \ldots, Y_{d_2 z_K}, \ldots),$$
$$\tilde{D} = (D_{z_1}, D_{z_2}, \ldots, D_{z_K}).$$

(iii) Instrument Monotonicity (No defier): The potential treatment response indicators satisfy $D_{z_{k+1}} \geq D_{z_k}$ *with probability* 1 *for all* $k = 1, 2, \ldots, K-1$.

Assumption 1.2.2 is similar to that in Angrist & Imbens (1995). Since we allow multivalued Z, the monotonicity assumption needs to hold for each pair D_{z_k} and $D_{z_{k+1}}$. Define conditional probabilities

$$P_k(B,C) = \mathbb{P}(Y \in B, D \in C | Z = z_k)$$

for all Borel sets $B, C \in \mathscr{B}_{\mathbb{R}}$ and all $z_k \in \mathscr{Z}_K$. The next lemma establishes a testable implication of IV validity in the multivalued treatment LATE when the treatment variable has a maximum value and/or a minimum value.

Lemma 1.2.1 *A testable implication of Assumption 1.2.2 is*

$$\begin{aligned} &P_1(B, \{d_{\max}\}) \leq P_2(B, \{d_{\max}\}) \leq \cdots \leq P_K(B, \{d_{\max}\}), \text{ if } d_{\max} \text{ exists,} \\ &P_1(B, \{d_{\min}\}) \geq P_2(B, \{d_{\min}\}) \geq \cdots \geq P_K(B, \{d_{\min}\}), \text{ if } d_{\min} \text{ exists,} \end{aligned} \tag{1.3}$$

for all $B \in \mathscr{B}_{\mathbb{R}}$, *and*

$$P_1(\mathbb{R}, C) \geq P_2(\mathbb{R}, C) \geq \cdots \geq P_K(\mathbb{R}, C) \tag{1.4}$$

for all $C = (-\infty, c]$ *with* $c \in \mathbb{R}$.

Lemma 1.2.1 generalizes the testable implication (1.2) to the case of a multivalued instrument and the more interesting case of a multivalued treatment. Clearly, when D and Z are both binary, $d_{\max} = 1$ and $d_{\min} = 0$ and (1.3) is equivalent to (1.2). The testable implication (first-order stochastic dominance) discussed by Angrist & Imbens (1995) for Assumption 1.2.2 is equivalent to (1.4). To the best of our knowledge, (1.3) is new in the literature.

1.3 Binary Treatment and Instrument

To highlight the basic idea of our test, this section examines instrument validity in LATE with a binary treatment D and a binary instrument Z based on the testable implication (1.2). We will generalize this test to accommodate a multivalued D and a multivalued Z in the next section.

1.3.1 Hypothesis Formulation

Based on the testable implication (1.2), the hypothesis of the test is formulated as follows:

$$\begin{aligned} &H_0 : P(B,1) - Q(B,1) \geq 0 \text{ and } Q(B,0) - P(B,0) \geq 0 \text{ for all } B \in \mathscr{B}_{\mathbb{R}}, \\ &H_1 : P(B,1) - Q(B,1) < 0 \text{ or } Q(B,0) - P(B,0) < 0 \text{ for some } B \subset \mathscr{B}_{\mathbb{R}}. \end{aligned} \tag{1.5}$$

By Lemma B.7 in Kitagawa (2015), hypothesis (1.5) is equivalent to

$$\begin{aligned} &H_0 : P(B,1) - Q(B,1) \geq 0 \text{ and } Q(B,0) - P(B,0) \geq 0 \text{ for all closed intervals } B \subset \mathbb{R}, \\ &H_1 : P(B,1) - Q(B,1) < 0 \text{ or } Q(B,0) - P(B,0) < 0 \text{ for some closed interval } B \subset \mathbb{R}. \end{aligned} \tag{1.6}$$

Suppose the data set consists of N observations, $\{(Y_i, D_i, Z_i)\}_{i=1}^N \subset \mathscr{Y} \times \{0,1\}^2$. We divide the sample into two subsamples, based on $Z = 0, 1$ respectively. Let $\left\{\left(Y_i^1, D_i^1\right)\right\}_{i=1}^m$ be the subsample for $Z = 1$ and $\left\{\left(Y_i^0, D_i^0\right)\right\}_{i=1}^n$ the subsample for $Z = 0$, with $N = m + n$. We assume that we have a simple random sample.

Assumption 1.3.1 *$\{(Y_i, D_i, Z_i)\}_{i=1}^N$ is an iid data set.*

Assumption 1.3.1 implies that $\left\{\left(Y_i^1, D_i^1\right)\right\}_{i=1}^m$ and $\left\{\left(Y_i^0, D_i^0\right)\right\}_{i=1}^n$ can be regarded as being drawn independently and identically from P and Q, respectively, and that $\left\{\left(Y_i^1, D_i^1\right)\right\}_{i=1}^m$ is independent of $\left\{\left(Y_i^0, D_i^0\right)\right\}_{i=1}^n$. The subsample sizes m, n could be correlated, since $m = \sum_{i=1}^N 1\{Z_i = 1\}$ and $n = \sum_{i=1}^N 1\{Z_i = 0\}$. This would not pose a problem for the performance of the test. The details can be found in the proofs.

Given the subsamples, it follows from Assumption 1.2.1 that $P(B,d)$ and $Q(B,d)$ can be written by

$$P(B,d) = E\left[1\left\{Y^1 \in B, D^1 = d\right\}\right],$$
$$Q(B,d) = E\left[1\left\{Y^0 \in B, D^0 = d\right\}\right],$$

for all closed intervals $B \subset \mathbb{R}$, $d = 0,1$. Define the indicator function $1_A(x) = 1\{x \in A\}$ for every set $A \in \mathbb{R}^k$ and every variable $x \in \mathbb{R}^k$ with $k \in \mathbb{N}$. Then we have

$$P(B,1) - Q(B,1) = E\left[1_{B\times\{1\}}\left(Y^1, D^1\right)\right] - E\left[1_{B\times\{1\}}\left(Y^0, D^0\right)\right],$$
$$Q(B,0) - P(B,0) = E\left[1_{B\times\{0\}}\left(Y^0, D^0\right)\right] - E\left[1_{B\times\{0\}}\left(Y^1, D^1\right)\right].$$

With the above setup, we define a set of functions by

$$\mathscr{H} = \left\{h = (-1)^d \cdot 1_{B\times\{d\}} : B \text{ is a closed interval in } \mathbb{R},\ d \in \{0,1\}\right\}. \tag{1.7}$$

Also, we define $\phi : \mathscr{H} \to \mathbb{R}$ by

$$\phi(h) = E\left[h\left(Y^1, D^1\right)\right] - E\left[h\left(Y^0, D^0\right)\right] \tag{1.8}$$

for all $h \in \mathscr{H}$. Then hypothesis (1.5) is equivalent to

$$H_0 : \sup_{h\in\mathscr{H}} \phi(h) \leq 0,$$
$$H_1 : \sup_{h\in\mathscr{H}} \phi(h) > 0. \tag{1.9}$$

We introduce the following notations, which will also be used later in the paper. For a set

$\mathbb{D}$, denote the space of bounded functions on $\mathbb{D}$ by ℓ^∞:

$$\ell^\infty(\mathbb{D}) = \{f : \mathbb{D} \to \mathbb{R} : \|f\|_\infty < \infty\}, \|f\|_\infty = \sup_{x \in \mathbb{D}} |f(x)|.$$

Then $\ell^\infty(\mathbb{D})$ is a Banach space under $\|\cdot\|_\infty$. If $\mathbb{D}$ is a compact Hausdorff topological space, let $C(\mathbb{D})$ denote the set of continuous maps on $\mathbb{D}$:

$$C(\mathbb{D}) = \{f : \mathbb{D} \to \mathbb{R} : f \text{ is continuous}\}.$$

Then $C(\mathbb{D}) \subset \ell^\infty(\mathbb{D})$ and is also a Banach space under $\|\cdot\|_\infty$. If $\mathbb{D}$ is a metric space with metric d, let $BL_1(\mathbb{D})$ denote the set of all real functions on $\mathbb{D}$ with a Lipschitz norm bounded by 1:

$$BL_1(\mathbb{D}) = \{f : \mathbb{D} \to \mathbb{R} : \|f\|_\infty < \infty, |f(x) - f(z)| \leq d(x,z) \text{ for all } x, z \in \mathbb{D}\}.$$

1.3.2 Test Statistic and Asymptotic Distribution

The test statistic used in this section is a modified version of that used in Kitagawa (2015). Let $T_N = mn/N$.

Assumption 1.3.2 *$m/N \to \lambda$ a.s. as $N \to \infty$, where $\lambda \in (0,1)$.*

By Assumption 1.3.2, m and n grow as $N \to \infty$ in a balanced way. According to our approach to splitting the sample, Assumption 1.3.2 is equivalent to assuming that $\mathbb{P}(Z = 1) = \lambda$. The almost sure convergence holds naturally for and iid data set. We define a probability measure $R = 1/2 \cdot P + 1/2 \cdot Q$.

For every measurable function h, define

$$P(h) = E\left[h\left(Y^1, D^1\right)\right], Q(h) = E\left[h\left(Y^0, D^0\right)\right],$$

and the sample analogue

$$P_m(h) = \frac{1}{m}\sum_{i=1}^{m} h\left(Y_i^1, D_i^1\right), Q_n(h) = \frac{1}{n}\sum_{i=1}^{n} h\left(Y_i^0, D_i^0\right).$$

By definition,

$$\phi(h) = E\left[h\left(Y^1, D^1\right)\right] - E\left[h\left(Y^0, D^0\right)\right] = P(h) - Q(h)$$

for every $h \in \mathscr{H}$. Define the sample analogue

$$\hat{\phi}(h) = P_m(h) - Q_n(h) = \frac{1}{m}\sum_{i=1}^{m} h\left(Y_i^1, D_i^1\right) - \frac{1}{n}\sum_{i=1}^{n} h\left(Y_i^0, D_i^0\right).$$

Then define the asymptotic variance of $\sqrt{T_N}\hat{\phi}(h)$ by

$$\sigma^2(h) = (1-\lambda)|P(h)|(1-|P(h)|) + \lambda|Q(h)|(1-|Q(h)|)$$

and the sample analogue

$$\hat{\sigma}_N^2(h) = (1-\hat{\lambda})|P_m(h)|(1-|P_m(h)|) + \hat{\lambda}|Q_n(h)|(1-|Q_n(h)|)$$

for all h, where $\hat{\lambda} = m/N$.

Let $\ell^\infty(\mathscr{H}) = \{\varphi : \mathscr{H} \to \mathbb{R} : \|\varphi\|_\infty = \sup_{h\in\mathscr{H}} |\varphi(h)| < \infty\}$, and define a map

$$\mathscr{S} : \ell^\infty(\mathscr{H}) \to \mathbb{R}$$

such that for all $\varphi \in \ell^\infty(\mathscr{H})$,

$$\mathscr{S}(\varphi) = \sup_{h\in\mathscr{H}} \varphi(h). \tag{1.10}$$

Now consider the pointwise ratios $[\phi/(\xi \vee \sigma)](h)$ and $[\hat{\phi}/(\xi \vee \hat{\sigma}_N)](h)$ on $\mathscr{H}$, where $\xi \in (0,1)$ is a user specified parameter. Clearly, $\phi/(\xi \vee \sigma), \hat{\phi}/(\xi \vee \hat{\sigma}_N) \in \ell^\infty(\mathscr{H})$, and set the

test statistic, TS_N, to

$$TS_N = \sqrt{T_N}\mathscr{S}\left(\frac{\hat{\phi}}{\xi \vee \hat{\sigma}_N}\right). \tag{1.11}$$

Now we introduce a theorem that establishes the asymptotic distribution of the test statistic under H_0. Define $\Psi_{\mathscr{H}} = \{h \in \mathscr{H} : \phi(h) = \mathscr{S}(\phi)\}$ and the map $\mathscr{S}_{\Psi_{\mathscr{H}}} : \ell^\infty(\mathscr{H}) \to \mathbb{R}$ such that for all $\psi \in \ell^\infty(\mathscr{H})$,

$$\mathscr{S}_{\Psi_{\mathscr{H}}}(\psi) = \sup_{h \in \Psi_{\mathscr{H}}} \psi(h). \tag{1.12}$$

Theorem 1.3.1 *Suppose the underlying probabilities P and Q are fixed as $N \to \infty$. Under Assumptions 1.3.1 and 1.3.2,*

$$\sqrt{T_N}\left(\hat{\phi} - \phi\right) \rightsquigarrow \sqrt{1-\lambda}\mathbb{G}_P - \sqrt{\lambda}\mathbb{G}_Q, \tag{1.13}$$

and under H_0, we obtain the asymptotic distribution of the test statistic:

$$\sqrt{T_N}\left(\mathscr{S}\left(\frac{\hat{\phi}}{\xi \vee \hat{\sigma}_N}\right) - \mathscr{S}\left(\frac{\phi}{\xi \vee \hat{\sigma}_N}\right)\right) \rightsquigarrow \mathscr{S}_{\Psi_{\mathscr{H}}}\left(\frac{\sqrt{1-\lambda}\mathbb{G}_P - \sqrt{\lambda}\mathbb{G}_Q}{\xi \vee \sigma}\right), \tag{1.14}$$

with

$$\mathscr{S}_{\Psi_{\mathscr{H}}}\left(\frac{\sqrt{1-\lambda}\mathbb{G}_P - \sqrt{\lambda}\mathbb{G}_Q}{\xi \vee \sigma}\right) \stackrel{L}{=} \mathscr{S}_{\Psi_{\mathscr{H}}}\left(\frac{\mathbb{G}_H}{\xi \vee \sigma}\right), \tag{1.15}$$

where $\mathbb{G}_P, \mathbb{G}_Q$ are a P-Browian bridge and a Q-Browian bridge, respectively, $H = \lambda P + (1-\lambda)Q$, $\mathbb{G}_H$ is an H-Browian bridge, "$\rightsquigarrow$" denotes weak convergence, and "$\stackrel{L}{=}$" denotes equivalence in law.

The weak convergence in (1.13) is basically due to the fact that $\mathscr{H}$ is a VC class, as established by Lemma A.1.4 in the appendix. It is worth noting that because $\phi/(\xi \vee \hat{\sigma}_N)$ is random, it is not straightforward to apply the standard delta methods to obtain the weak

convergence in (1.14). We extend the standard delta method in Lemma A.1.2 in the appendix so that it can be applied to such a "random parameter" situation. This might be of independent interest. The details can be found in Lemma A.1.2.

1.3.3 Bootstrap-Based Inference

With the limiting distribution in (1.14), we construct the critical value by a bootstrap method and establish the testing theory. As discussed in Theorems 3.1 and 3.2 in Fang & Santos (2014), since $\mathbb{G}_H/(\xi \vee \sigma)$ is centered Gaussian and $\mathscr{S}_{\Psi_{\mathscr{H}}}$ is nonlinear, the standard bootstrap method may fail to approximate the limiting distribution in (1.14) consistently. Thus we employ the bootstrap method proposed in Fang & Santos (2014).

First, we need to obtain an estimation of $\mathscr{S}_{\Psi_{\mathscr{H}}}$. This is because $\mathscr{S}_{\Psi_{\mathscr{H}}}$ is determined by ϕ, which is unknown and has to be estimated. By (1.12), $\mathscr{S}_{\Psi_{\mathscr{H}}}$ is an operator that involves the set $\Psi_{\mathscr{H}}$. Thus if we can find a "valid" estimator $\hat{\Psi}_{\mathscr{H}}$ for $\Psi_{\mathscr{H}}$, then a natural approximation of $\mathscr{S}_{\Psi_{\mathscr{H}}}$ denoted by $\hat{\mathscr{S}}_N$ can be constructed by

$$\hat{\mathscr{S}}_N(\psi) = \sup_{h \in \hat{\Psi}_{\mathscr{H}}} \psi(h), \psi \in C(\mathscr{H}).$$

If H_0 is true, then since $1_{\{a\}\times\{0\}}, -1_{\{a\}\times\{1\}} \in \mathscr{H}$ for all $a \in \mathbb{R}$, we have $\mathscr{S}(\phi) = 0$. By the definition of $\Psi_{\mathscr{H}}$, we can conclude that under H_0,

$$\Psi_{\mathscr{H}} = \{h \in \mathscr{H} : \phi(h) = 0\}.$$

This is similar to what is called the contact set in Linton *et al.* (2010). Then we construct $\hat{\Psi}_{\mathscr{H}}$ naturally by

$$\hat{\Psi}_{\mathscr{H}} = \left\{h \in \mathscr{H} : |\hat{\phi}(h)| \leq \tau_N\right\}, \tag{1.16}$$

where $\tau_N \downarrow 0$ but $\tau_N\sqrt{T_N} \to \infty$. This rate follows from the weak convergence in (1.13). Intuitively, we do not want to exclude too many h from $\hat{\Psi}_{\mathscr{H}}$ as $\hat{\phi}$ converges to ϕ. Lemma A.2.1 in the

appendix shows that $\hat{\Psi}_{\mathscr{H}}$ is a valid estimator for $\Psi_{\mathscr{H}}$, so we can construct $\hat{\mathscr{S}}_N$ by plugging in $\hat{\Psi}_{\mathscr{H}}$.

Test Procedure

With the estimator $\hat{\mathscr{S}}_N$, we introduce the procedure for the bootstrap-based test.

(1) Obtain the bootstrap samples $\left\{\left(Y_i^{1*}, D_i^{1*}\right)\right\}_{i=1}^m$ and $\left\{\left(Y_i^{0*}, D_i^{0*}\right)\right\}_{i=1}^n$ drawn with replacement from the subsample $\left\{\left(Y_i^1, D_i^1\right)\right\}_{i=1}^m$ and $\left\{\left(Y_i^0, D_i^0\right)\right\}_{i=1}^n$ respectively.

(2) Calculate the bootstrap version of ϕ by

$$\hat{\phi}^*(h) = \hat{P}_m^*(h) - \hat{Q}_n^*(h),$$

and the bootstrap version of σ by

$$\hat{\sigma}_N^*(h) = \sqrt{\left(1-\hat{\lambda}\right)\left|\hat{P}_m^*(h)\right|\left(1-\left|\hat{P}_m^*(h)\right|\right) + \hat{\lambda}\left|\hat{Q}_n^*(h)\right|\left(1-\left|\hat{Q}_n^*(h)\right|\right)},$$

where $\hat{P}_m^*(h) = m^{-1}\sum_{i=1}^m h\left(Y_i^{1*}, D_i^{1*}\right)$ and $\hat{Q}_n^*(h) = n^{-1}\sum_{i=1}^n h\left(Y_i^{0*}, D_i^{0*}\right)$.

(3) Calculate the bootstrap version of the test statistic by $\hat{\mathscr{S}}_N(\sqrt{T_N}(\hat{\phi}^* - \hat{\phi})/(\xi \vee \hat{\sigma}_N^*))$.

(4) Repeat (1), (2), and (3) many times and obtain the empirical distribution of $\hat{\mathscr{S}}_N(\sqrt{T_N}(\hat{\phi}^* - \hat{\phi})/(\xi \vee \hat{\sigma}_N^*))$. Given nominal significance level α, calculate the bootstrap critical value $\hat{c}_{1-\alpha}$ by

$$\hat{c}_{1-\alpha} = \inf\left\{c : \mathbb{P}\left(\hat{\mathscr{S}}_N\left(\frac{\sqrt{T_N}(\hat{\phi}^* - \hat{\phi})}{\xi \vee \hat{\sigma}_N^*}\right) \leq c | \{(Y_i^1, D_i^1)\}_{i=1}^m, \{(Y_i^0, D_i^0)\}_{i=1}^n\right) \geq 1-\alpha\right\}.$$

(5) The decision rule for the test is:

$$\text{Reject } H_0 \text{ if } \sqrt{T_N}\mathscr{S}(\hat{\phi}) > \hat{c}_{1-\alpha}. \tag{1.17}$$

The difference between this bootstrap method and the standard bootstrap method is that we use $\hat{\mathscr{S}}_N(\sqrt{T_N}(\hat{\phi}^* - \hat{\phi})/(\xi \vee \hat{\sigma}_N^*))$ instead of $\sqrt{T_N}(S(\hat{\phi}^*/(\xi \vee \hat{\sigma}_N^*)) - S(\hat{\phi}/(\xi \vee \hat{\sigma}_N^*)))$ to construct critical values.

Theorem 1.3.2 *Suppose Assumptions 1.3.1 and 1.3.2 hold. Then under decision rule* (2.33)*:*

(i) If H_0 is true and the CDF of $\mathscr{S}_{\Psi_{\mathscr{H}}}(\mathbb{G}_H/(\xi \vee \sigma))$ is strictly increasing and continuous at its $1-\alpha$ quantile $c_{1-\alpha}$, then $\lim_{N\to\infty}\mathbb{P}(\text{reject } H_0) = \alpha$.

(ii) If H_0 is false, then $\lim_{N\to\infty}\mathbb{P}(\text{reject } H_0) = 1$.

Theorem 11.1 in Davydov *et al.* (1998) implies that the CDF of $\mathscr{S}_{\Psi_{\mathscr{H}}}(\mathbb{G}_H/(\xi \vee \sigma))$ is differentiable and has a positive derivative everywhere except at countably many points in its support, provided that $\mathscr{S}_{\Psi_{\mathscr{H}}}(\mathbb{G}_H/(\xi \vee \sigma)) \neq 0$. Thus Theorem 1.3.2(i) shows that the asymptotic size of the test can be promoted to the nominal significance level α over much of the null. This suggests a good power property of the test. The remark below shows theoretically that the finite sample power of the proposed test is higher than that in Kitagawa (2015).

Remark 1.3.1 *Kitagawa (2015) approximates the distribution of $\mathbb{G}_H$ using a different bootstrap estimator, denoted by $\sqrt{T_N}\hat{\phi}^\sharp$ here, rather than $\sqrt{T_N}(\hat{\phi}^* - \hat{\phi})$. It can be shown that Theorem 1.3.2 also holds if we use $\sqrt{T_N}\hat{\phi}^\sharp$ instead of $\sqrt{T_N}(\hat{\phi}^* - \hat{\phi})$. By definition,*

$$\hat{\mathscr{S}}_N\left(\frac{\sqrt{T_N}\hat{\phi}^\sharp}{\xi \vee \hat{\sigma}_N^*}\right) \leq \mathscr{S}\left(\frac{\sqrt{T_N}\hat{\phi}^\sharp}{\xi \vee \hat{\sigma}_N^*}\right) \text{ a.s.} \tag{1.18}$$

Since the test statistic used in this paper is equivalent to that in Kitagawa (2015), (1.18) *shows that the proposed test has a larger finite sample power because Kitagawa (2015) uses the quantity on the right-hand side of* (1.18) *to construct the bootstrap critical value, while we use the quantity on the left-hand side.*

1.4 Multivalued Treatment and Instrument

In this section, we extend the testing theory from the previous section to the case where the treatment and/or the instrument is a multivalued discrete variable. The test is constructed on the testable implication (1.3). Without loss of generality, we assume that both $d_{\max}$ and $d_{\min}$ exist and that $d_{\max} = 1$ and $d_{\min} = 0$. By definition, for all $B, C \in \mathscr{B}_{\mathbb{R}}$,

$$P_k(B,C) = \mathbb{P}(Y \in B, D \in C | Z = z_k) = \frac{\mathbb{P}(Y \in B, D \in C, Z = z_k)}{\mathbb{P}(Z = z_k)}.$$

Define function spaces

$$\begin{aligned}
\mathscr{G}_K &= \left\{ 1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}} : k = 1, 2, \cdots, K \right\}, \\
\mathscr{G} &= \left\{ \left(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z_{k+1}\}} \right) : k = 1, 2, \cdots, K-1 \right\}, \\
\mathscr{H}_{K1} &= \left\{ (-1)^d \cdot 1_{B\times\{d\}\times\mathscr{Z}_K} : B \text{ is a closed interval, } d = 0, 1 \right\}, \\
\mathscr{H}_{K2} &= \left\{ 1_{\mathbb{R}\times C\times\mathscr{Z}_K} : C = (-\infty, c], c \in \mathbb{R} \right\}, \\
\mathscr{H}_K &= \mathscr{H}_{K1} \cup \mathscr{H}_{K2}.
\end{aligned}$$

By Lemma B.7 in Kitagawa (2015), we use all closed intervals $B \subset \mathbb{R}$ to construct $\mathscr{H}_{K1}$ instead of all Borel sets.

Let $\mathscr{P}_3$ be the set of probability measures on $\mathbb{R}^3$ and let $P \in \mathscr{P}_3$ be the probability measure induced by the joint distribution of (Y, D, Z). For every measurable function h, define

$$P(h) = \int h dP.$$

Define for every $(h, g) \in \mathscr{H}_K \times \mathscr{G}$ with $g = (g_1, g_2)$,

$$\phi_K(h, g) = \frac{P(h \cdot g_2)}{P(g_2)} - \frac{P(h \cdot g_1)}{P(g_1)}$$

and the sample analogue

$$\hat{\phi}_K(h,g) = \frac{\frac{1}{N}\sum_{i=1}^N (h\cdot g_2)(Y_i,D_i,Z_i)}{\frac{1}{N}\sum_{i=1}^N g_2(Y_i,D_i,Z_i)} - \frac{\frac{1}{N}\sum_{i=1}^N (h\cdot g_1)(Y_i,D_i,Z_i)}{\frac{1}{N}\sum_{i=1}^N g_1(Y_i,D_i,Z_i)}.$$

For example, for every closed interval B, $d \in \{0,1\}$, and $k = 1,2,\ldots,K-1$, and for $h = (-1)^d \cdot 1_{B\times\{d\}\times\mathscr{Z}_K}$ and $g = (g_1,g_2) = \left(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z_{k+1}\}}\right)$,

$$\begin{aligned}\phi_K(h,g) &= (-1)^d \cdot \frac{\mathbb{P}(Y\in B, D=d, Z=z_{k+1})}{\mathbb{P}(Z=z_{k+1})} - (-1)^d \cdot \frac{\mathbb{P}(Y\in B, D=d, Z=z_k)}{\mathbb{P}(Z=z_k)} \\ &= (-1)^d \cdot P_{k+1}(B,d) - (-1)^d \cdot P_k(B,d).\end{aligned}$$

Obviously, if D,Z are binary and we let $g_1 = 1_{\mathbb{R}\times\mathbb{R}\times\{0\}}$ and $g_2 = 1_{\mathbb{R}\times\mathbb{R}\times\{1\}}$, then

$$\sum_{i=1}^N 1_{\mathbb{R}\times\mathbb{R}\times\{1\}}(Y_i,D_i,Z_i) = m, \quad \sum_{i=1}^N 1_{\mathbb{R}\times\mathbb{R}\times\{0\}}(Y_i,D_i,Z_i) = n.$$

In that case, m and n are subsample sizes defined in Section 1.3.

Define a map $\mathscr{S}_K : \ell^\infty(\mathscr{H}_K\times\mathscr{G}) \to \mathbb{R}$ by

$$\mathscr{S}_K(\psi) = \sup_{(h,g)\in\mathscr{H}_K\times\mathscr{G}} \psi(h,g),$$

for every $\psi \in \ell^\infty(\mathscr{H}_K\times\mathscr{G})$. Then the testable implication (1.3) is equivalent to

$$H_0 : \mathscr{S}_K(\phi_K) \le 0,$$
$$H_1 : \mathscr{S}_K(\phi_K) > 0.$$

Define a metric on $\mathscr{H}_K\times\mathscr{G}$ such that for all $(h,g),(h',g') \in \mathscr{H}_K\times\mathscr{G}$,

$$\rho_P\left((h,g),(h',g')\right) = \left\|h-h'\right\|_{L^2(P)} + \left\|g_1-g_1'\right\|_{L^2(P)} + \left\|g_2-g_2'\right\|_{L^2(P)}.$$

Lemma 1.4.1 *Under Assumption 1.3.1,* $\sqrt{N}(\hat{\phi}_K - \phi_K) \rightsquigarrow \mathbb{G}_K$ *for some Gaussian process* $\mathbb{G}_K$*, and for all* $(h,g) \in \mathscr{H}_K \times \mathscr{G}$*,*

$$Var\left(\mathbb{G}_K(g,h)\right) = \frac{|P(h\cdot g_2)|}{P^2(g_2)}\left(1 - \frac{|P(h\cdot g_2)|}{P(g_2)}\right) + \frac{|P(h\cdot g_1)|}{P^2(g_1)}\left(1 - \frac{|P(h\cdot g_1)|}{P(g_1)}\right). \tag{1.19}$$

Lemma 1.4.1 provides the asymptotic distribution of $\sqrt{N}(\hat{\phi}_K - \phi_K)$ and its asymptotic variance. This asymptotic variance will be used later, when we construct the test statistic later.

By (1.19), for every $(h,g) \in \mathscr{H}_K \times \mathscr{G}$ and $g = (g_1, g_2)$, let

$$\sigma_K^2(h,g) = \frac{|P(h\cdot g_2)|}{P^2(g_2)}\left(1 - \frac{|P(h\cdot g_2)|}{P(g_2)}\right) + \frac{|P(h\cdot g_1)|}{P^2(g_1)}\left(1 - \frac{|P(h\cdot g_1)|}{P(g_1)}\right).$$

Similarly, define the sample analogue of $\sigma_K^2(h,g)$ by

$$\hat{\sigma}_{KN}^2(h,g) = \frac{|P_N(h\cdot g_2)|}{P_N^2(g_2)}\left(1 - \frac{|P_N(h\cdot g_2)|}{P_N(g_2)}\right) + \frac{|P_N(h\cdot g_1)|}{P_N^2(g_1)}\left(1 - \frac{|P_N(h\cdot g_1)|}{P_N(g_1)}\right),$$

where P_N is the empirical probability measure of P such that for every measurable function f,

$$P_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(Y_i, D_i, Z_i).$$

For multivalued D and Z, set the test statistic, MTS_N, to

$$MTS_N = \sqrt{N}\mathscr{S}_K\left(\frac{\hat{\phi}_K}{\xi \vee \hat{\sigma}_{KN}}\right). \tag{1.20}$$

Define $\Psi_{\mathscr{H}_K\times\mathscr{G}} = \{(h,g) \in \mathscr{H}_K \times \mathscr{G} : \phi_K(h,g) = \mathscr{S}_K(\phi_K)\}$. It is not hard to see that $\Psi_{\mathscr{H}_K\times\mathscr{G}} \neq \varnothing$. Also, define $\mathscr{S}_{\Psi_{\mathscr{H}_K\times\mathscr{G}}} : \ell^\infty(\mathscr{H}_K \times \mathscr{G}) \to \mathbb{R}$ such that for all $\psi \in \ell^\infty(\mathscr{H}_K \times \mathscr{G})$,

$$\mathscr{S}_{\Psi_{\mathscr{H}_K\times\mathscr{G}}}(\psi) = \sup_{(h,g)\in\Psi_{\mathscr{H}_K\times\mathscr{G}}} \psi(h,g).$$

Theorem 1.4.1 *Suppose Assumption 1.3.1 holds. Then under H_0,*

$$\sqrt{N}\left\{\mathscr{S}_K\left(\frac{\hat{\phi}_K}{\xi \vee \hat{\sigma}_{KN}}\right) - \mathscr{S}_K\left(\frac{\phi_K}{\xi \vee \hat{\sigma}_{KN}}\right)\right\} \rightsquigarrow \mathscr{S}_{\Psi_{\mathscr{H}_K \times \mathscr{G}}}\left(\frac{\mathbb{G}_K}{\xi \vee \sigma_K}\right).$$

Theorem 1.4.1 shows the asymptotic distribution of the test statistic for multivalued treatments and instruments. By Lemma 1.4.1 and Lemma A.1.12 in the appendix, we establish the weak convergence $(\hat{\phi}_K - \phi_K)/(\xi \vee \hat{\sigma}_{KN}) \rightsquigarrow \mathbb{G}_K/(\xi \vee \sigma_K)$. Then similarly to Theorem 1.3.1, the extended delta method in Lemma A.1.2 in the appendix is applied to obtain the limiting distribution in Theorem 1.4.1.

1.4.1 Bootstrap-Based Inference

Similarly to the binary treatment and binary instrument case, we need to obtain an estimation of the map $\mathscr{S}_{\Psi_{\mathscr{H}_K \times \mathscr{G}}}$. Equivalently we need to find a "valid" estimator $\hat{\Psi}_{\mathscr{H}_K \times \mathscr{G}}$ for $\Psi_{\mathscr{H}_K \times \mathscr{G}}$. Then a natural approximation of $\mathscr{S}_{\Psi_{\mathscr{H}_K \times \mathscr{G}}}$, which we denote by $\hat{\mathscr{S}}_{KN}$, can be constructed by

$$\hat{\mathscr{S}}_{KN}(\psi) = \sup_{h \in \hat{\Psi}_{\mathscr{H}_K \times \mathscr{G}}} \psi(h), \psi \in C(\mathscr{H}).$$

If H_0 is true, then since $1_{\{a\}\times\{0\}\times\mathscr{Z}_K}, -1_{\{a\}\times\{1\}\times\mathscr{Z}_K} \in \mathscr{H}_K$ for all $a \in \mathbb{R}$, we have $\mathscr{S}_K(\phi_K) = 0$. By the definition of $\Psi_{\mathscr{H}_K \times \mathscr{G}}$, we can conclude that under H_0,

$$\Psi_{\mathscr{H}_K \times \mathscr{G}} = \{(h,g) \in \mathscr{H}_K \times \mathscr{G} : \phi_K(h,g) = 0\}.$$

Then we construct $\hat{\Psi}_{\mathscr{H}}$ naturally by

$$\hat{\Psi}_{\mathscr{H}_K \times \mathscr{G}} = \left\{(h,g) \in \mathscr{H}_K \times \mathscr{G} : |\hat{\phi}_K(h,g)| \leq \tau_N\right\}, \tag{1.21}$$

where $\tau_N \downarrow 0$ but $\tau_N\sqrt{N} \to \infty$.

Lemma A.3.1 in the appendix is a result similar to Lemma A.2.1 and shows that $\hat{\Psi}_{\mathscr{H}_K \times \mathscr{G}}$

is a valid estimator for $\Psi_{\mathscr{H}_K \times \mathscr{G}}$, so we can construct $\hat{\mathscr{S}}_{KN}$ by plugging in $\hat{\Psi}_{\mathscr{H}_K \times \mathscr{G}}$.

Test Procedure

Now we introduce the procedure for the test in the case of multivalued D and multivalued Z.

(1) Obtain the bootstrap samples $\{(Y_i^*, D_i^*, Z_i^*)\}_{i=1}^N$ drawn with replacement from the sample $\{(Y_i, D_i, Z_i)\}_{i=1}^N$.

(2) Calculate the bootstrap version of ϕ_K by

$$\hat{\phi}_K^*(h,g) = \frac{\hat{P}_N^*(h \cdot g_2)}{\hat{P}_N^*(g_2)} - \frac{\hat{P}_N^*(h \cdot g_1)}{\hat{P}_N^*(g_1)}$$

for all $h \in \mathscr{H}_K$ and $g \in \mathscr{G}$, and calculate the bootstrap version of σ by

$$\hat{\sigma}_{KN}^*(h,g) = \sqrt{\frac{\left|\hat{P}_N^*(h \cdot g_2)\right|}{\hat{P}_N^{*2}(g_2)}\left(1 - \frac{\left|\hat{P}_N^*(h \cdot g_2)\right|}{\hat{P}_N^*(g_2)}\right) + \frac{\left|\hat{P}_N^*(h \cdot g_1)\right|}{\hat{P}_N^{*2}(g_1)}\left(1 - \frac{\left|\hat{P}_N^*(h \cdot g_1)\right|}{\hat{P}_N^*(g_1)}\right)},$$

where $\hat{P}_N^*(v) = N^{-1}\sum_{i=1}^N v(Y_i^*, D_i^*, Z_i^*)$.

(3) Calculate the bootstrap version of the test statistic by $\hat{\mathscr{S}}_{KN}(\sqrt{N}(\hat{\phi}_K^* - \hat{\phi}_K)/(\xi \vee \hat{\sigma}_{KN}^*))$.

(4) Repeat (1), (2), and (3) many times and obtain the empirical distribution of $\hat{\mathscr{S}}_{KN}(\sqrt{N}(\hat{\phi}_K^* - \hat{\phi}_K)/(\xi \vee \hat{\sigma}_{KN}^*))$. Given the nominal significance level α, calculate the bootstrap critical value $\hat{c}_{1-\alpha}$ by

$$\hat{c}_{1-\alpha} = \inf\left\{c : \mathbb{P}\left(\hat{\mathscr{S}}_{KN}\left(\frac{\sqrt{N}(\hat{\phi}_K^* - \hat{\phi}_K)}{\xi \vee \hat{\sigma}_{KN}^*}\right) \leq c | \{(Y_i, D_i, Z_i)\}_i^N\right) \geq 1 - \alpha\right\}.$$

(5) The decision rule for the test is:

$$\text{Reject } H_0 \text{ if } \sqrt{N}\mathscr{S}_K(\hat{\phi}) > \hat{c}_{1-\alpha}. \tag{1.22}$$

Theorem 1.4.2 *Suppose Assumption 1.3.1 holds. Then under decision rule* (1.22)*:*

(i) If H_0 is true and the CDF of $\mathscr{S}_{\Psi_{\mathscr{H}_K \times \mathscr{G}}}(\mathbb{G}_K/(\xi \vee \sigma_K))$ is strictly increasing and continuous at its $1-\alpha$ quantile $c_{1-\alpha}$, then $\lim_{N\to\infty}\mathbb{P}(\text{reject } H_0) = \alpha$.

(ii) If H_0 is false, then $\lim_{N\to\infty}\mathbb{P}(\text{reject } H_0) = 1$.

The proof of Theorem 1.4.2 is similar to that of Theorem 1.3.2, so we won't repeat it. Theorem 1.4.2 establishes the testing theory for IV validity in the LATE framework with a multivalued treatment and a multivalued instrument.

1.4.2 Continuous Instrument

In this section, we briefly discuss the case where Z is continuous. For simplicity, suppose $D = 0,1$ and $Z \in \mathscr{Z} \subset \mathbb{R}$. As mentioned in Cornelissen *et al.* (2016), when Z is continuous the monotonicity assumption needs to hold between all pairs of values $z, z' \in \mathscr{Z}$ so that the IV LATE estimators can capture the average treatment effect for compliers with a change in the instrument from z to z'. However, it is not quite possible to compute all pairwise LATEs with a continuous instrument, because the number of observations in a sample for every pair (z,z') is likely to be small. For the same reason, it is not straightforward to do the test for the continuous instrument case based on the framework we introduced earlier. A practical way to exploit a continuous instrument is to partition its support into discrete groups, since we would be interested in the average treatment effect for compliers with a change of the instrument from one group to another, provided that there is additional information about the treatment variable D in each group. Suppose we are interested in a partition

$$\mathscr{Z} = C_1 \cup C_2 \cup \cdots \cup C_K,$$

where $C_1, C_2, \ldots, C_K$ are disjoint subsets of $\mathbb{R}$. Suppose there exist potential variables $Y_d(z) \in \mathscr{Y}$ for $d = 0,1$ and $z \in \mathscr{Z}$, and $D(z)$ for $z \in \mathscr{Z}$.

Assumption 1.4.1 *IV Validity for continuous Z:*

(i) Instrument Exclusion: With probability 1, $Y_d(z) = Y_d(z')$ *for* $d = 0,1$ *and all* $z, z' \in \mathscr{Z}$.

(ii) Random Assignment: The variable Z is jointly independent of $(Y_1(z), Y_0(z), D(z))$ *with* $z \in \mathscr{Z}$.

(iii) Instrument Monotonicity (No defier): The potential treatment response indicators satisfy $D(z') \geq D(z)$ *with probability* 1, *where* z' *and z are prespecified.*

Assumption 1.4.2 $D(z) = D(z')$ *for all* $z, z' \in C_k$ *and all k.*

Assumption 1.4.2 requires D to be grouped by the partition of $\mathscr{Z}$. If Assumptions 1.4.1 and 1.4.2 hold, we can construct potential variables D_k for $Z \in C_k$, that is, $D_k = D(z)\,1\{z \in C_k\}$. Then another IV validity condition for continuous instrument Z is formalized by the following assumption.

Assumption 1.4.3 *IV Validity for Continuous Z with D grouped by the partition of Z:*

(i) Instrument Exclusion: With probability 1, $Y_d(z) = Y_d(z')$, *for* $d = 0,1$ *and all* $z, z' \in \mathscr{Z}$.

(ii) Random Assignment: The variable Z is jointly independent of $(Y_1(z), Y_0(z), D_1, D_2, \cdots, D_K)$.

(iii) Instrument Monotonicity (No defier): The potential treatment response indicators satisfy $D_{k+1} \geq D_k$ *with probability* 1 *for all* $k \in \{1, 2, \ldots, K-1\}$.

Define probability measures

$$P_k(B, d) = \mathbb{P}(Y \in B, D = d | Z \in C_k),$$

for every Borel set $B \in \mathscr{B}_{\mathbb{R}}$ and $d = 0, 1$. The testable implication for Assumption 1.4.3 can be constructed as

$$\begin{aligned} &P_1(B,1) \leq P_2(B,1) \leq \cdots \leq P_K(B,1), \\ &P_1(B,0) \geq P_2(B,0) \geq \cdots \geq P_K(B,0). \end{aligned} \tag{1.23}$$

By definition,

$$P_k(B,d) = \frac{\mathbb{P}(Y \in B, D = d, Z \in C_k)}{\mathbb{P}(Z \in C_k)}.$$

Define

$$\mathscr{G} = \left\{ \left(1_{B\times\{d\}\times C_k}, 1_{B\times\{d\}\times C_{k+1}}\right) : k = 1, 2, \ldots, K-1 \right\}$$

and

$$\mathscr{H}_K = \left\{ (-1)^d \cdot 1_{B\times\{d\}\times\mathbb{R}} : B \text{ is a closed interval, } d = 0, 1 \right\}.$$

Let P be the probability measure on $\mathbb{R}^3$ corresponding to the joint distribution of (Y, D, Z). For any h, define

$$P(h) = \int h dP.$$

For every closed interval B, $d \in \{0,1\}$ and $k \in \{1, 2, \cdots, K-1\}$, and for any $(h, g) \in \mathscr{H}_K \times \mathscr{G}$ with $h = (-1)^d \cdot 1_{B\times\{d\}\times\mathbb{R}}$ and $g = (g_1, g_2) = \left(1_{\mathbb{R}\times\{0,1\}\times C_k}, 1_{\mathbb{R}\times\{0,1\}\times C_{k+1}}\right)$,

$$\begin{aligned} \phi_K(h,g) &= \frac{P(h\cdot g_2)}{P(g_2)} - \frac{P(h\cdot g_1)}{P(g_1)} \\ &= (-1)^d \cdot \frac{\mathbb{P}(Y \in B, D = d, Z \in C_{k+1})}{\mathbb{P}(Z \in C_{k+1})} - (-1)^d \cdot \frac{\mathbb{P}(Y \in B, D = d, Z \in C_k)}{\mathbb{P}(Z \in C_k)} \\ &= (-1)^d \cdot P_{k+1}(B,d) - (-1)^d \cdot P_k(B,d). \end{aligned}$$

Then define a map $\mathscr{S}_K : \ell^\infty(\mathscr{H}_K \times \mathscr{G}) \to \mathbb{R}$ by

$$\mathscr{S}_K(\phi) = \sup_{(h,g)\in\mathscr{H}_K\times\mathscr{G}} \phi(h,g),$$

for all $\phi \in \ell^\infty(\mathscr{H}_K \times \mathscr{G})$. Then the testable implication (1.23) is equivalent to

$$H_0 : \mathscr{S}_K(\phi_K) \leq 0,$$
$$H_1 : \mathscr{S}_K(\phi_K) > 0.$$

The testing process and results are then similar to those for the case of multivalued D and multivalued Z.

1.5 Conditional on Discrete Covariates

For simplicity, we consider the case where $D = 0,1$ and assume that X is a one-dimensional variable. A testable implication for the conditional version of the inequalities in (1.2) is given by

$$\begin{aligned}
&\mathbb{P}(Y \in B, D = 1 | Z = z_k, X) \leq \mathbb{P}(Y \in B, D = 1 | Z = z_{k+1}, X) \text{ a.s.} \\
&\mathbb{P}(Y \in B, D = 0 | Z = z_k, X) \geq \mathbb{P}(Y \in B, D = 0 | Z = z_{k+1}, X) \text{ a.s.}
\end{aligned} \tag{1.24}$$

Suppose X is discrete and let $\mathscr{X}_L$ be the set of possible values of X with $\mathscr{X}_L = \{x_1, x_2, \ldots, x_L\}$. Then for every Borel set B and $d = 0,1$,

$$\mathbb{P}(Y \in B, D = d | Z = z_k, X = x_l) = \frac{\mathbb{P}(Y \in B, D = d, Z = z_k, X = x_l)}{\mathbb{P}(Z = z_k, X = x_l)}.$$

Define

$$\mathscr{G}_{KL} = \left\{ 1_{\mathbb{R}\times\{0,1\}\times\{z_k\}\times\{x_l\}} : k = 1,2,\ldots,K, l = 1,2,\ldots,L \right\}$$

and

$$\mathscr{G} = \left\{ \left(1_{\mathbb{R}\times\{0,1\}\times\{z_k\}\times\{x_l\}}, 1_{\mathbb{R}\times\{0,1\}\times\{z_{k+1}\}\times\{x_l\}} \right) : k = 1,2,\ldots,K-1, l = 1,2,\ldots,L \right\}.$$

Then define

$$\mathscr{H}_{KL} = \left\{ (-1)^d \cdot 1_{B\times\{d\}\times\mathscr{Z}_K\times\mathscr{X}_L} : B \text{ is a closed interval, } d = 0,1 \right\}.$$

Let $\mathscr{P}_4$ be the set of probability measures on $\mathbb{R}^4$. Let $P \in \mathscr{P}_4$ be the probability measure corresponding to the joint distribution of (Y,D,Z,X). For every measurable h, define

$$P(h) = \int h dP.$$

For every interval B, $d \in \{0,1\}$, $k = 1,2,\ldots,K-1$, and $l = 1,2,\ldots,L$, and for $(h,g) \in \mathscr{H}_{KL}\times\mathscr{G}$ with $h = (-1)^d \cdot 1_{B\times\{d\}\times\mathscr{Z}_K\times\mathscr{X}_L}$ and $g = (g_1,g_2) = \left(1_{\mathbb{R}\times\{0,1\}\times\{z_k\}\times\{x_l\}}, 1_{\mathbb{R}\times\{0,1\}\times\{z_{k+1}\}\times\{x_l\}} \right)$, define

$$\begin{aligned}
\phi_{KL}(h,g) =& \frac{P(h\cdot g_2)}{P(g_2)} - \frac{P(h\cdot g_1)}{P(g_1)} \\
=& (-1)^d \cdot \mathbb{P}(Y \in B, D = d | Z = z_{k+1}, X = x_l) \\
& - (-1)^d \cdot \mathbb{P}(Y \in B, D = d | Z = z_k, X = x_l).
\end{aligned}$$

The define a map $\mathscr{S}_{KL} : \ell^\infty(\mathscr{H}_{KL}\times\mathscr{G}) \to \mathbb{R}$ by

$$\mathscr{S}_{KL}(\phi) = \sup_{(h,g)\in\mathscr{H}_{KL}\times\mathscr{G}} \phi(h,g)$$

for all $\phi \in \ell^{\infty}(\mathscr{H}_{KL} \times \mathscr{G})$. Then the testable implication (1.24) is equivalent to

$$H_0 : \mathscr{S}_{KL}(\phi_{KL}) \leq 0,$$
$$H_1 : \mathscr{S}_{KL}(\phi_{KL}) > 0.$$

The testing process and results are then similar to those for the case of multivalued D and multivalued Z.

1.6 Tuning Parameter Selection

For simplicity, we discuss the approach for choosing the tuning parameter under the binary treatment and binary instrument framework in Section 1.3. It is straightforward to extend this approach to the case of a multivalued treatment and a multivalued instrument. As we see in Lemma A.2.1 in the appendix, we need to let τ_N to decay to 0 at a certain rate in order to obtain a consistent estimator $\hat{\mathscr{S}}_N$. However, it is not obvious how to choose the value of τ_N for a given sample size. Notice that under H_0, if $\phi = 0$ everywhere, then $\mathscr{H} = \Psi_{\mathscr{H}}$ and therefore

$$\mathscr{S}\left(\frac{\mathbb{G}_H}{\xi \vee \sigma}\right) = \mathscr{S}_{\Psi_{\mathscr{H}}}\left(\frac{\mathbb{G}_H}{\xi \vee \sigma}\right) \text{ a.s.} \tag{1.25}$$

As we stated earlier, we can construct the bootstrap critical value by two methods. The first one is the method proposed in the present paper using $\hat{\mathscr{S}}_N(\sqrt{T_N}\hat{\phi}^{\sharp}/(\xi \vee \hat{\sigma}_N^*))$, and the second one is the method proposed in Kitagawa (2015) using $\mathscr{S}(\sqrt{T_N}\hat{\phi}^{\sharp}/(\xi \vee \hat{\sigma}_N^*))$; the relationship between them is given in (1.18). When (1.25) holds, the bootstrap critical values constructed with the two method should be close to each other, since they converge to the $1-\alpha$ quantile of the same distribution in (1.25). Thus given that $\phi = 0$ everywhere, we can choose τ_N so that the two critical values are sufficiently close.

(1) Predetermine a set of candidate values of τ_N, denoted by C_τ.

(2) Given a data set $\left\{\left(Y_i^1, D_i^1\right)\right\}_{i=1}^m$ from P, draw a sample of size m with replacement from $\left\{\left(Y_i^1, D_i^1\right)\right\}_{i=1}^m$, denoted by $\left\{\left(Y_i^{B1}, D_i^{B1}\right)\right\}_{i=1}^m$, and draw a sample of size n with replacement from $\left\{\left(Y_i^1, D_i^1\right)\right\}_{i=1}^m$, denoted by $\left\{\left(Y_i^{B0}, D_i^{B0}\right)\right\}_{i=1}^n$. Pretend that $\left\{\left(Y_i^{B1}, D_i^{B1}\right)\right\}_{i=1}^m$ and $\left\{\left(Y_i^{B0}, D_i^{B0}\right)\right\}_{i=1}^n$ are the samples from the population distribution P and Q, respectively.

(3) Compute $\hat{c}_{1-\alpha}$ with the data set $\left\{\left(Y_i^{B1}, D_i^{B1}\right)\right\}_{i=1}^m, \left\{\left(Y_i^{B0}, D_i^{B0}\right)\right\}_{i=1}^n$ by using each value from the candidate set C_τ. Also, compute the bootstrap critical value from Kitagawa (2015) with the same data set.

(4) Choose the value of τ_N such that the two bootstrap critical values are sufficiently close.

In Step (2), by generating the bootstrap samples $\left\{\left(Y_i^{B1}, D_i^{B1}\right)\right\}_{i=1}^m$ and $\left\{\left(Y_i^{B0}, D_i^{B0}\right)\right\}_{i=1}^n$ by resampling from $\left\{\left(Y_i^1, D_i^1\right)\right\}_{i=1}^m$, we approximate the setting where $\phi = 0$ everywhere.

1.7 Simulation Evidence

The Monte Carlo experiments conducted in this section follow the construction in Kitagawa (2015), so we can compare the results and show the improvement in the power of the test when the test is applied in the LATE framework with a binary treatment. We simulated the limiting rejection rates from the test proposed in the present paper and that proposed in Kitagawa (2015), using the same randomly generated data.

There were a total of 6 data-generating processes for H_0 and H_1. Each simulation consisted of 1000 Monte Carlo iterations and 1000 bootstrap iterations. The user-specified trimming parameter ξ was set to 0.07, as suggested by Kitagawa (2015). 5 sets of sample sizes were considered: $(m,n) = (100,100)$, $(100,500)$, $(500,500)$, $(100,1000)$, and $(1000,1000)$. The set of candidate values of τ_N was $\{0.00, 1.02, \ldots, 0.10\}$ for each of the sample sizes, and we chose τ_N by the approach proposed earlier. When calculating the supremum value in test statistics and the bootstrap critical values, we followed the numerical computation approach used in Kitagawa (2015). Specifically, we considered all the closed intervals with the values of

Y observed in the data as endpoints. To expedite the simulation, we employed the warp-speed method in Giacomini *et al.* (2013). Also, when we calculated the bootstrap version of $\hat{\phi}$, we followed the method used in Kitagawa (2015) because that reduced the amount of computation and made it easy for us to compare the results.

1.7.1 Data-Generating Processes

The data-generating processes under H_0 and H_1 in the binary D and binary Z framework:

(1) H_0 is true:

DGP 1: For $z \in \{0,1\}$, $\mathbb{P}(D^z = 1) = 0.5$ and $Y^z = D^z \cdot N(1,1) + (1-D^z) \cdot N(0,1)$, where the superscripts denote the subsamples as before.

DGP 2: $\mathbb{P}\left(D^1 = 1\right) = 0.5$ and $\mathbb{P}\left(D^0 = 1\right) = 0.48$. For $z \in \{0,1\}$, $Y^z = D^z \cdot N(1,1) + (1-D^z) \cdot N(0,1)$.

(2) H_1 is true:

Let $\mathbb{P}\left(D^1 = 1\right) = 0.55$, and $\mathbb{P}\left(D^0 = 1\right) = 0.45$, and $Y^1 = D^1 \cdot N(0,1) + \left(1-D^1\right) \cdot N(0,1)$.

DGP 1: $Y^0 = D^0 \cdot N(-0.7,1) + \left(1-D^0\right) \cdot N(0,1)$.

DGP 2: $Y^0 = D^0 \cdot N\left(0, 1.675^2\right) + \left(1-D^0\right) \cdot N(0,1)$.

DGP 3: $Y^0 = D^0 \cdot N\left(0, 0.515^2\right) + \left(1-D^0\right) \cdot N(0,1)$.

DGP 4: $Y^0 = D^0 \cdot W + \left(1-D^0\right) \cdot N(0,1)$, where $W = \sum_{k=1}^{5} 1\{K=k\} N\left(\mu_k, 0.125^2\right)$, $(\mathbb{P}(K=1), \mathbb{P}(K=2), \mathbb{P}(K=3), \mathbb{P}(K=4), \mathbb{P}(K=5)) = (0.15, 0.2, 0.3, 0.2, 0.15)$, and $(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5) = (-1, -0.5, 0, 0.5, 1)$.

Figure 1.2 shows how DGPs 1–4 violate H_0.

1.7.2 Simulation Results

Tables 1.1 and 1.2 show the simulated rejection rates in the two cases under H_0 from the test proposed in the present paper and that proposed by Kitagawa (2015). The rejection rates are

slightly upwardly biased but close to the nominal significance levels. With the chosen τ_N's, the limiting rejection rates under the null or the empirical significance levels of the proposed test are slightly higher than those in Kitagawa (2015).

Table 1.1. Rejection Rates under H_0.

DGPs:		DGP1			DGP2		
α:		0.01	0.05	0.10	0.01	0.05	0.10
(m,n):	$(100,100)$	0.017	0.069	0.134	0.009	0.055	0.108
	$(100,500)$	0.023	0.062	0.114	0.011	0.061	0.122
	$(500,500)$	0.014	0.076	0.132	0.013	0.040	0.091
	$(100,1000)$	0.017	0.049	0.127	0.014	0.032	0.095
	$(1000,1000)$	0.016	0.062	0.114	0.010	0.038	0.080

Table 1.2. Rejection Rates under H_0 by Kitagawa (2015).

DGPs:		DGP1			DGP2		
α:		0.01	0.05	0.10	0.01	0.05	0.10
(m,n):	$(100,100)$	0.017	0.065	0.111	0.009	0.048	0.089
	$(100,500)$	0.023	0.062	0.097	0.011	0.047	0.074
	$(500,500)$	0.014	0.071	0.129	0.013	0.036	0.090
	$(100,1000)$	0.017	0.043	0.127	0.014	0.032	0.095
	$(1000,1000)$	0.014	0.062	0.114	0.007	0.034	0.080

Tables 1.3 and 1.4 show the simulated rejection rates in the two cases under H_1. We find that most of the rejection rates are larger than those obtained by the test of Kitagawa (2015), which shows an improvement in the test power.

Table 1.3. Rejection Rates under H_1.

DGPs:		DGP1			DGP2			DGP3			DGP4		
α:		0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
(m,n):	$(100,100)$	0.053	0.193	0.344	0.013	0.045	0.209	0.069	0.188	0.404	0.017	0.095	0.181
	$(100,500)$	0.130	0.354	0.449	0.079	0.283	0.428	0.103	0.270	0.322	0.042	0.183	0.226
	$(500,500)$	0.708	0.847	0.919	0.740	0.897	0.974	0.536	0.706	0.860	0.103	0.363	0.562
	$(100,1000)$	0.157	0.299	0.460	0.191	0.270	0.446	0.128	0.184	0.373	0.054	0.165	0.266
	$(1000,1000)$	0.986	0.999	0.998	0.996	0.999	1.000	0.881	0.958	0.961	0.381	0.671	0.759

Table 1.4. Rejection Rates under H_1 by Kitagawa (2015).

DGPs:		DGP1			DGP2			DGP3			DGP4		
α:		0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
(m,n):	$(100,100)$	0.049	0.187	0.263	0.013	0.044	0.117	0.068	0.176	0.315	0.017	0.084	0.132
	$(100,500)$	0.117	0.294	0.363	0.070	0.225	0.339	0.086	0.239	0.322	0.042	0.132	0.226
	$(500,500)$	0.699	0.847	0.879	0.732	0.897	0.943	0.536	0.695	0.804	0.103	0.363	0.458
	$(100,1000)$	0.124	0.257	0.391	0.098	0.174	0.379	0.115	0.184	0.322	0.035	0.151	0.229
	$(1000,1000)$	0.986	0.999	0.997	0.996	0.999	1.000	0.881	0.957	0.961	0.381	0.664	0.756

Additional Results for Randomly Chosen Intervals

When the sample size is large, the numerical computation approach used in Kitagawa (2015) for calculating the supremum value in the test statistic and the bootstrap critical value is time-consuming. Here we employ another approach. We randomly choose 40,000 closed intervals with endpoints between the maximum and minimum values of Y obtained in the data. Then calculate the supremum value over all these 40,000 closed intervals.

The data-generating processes were the same as earlier. Each simulation consisted of 1000 Monte Carlo iterations and 1000 bootstrap iterations. The user-specified trimming parameter ξ was set to 0.07, 0.30 and 1.00. The sample size was $(m,n) = (3000, 8000)$. The set of candidate values of τ_N was $\{0.00, 1.02, \ldots, 0.10\}$.

Tables 1.5 and 1.6 show that under H_0 all the rejection rates are close to or below the nominal significance levels and under H_1 all the rejection rates are close to 1.

Table 1.5. Rejection Rates under H_0 with Randomly Chosen Intervals.

Measures:		DGP 1			DGP 2		
α:		0.01	0.05	0.10	0.01	0.05	0.10
ξ:	0.07	0.010	0.047	0.097	0.007	0.013	0.039
	0.30	0.010	0.058	0.089	0.000	0.011	0.007
	1.00	0.006	0.058	0.089	0.000	0.000	0.009

Table 1.6. Rejection Rates under H_1 with Randomly Chosen Intervals.

DGPs:		DGP1			DGP2			DGP3			DGP4		
α:		0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
(ξ):	0.07	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000
	0.30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	1.000
	1.00	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.969	0.988	0.998

1.8 Empirical Applications

We illustrate the performance of the proposed test in practice by examining the instrument of the Vietnam-era draft lottery used in Angrist & Krueger (1992) and Angrist & Krueger (1995). Details of the Vietnam era draft lottery can be found in Angrist (1990). We follow Abadie (2002) and define a binary draft eligibility instrument (Z) by a dummy variable that indicates whether one's lottery number is less than or equal to 100. The data set we used is a subsample of the data used in Angrist & Krueger (1992) and Angrist & Krueger (1995), which was taken from the March Current Population Surveys in 1979 and 1981–1985. There are a total of 30,967 men in the sample. After eliminating the people who had missing values or did not work during the year, the sample size was 26,119. Finally, we kept only the people who were born in 1950 through 1953. The final sample size was 11,291. Similarly to Kitagawa (2015), two outcome measures (Y) were used in the test, annual labor earnings and weekly wages, which were measured in terms of 1978 dollars using the consumer price index (CPI). Weekly earnings were imputed by

the annual labor earnings divided by the number of weeks worked. The treatment variable D indicates whether a man had Vietnam veteran status.

By using the data above, the number of people with $Z = 1$ was $m = 3125$, and the number of people with $Z = 0$ was $n = 8166$. $\hat{\mathbb{P}}(D = 1|Z = 1) = 0.3094$ and $\hat{\mathbb{P}}(D = 1|Z = 0) = 0.1876$. In Table 1.7, we show the empirical p-values obtained from the test for each $\xi = 0.07, 0.30, 1.00$ and $\alpha = 0.00, 1.05, 0.10$. For each pair (ξ, α), we chose the value of the tuning parameter τ_N. When we calculated the test statistic and the bootstrap test statistic, we didn't apply the numerical method mentioned earlier because the sample size was large and as a result the calculation would have been slow. Instead, we randomly chose 40,000 closed intervals with endpoints between the maximum and minimum values of Y observed in the data. As shown in the table, all the empirical p-values are close to 1, so we failed to reject the validity of the instrument.

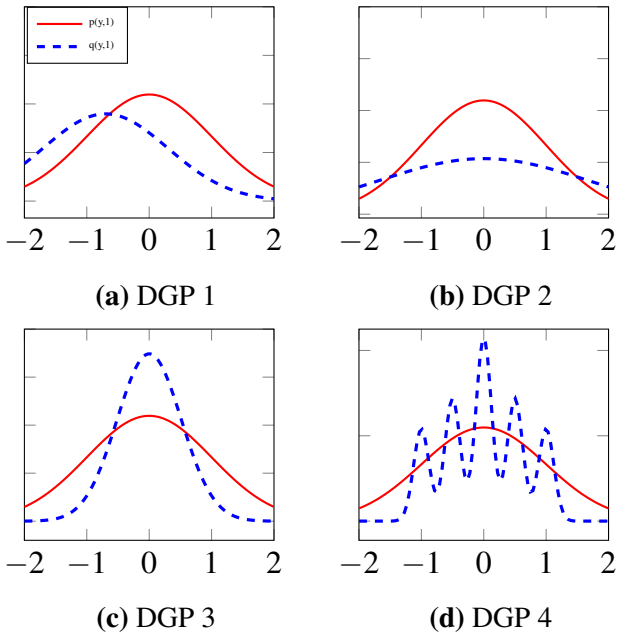
Table 1.7. p-Values of Validity Test for Draft Lottery.

Measures:		Annual Earnings			Weekly Wages		
α:		0.01	0.05	0.10	0.01	0.05	0.10
ξ:	0.07	1.000	0.999	0.989	0.973	1.000	0.996
	0.30	1.000	0.995	0.999	0.996	1.000	1.000
	1.00	1.000	1.000	1.000	0.996	1.000	0.998

1.9 Conclusion

In this paper, we have provided a testable implication for instrument validity in the LATE framework with multivalued treatments. Based on this testable implication, we have constructed a nonparametric test of instrument validity for the multivalued treatment LATE. We extended the delta method and established the asymptotic distribution of the test statistic. We then constructed the critical value for this asymptotic distribution using a modified bootstrap method and showed that the test is asymptotically consistent. The size of the test can be promoted to the nominal significance level over much of the null, indicating a good power property. We also showed that with a minor modification the proposed test can easily be applied when there are conditioning

covariates with finitely many possible values.



(a) DGP 1 **(b)** DGP 2

(c) DGP 3 **(d)** DGP 4

Figure 1.2. Graphs of $p(y,1)$ and $q(y,1)$ under H_1.

Chapter 1, in part is currently being prepared for submission for publication of the material. Sun, Zhenting. The dissertation author was the primary investigator and author of this material.

Chapter 2

Improved Nonparametric Bootstrap Tests of Lorenz Dominance

Abstract

One income or wealth distribution is said to Lorenz dominate another when the Lorenz curve for the former distribution is nowhere below that of the latter, indicating a (weakly) more equitable allocation of resources. Existing tests of the null of Lorenz dominance based on pairs of samples of income or wealth achieve the nominal rejection rate asymptotically when the two Lorenz curves are equal, but are conservative at other points in the null. We propose new nonparametric bootstrap tests of Lorenz dominance based on preliminary estimation of a contact set. Our tests achieve the nominal rejection rate asymptotically on the boundary of the null; that is, when Lorenz dominance is satisfied, and the Lorenz curves coincide on some interval. Numerical simulations indicate that our tests enjoy substantially improved power compared to existing procedures at relevant sample sizes.

2.1 Introduction

Lorenz curves are widely used for the analysis of economic inequality. A Lorenz curve is a function of the distribution of wealth (or income) across a population, which graphs the cumulative proportion of total wealth by cumulative proportion of the population ordered

from poorest to richest. In practice, people are interested in comparing the Lorenz curves between different populations. If one Lorenz curve is below another one, the wealth in the former population is more unequally distributed toward the rich. We use the concept of Lorenz dominance to formalize the comparison of two Lorenz curves: distribution A Lorenz dominates distribution B if the Lorenz curve for A is nowhere below that for B. So if distribution A Lorenz dominates distribution B, then the allocation of resources is more equitable in distribution A than in distribution B.

In Figure 2.1, the vertical axis measures the cumulative share of wealth owned and the horizontal axis measures the cumulative share of people ordered from lowest to highest income. Distribution A Lorenz dominates distribution B, hence distribution A exhibits more economic equality than distribution B. The line of equality (45 degree line) is the Lorenz curve representing perfect equality, i.e. wealth is uniformly distributed.

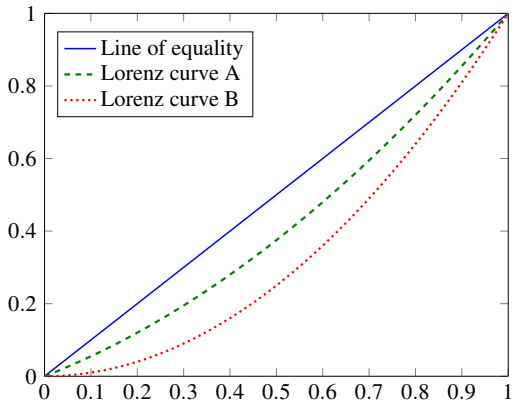


Figure 2.1. Lorenz Curves and Lorenz Dominance.

Because of the economic importance of Lorenz dominance, methods of statistically testing for Lorenz dominance are of interest. Bishop *et al.* (1991a) and Bishop *et al.* (1991b) employed pair-wise multiple comparisons of sample Lorenz ordinates to test for differences between Lorenz curves and then determine Lorenz dominance. Dardanoni & Forcina (1999) and Davidson & Duclos (2000) proposed tests of Lorenz dominance at a chosen set of points. Barrett *et al.* (2014) pointed out that these tests are potentially inconsistent because they limit attention

to a small fixed set of grid points. They proposed a new class of consistent nonparametric tests for testing the null hypothesis of Lorenz dominance, which are analogous to tests of stochastic dominance proposed by McFadden (1989) and elaborated and extended by Barrett & Donald (2003). The tests are constructed based on a general functional $\mathscr{F}$ applied to $\hat{\phi}$, a function on $[0,1]$ equal to the difference between two empirical Lorenz curves. Two specific functionals used to construct test statistics are $\mathscr{S}$, which computes the supremum of $\hat{\phi}$, and $\mathscr{I}$, which computes the integral of $\hat{\phi}$ over the region where $\hat{\phi}$ is positive. The $\mathscr{I}$-based test statistic was first proposed by Bhattacharya (2007).

A pair of distributions satisfying the null of Lorenz dominance is said to be on the boundary of the null whenever the corresponding Lorenz curves coincide over some interval. To obtain critical values, Barrett *et al.* (2014) employ a bootstrap procedure that leads to a test with limiting rejection rate equal to the nominal level when the two Lorenz curves are equal, and below the nominal level elsewhere in the null. If we are at a point on the boundary of the null where the Lorenz curves are not equal, then their test has limiting rejection rate below the nominal level, and thus lacks power against nearby points in the alternative. Our main contribution is an alternative construction of bootstrap critical values for the test statistics of Barrett *et al.* (2014) that achieves a limiting rejection rate equal to the nominal level over the boundary of the null, thereby improving power. Numerical simulations indicate that the improvement to power can be large.

The primary technical obstacle to obtaining a valid bootstrap approximation over the boundary of the null is that the functional $\mathscr{F}$ typically fails to be Hadamard differentiable in this region, which is known to imply inconsistency of standard bootstrap approximations (Dümbgen, 1993). By applying recent results of Fang & Santos (2014) on bootstrap inference under nondifferentiability, we show that a modified bootstrap procedure based on preliminary estimation of a contact set can deliver consistent approximation over the boundary of the null. Our power-improving modification to the tests of Lorenz dominance proposed by Barrett *et al.* (2014) can be viewed as analogous to the modifications made by Linton *et al.* (2010) and Donald

& Hsu (2016) to the tests of stochastic dominance proposed by Barrett & Donald (2003), or to the modification made by Beare & Shi (2018) to the tests of density ratio ordering proposed by Carolan & Tebbs (2005) and Beare & Moon (2015), or to the modifications made by Seo (Forthcoming) to the tests of stochastic monotonicity and conditional stochastic dominance proposed by Delgado & Escanciano (2012, 2013).

Our asymptotic results exploit important recent work by Kaji (2017), who has established weak convergence of the empirical quantile process and bootstrap empirical quantile process in the L_1-semimetric under mild technical conditions. Such convergence implies weak convergence of the empirical Lorenz process and bootstrap empirical Lorenz process in the uniform metric under the same conditions, greatly facilitating our analysis.

In this paper, given a set A, we let $\ell^\infty(A)$ denote the Banach space of bounded real functions on A equipped with the uniform norm $\|\cdot\|_\infty$. When A is a metric space, we let $C(A)$ denote the subspace of $\ell^\infty(A)$ consisting of continuous functions. If $h \in C(A)$, we say that h vanishes at infinity if for every $\varepsilon > 0$ the set $\{a \in A : |h(x)| \geq \varepsilon\}$ is compact, and we define

$$C_0(A) = \{h \in C(A) : h \text{ vanishes at infinity}\}.$$

We let $\rightsquigarrow$ denote weak convergence in a metric space in the sense of Hoffman-Jørgensen.

2.2 Hypothesis Tests of Lorenz Dominance

Suppose that F_1 and F_2 are the cumulative distribution functions (CDFs) of income in two populations. Let $\mathbb{L}$ be the space of Lebesgue measurable functions $h : [0,\infty) \to \mathbb{R}$ with limit $h(\infty) = \lim_{x\to\infty} h(x) \in \mathbb{R}$ and $\|h\|_{\mathbb{L}} < \infty$, where

$$\|h\|_{\mathbb{L}} = \max\{\|h\|_\infty, \|h - h(\infty)\|_1\}.$$

Let $\mathbb{L}_F \subset \mathbb{L}$ be the set of distribution functions that are monotone and cadlag with $h(0) = 0$ and $h(\infty) = 1$. We impose the following regularity conditions on F_1 and F_2.

Assumption 2.2.1 *For $j = 1, 2$, $F_j \in \mathbb{L}_F$ is continuously differentiable with strictly positive derivative f_j. Also, F_j has a $(2+\varepsilon)$th moment for some $\varepsilon > 0$.*

Assumption 2.2.1 is a low level condition and guarantees that we can obtain the differentiability of generalized inverse transformations by using the results in Kaji (2017). More details will be discussed later.

We denote the quantile function by the generalized inverse transformation of a CDF, i.e. for any CDF F with support $[0, \infty)$, the quantile function is

$$Q(p) = \mathscr{V}(F)(p) = \inf\{x \in [0,\infty) : F(x) \geq p\}, \tag{2.1}$$

for all $p \in [0, 1]$, where $\mathscr{V}$ denotes the generalized inverse map.

Definition 2.2.1 *With the existence of nonzero first moment of continuously differentiable distribution F_j for $j = 1, 2$, the Lorenz curve (LC) for the respective population is,*

$$L_j(p) = \frac{\int_0^{Q_j(p)} x f_j(x) \mathrm{d}x}{\int_0^{\infty} x f_j(x) \mathrm{d}x} = \frac{\int_0^p Q_j(t) \mathrm{d}t}{\mu_j}, \tag{2.2}$$

where Q_j is the quantile function of F_j, and μ_j is the mean of the distribution.

One implication of Assumption 2.2.1 is that the quantile function can be defined as $Q_j(p) = F_j^{-1}(p), (0 \leq p \leq 1)$, and it is also continuously differentiable within $(0, 1)$. Here F_j^{-1} is the standard inverse function of F_j, which is well defined because F_j is strictly increasing under Assumption 2.2.1.

Definition 2.2.2 *Given two distributions F_1 and F_2, we say that F_1 weakly Lorenz dominates F_2 if the Lorenz curve L_1 for F_1 is nowhere below L_2 for F_2, i.e. $L_1(p) - L_2(p) \geq 0$ for all $p \in [0, 1]$.*

Notice that by Definition 2.2.1 the LC is a special type of distribution function with support $[0,1]$. In this sense, Definition 2.2.2 is similar to the first order stochastic dominance for associated Lorenz curves, while Lorenz dominance has an economic background. If one Lorenz curve is nowhere below another, the former one implies less inequality in economy for the population. For example, in Figure 2.1, distribution A weakly Lorenz dominates B by Definition 2.2.2, which indicates A is more economically equal than B.

2.2.1 Hypothesis Formulation

Under Assumption 2.2.1, Definition 2.2.1 and 2.2.2, Barrett *et al.* (2014) proposed consistent nonparametric tests of Lorenz dominance. This paper follows the basic setup of Barrett *et al.* (2014).

The hypothesis of interest in this paper is

$$H_0 : L_2(p) \leq L_1(p) \text{ for all } p \in [0,1],$$
$$H_1 : L_2(p) > L_1(p) \text{ for some } p \in [0,1].$$

The null hypothesis H_0 is satisfied when F_1 weakly Lorenz dominates F_2, while the alternative hypothesis H_1 is satisfied when such dominance does not occur.

We define the point-wise difference between the two Lorenz curves by

$$\phi(p) = L_2(p) - L_1(p) \text{ for all } p \in [0,1]. \tag{2.3}$$

And by Definition 2.2.1, under Assumption 2.2.1, $\phi \in C[0,1]$.

To test the hypothesis, we consider functionals which transform ϕ into a scalar value. Suppose there exists a functional $\mathscr{F} : C[0,1] \to \mathbb{R}$. We introduce necessary assumptions on $\mathscr{F}$ to establish the testing theory.

Assumption 2.2.2 *Properties of functional $\mathscr{F}$: For any $h \in C[0,1]$,*

(i) if $h(p) \leq 0$ *and* $h(p) = 0$ *for some* $p \in [0,1]$*, then* $\mathscr{F}(h) = 0$*;*

(ii) if $h(p) > 0$ *for some* $p \in (0,1)$*, then* $\mathscr{F}(h) > 0$*.*

Because by Definition 2.2.1 $\phi(0) = \phi(1) = 0$, under Assumptions 2.2.2(i) and (ii), H_0 (H_1) is equivalent to $\mathscr{F}(\phi) = 0$ ($\mathscr{F}(\phi) > 0$).

Two specific examples of $\mathscr{F}$ we will mainly focus on in this paper are

$$\mathscr{S}(\phi) = \sup_{p \in [0,1]} \phi(p), \tag{2.4}$$

and

$$\mathscr{I}(\phi) = \int_0^1 \phi(p) 1\{\phi(p) > 0\} \mathrm{d}p, \tag{2.5}$$

which can be proved to both satisfy Assumption 2.2.2.

The assumptions about sample data are given below.

Assumption 2.2.3 *For* $j = 1,2$*,* $\{X_i^j\}_{i=1}^{n_j}$ *is an independent and identically distributed (iid) collection of random variables drawn from* F_j*. And* $\{X_i^1\}_{i=1}^{n_1}$ *is independent from* $\{X_i^2\}_{i=1}^{n_2}$*.*

Formally, we treat the first sample size n_1 as a function of the second sample size n_2, such that $n_1 = n_1(n_2) \to \infty$ as $n_2 \to \infty$. We suppose further that the sample sizes n_1, n_2 satisfy:

$$\lim_{n_2 \to \infty} \frac{n_1}{n_1 + n_2} = \lambda \in (0,1). \tag{2.6}$$

Basically, (2.6) requires that the sample sizes n_1 and n_2 grow at comparable rates which can be extended in certain cases. We let $n = n_1 + n_2$. Then $n_2 \to \infty$ is equivalent to $n \to \infty$ under (2.6). We define $T_n = n_1 n_2 / (n_1 + n_2)$. With Assumption 3.3.1 for sample data, the empirical notations are defined as below.

Definition 2.2.3 *With sample data* $\{X_i^j\}_{i=1}^{n_j}$*, for* $j = 1,2$*, define*

(i) Empirical CDF: $\hat{F}_j(z) = n_j^{-1} \sum_{i=1}^{n_j} 1\{X_i^j \le z\}$;

(ii) Empirical quantile: $\hat{Q}_j(p) = \inf\{z \in [0,\infty) : \hat{F}_j(z) \ge p\}$ *for some* $p \in [0,1]$ *for* $z \in [0,\infty)$;

(iii) Empirical LC: $\hat{L}_j(p) = \hat{\mu}_j^{-1} \int_0^p \hat{Q}_j(t)\mathrm{d}t$ *for all* $p \in [0,1]$, *where* $\hat{\mu}_j$ *is the sample mean for sample* j.

Given sample data $\{X_i^j\}_{i=1}^{n_j}$, for $j = 1,2$, we can order $\{X_i^j\}$ from smallest to largest by $X_{(1)}^j \le X_{(2)}^j \cdots \le X_{(n_j)}^j$. By Definition 2.2.3(i), (ii), $\hat{Q}_j(p) = X_{(i)}^j$, for $p \in ((i-1)/n_j, i/n_j]$. Then the empirical Lorenz curve can be calculated exactly by plugging $\hat{Q}_j$ into $\hat{L}_j$. For $p \le n_j^{-1}$, we have $\hat{L}_j(p) = \hat{\mu_j}^{-1} p X_{(1)}^j$. For any $p \in ((i-1)/n_j, i/n_j]$ with $i = 2,3,\cdots,n_j$,

$$\hat{L}_j(p) = \frac{\frac{1}{n_j}\sum_{k=1}^{i-1} X_{(k)}^j + (p - \frac{i-1}{n_j}) X_{(i)}^j}{\hat{\mu}_j}. \tag{2.7}$$

Now we introduce the weak convergence of the empirical Lorenz process $\sqrt{n_j}(\hat{L}_j - L_j)$ for $j = 1,2$. Similar results can be found in Goldie (1977) and Chernozhukov *et al.* (2010). We obtain the asymptotic distribution by first deriving the Hadamard differentiability of Lorenz curves with respect to the corresponding CDFs. This differentiability will also be helpful when we apply the bootstrap method later. To this point, we first introduce the concept of Hadamard differentiability.

Definition 2.2.4 *Let* $\mathbb{D}$ *and* $\mathbb{E}$ *be normed spaces, and* $\mathscr{F} : \mathbb{D}_{\mathscr{F}} \subset \mathbb{D} \mapsto \mathbb{E}$. *A map* $\mathscr{F}$ *is said to be Hadamard differentiable at* $\phi \in \mathbb{D}_{\mathscr{F}}$ *tangentially to a set* $\mathbb{D}_0 \subset \mathbb{D}$, *if there is a continuous linear map* $\mathscr{F}'_\phi : \mathbb{D}_0 \to \mathbb{E}$ *s.t.*

$$\lim_{n\to\infty} \left\| \frac{\mathscr{F}(\phi + t_n h_n) - \mathscr{F}(\phi)}{t_n} - \mathscr{F}'_\phi(h) \right\|_{\mathbb{E}} = 0, \tag{2.8}$$

for all sequences $\{h_n\} \subset \mathbb{D}$ *and* $\{t_n\} \in \mathbb{R}$ *s.t.* $t_n \to 0$, $h_n \to h \in \mathbb{D}_0$ *as* $n \to \infty$ *and* $\phi + t_n h_n \in \mathbb{D}_{\mathscr{F}}$ *for all n.*

Hadamard differentiability is an important property when we use the delta method to derive the asymptotic distribution of Lorenz curves. The next lemma together show the Hadamard differentiability of Lorenz curve defined by 2.2.1 with respect to the corresponding CDFs.

Let $\mathscr{Z} : \mathbb{L}_F \to \ell^\infty[0,1]$ be such that

$$\mathscr{Z}(F)(p) = \frac{\int_0^p \mathscr{V}(F)(t)\mathrm{d}t}{\int_0^1 \mathscr{V}(F)(t)\mathrm{d}t}, \tag{2.9}$$

where $\mathscr{V}$ is the quantile map. Under this map,

$$L_j = \mathscr{Z}(F_j), \quad \hat{L}_j = \mathscr{Z}(\hat{F}_j). \tag{2.10}$$

Lemma 2.2.1 shows that the Lorenz curve is Hadamard differentiable with respect to F at F_j.

Lemma 2.2.1 *Under Assumption 2.2.1, $\mathscr{Z} : \mathbb{L}_F \to \ell^\infty[0,1]$ is Hadamard differentiable at F_j tangentially to $C_0[0,\infty) \cap \mathbb{L}$ with derivative*

$$\mathscr{Z}'_{F_j}(h)(p) = \frac{\int_0^p F_j^{-1}(t)\mathrm{d}t \int_0^1 \frac{h(F_j^{-1}(t))}{f_j(F_j^{-1}(t))}\mathrm{d}t - \int_0^p \frac{h(F_j^{-1}(t))}{f_j(F_j^{-1}(t))}\mathrm{d}t \int_0^1 F_j^{-1}(t)\mathrm{d}t}{(\int_0^1 F_j^{-1}(t)\mathrm{d}t)^2}, \tag{2.11}$$

for all $h \in C_0[0,\infty) \cap \mathbb{L}$.

With the Hadamard differentiability of Lorenz curves, we obtain the asymptotic distribution of $\sqrt{n_j}(\hat{L}_j - L_j)$ by first applying the weak convergence of $\sqrt{n_j}(\hat{F}_j - F_j)$ in $\mathbb{L}$ in Kaji (2017). And then we can show the asymptotic distribution of $\sqrt{T_n}(\hat{\phi} - \phi)$ based on Assumption 3.3.1.

Lemma 2.2.2 *Under Assumptions 2.2.1 and 3.3.1,*

$$\sqrt{n_j}(\hat{L}_j - L_j) \rightsquigarrow \mathscr{L}_j, \tag{2.12}$$

as $n_j \to \infty$ for $j = 1,2$, where $\mathscr{L}_j$ is a Gaussian process with continuous sample paths. Moreover,

as $n \to \infty$, we have

$$\sqrt{T_n}(\hat{\phi} - \phi) = \sqrt{T_n}(\hat{L}_2 - L_2) - \sqrt{T_n}(\hat{L}_1 - L_1) \rightsquigarrow \sqrt{\lambda}\mathscr{L}_2 - \sqrt{1-\lambda}\mathscr{L}_1. \tag{2.13}$$

We next derive the asymptotic distribution of $\sqrt{T_n}(\mathscr{F}(\hat{\phi}) - \mathscr{F}(\phi))$ to establish the testing theory. If the map $\mathscr{F}$ is Hadamard differentiable, we can easily obtain the limit distribution of $\sqrt{T_n}(\mathscr{F}(\hat{\phi}) - \mathscr{F}(\phi))$ by the Delta method with the weak convergence result in (2.13), and then approximate that limit distribution using the bootstrap law of $\mathscr{F}'_\phi(\sqrt{T_n}(\hat{\phi}^* - \hat{\phi}))$, where $\mathscr{F}'_\phi$ is the Hadamard derivative of $\mathscr{F}$ at ϕ and $\hat{\phi}^*$ is a bootstrap version of $\hat{\phi}$. However, in applications $\mathscr{F}$ is not always Hadamard differentiable with respect to ϕ, and as a result the bootstrap method may not work. For the two specific non-Hadamard differentiable functionals (2.4) and (2.5), Barrett *et al.* (2014) suggest instead to bootstrap the limit distribution of $\sqrt{T_n}\mathscr{F}(\hat{\phi} - \phi)$ such that under H_0:

$$\mathscr{F}(\sqrt{T_n}(\hat{\phi} - \phi)) = \sqrt{T_n}\mathscr{F}(\hat{\phi} - \phi) \geq \sqrt{T_n}\mathscr{F}(\hat{\phi}) \text{ a.s.} \tag{2.14}$$

And Barrett *et al.* (2014) bootstrap the p-value using

$$\hat{p}(\mathscr{F}) = \frac{1}{J}\sum_{t=1}^{J} 1\{\sqrt{T_n}\mathscr{F}(\hat{\phi}_j^* - \hat{\phi}) > \sqrt{T_n}\mathscr{F}(\hat{\phi})\}, \tag{2.15}$$

where $\hat{\phi}_j^*$ is the jth bootstrap version of $\hat{\phi}$ obtained from bootstrap sample, and J is the total number of bootstrap samples. With (2.15), the decision rule of the test in Barrett *et al.* (2014) is

$$\text{reject } H_0 \text{ if } \hat{p}(\mathscr{F}) < \alpha, \tag{2.16}$$

where α is the nominal significance level. Under this setting, Barrett *et al.* (2014) prove that the limit rejection rate is less than or equal to α under H_0 and converges to 1 under H_1. The reason why the limit rejection rate could be below α is that under H_0 we have $\mathscr{F}(\phi) = 0$, and so

$\sqrt{T_n}\mathscr{F}(\hat{\phi}-\phi) \geq \sqrt{T_n}(\mathscr{F}(\hat{\phi})-\mathscr{F}(\phi))$ a.s. as is shown in (2.14). And with a finite sample we can not rule out the strict inequality if the two Lorenz curves are not identical. The estimated p value is obtained from bootstrapping the distribution of an upper bound of the test statistic and therefore it is larger than that from bootstrapping the distribution of the test statistic itself. If we can find an effective way to derive and bootstrap the asymptotic distribution of $\sqrt{T_n}(\mathscr{F}(\hat{\phi})-\mathscr{F}(\phi))$, we may be able to construct a test with a superior power.

The bootstrap method proposed by Fang & Santos (2014) supports this idea. To proceed to illustrating how to construct the test, we now introduce the concept of Hadamard directional differentiability.

Definition 2.2.5 *Let $\mathbb{D}$ and $\mathbb{E}$ be normed spaces, and $\mathscr{F} : \mathbb{D}_{\mathscr{F}} \subset \mathbb{D} \mapsto \mathbb{E}$. A map $\mathscr{F}$ is said to be Hadamard directionally differentiable at $\phi \in \mathbb{D}_{\mathscr{F}}$ tangentially to a set $\mathbb{D}_0 \subset \mathbb{D}$, if there is a continuous map $\mathscr{F}'_\phi : \mathbb{D}_0 \to \mathbb{E}$ s.t.*

$$\lim_{n\to\infty} \left\| \frac{\mathscr{F}(\phi + t_n h_n) - \mathscr{F}(\phi)}{t_n} - \mathscr{F}'_\phi(h) \right\|_{\mathbb{E}} = 0, \tag{2.17}$$

for all sequences $\{h_n\} \subset \mathbb{D}$ and $\{t_n\} \in \mathbb{R}_+$ s.t. $t_n \downarrow 0$, $h_n \to h \in \mathbb{D}_0$ as $n \to \infty$ and $\phi + t_n h_n \in \mathbb{D}_{\mathscr{F}}$ for all n.

As is mentioned in Fang & Santos (2014), there are two differences between Hadamard differentiability and Hadamard directional differentiability. One is that in Definition 2.2.5, t_n must approach 0 from above. The other one is the map $\mathscr{F}'_\phi$ is not necessarily linear in (2.17). And it is because of the second difference, we need to find a new way to applying the bootstrap approach.

Hadamard Directional Derivatives for the Two Specific Functionals

In this part, the notations in Definition 2.2.5 are specified as $\mathbb{D} = \mathbb{D}_{\mathscr{F}} = \ell^\infty[0,1]$, $\mathbb{D}_0 = C[0,1]$ with $\|\cdot\|_{\mathbb{D}} = \|\cdot\|_\infty$, and $\mathbb{E} = \mathbb{R}$ with $\|\cdot\|_{\mathbb{E}} = |\cdot|$.

For $\mathscr{S}(\psi) = \sup_{p\in[0,1]} \psi(p)$, by Example 3 of Fang & Santos (2014), $\mathscr{S}$ is Hadamard directionally differentiable at ϕ tangentially to $C[0,1]$, and the directional derivative is

$$\mathscr{S}'_\phi(h) = \sup_{p\in\Psi_{[0,1]}(\phi)} h(p), \tag{2.18}$$

where $\Psi_{[0,1]}(\phi) = \arg\max_{p\in[0,1]} \phi(p)$. Then by Theorem 2.1 of Fang & Santos (2014) and (2.13),

$$\sqrt{T_n}(\mathscr{S}(\hat{\phi}) - \mathscr{S}(\phi)) \rightsquigarrow \mathscr{S}'_\phi(\sqrt{\lambda}\mathscr{L}_2 - \sqrt{1-\lambda}\mathscr{L}_1). \tag{2.19}$$

For $\mathscr{I}(\psi) = \int_0^1 \psi(p) 1\{\psi(p) > 0\} \mathrm{d}p$, by Example 5 of Fang & Santos (2014), $\mathscr{I}$ is Hadamard directionally differentiable at ϕ. Define

$$B_0(\phi) = \{p \in [0,1] : \phi(p) = 0\}$$

and

$$B_+(\phi) = \{p \in [0,1] : \phi(p) > 0\}.$$

The Hadamard directional derivative of $\mathscr{I}$ at ϕ is

$$\mathscr{I}'_\phi(h) = \int_{B_+(\phi)} h(p)\mathrm{d}p + \int_{B_0(\phi)} \max\{h(p), 0\}\mathrm{d}p. \tag{2.20}$$

By Theorem 2.1 of Fang & Santos (2014) and (2.13),

$$\sqrt{T_n}(\mathscr{I}(\hat{\phi}) - \mathscr{I}(\phi)) \rightsquigarrow \mathscr{I}'_\phi(\sqrt{\lambda}\mathscr{L}_2 - \sqrt{1-\lambda}\mathscr{L}_1). \tag{2.21}$$

Our next assumption, which imposes Hadamard directional differentiability upon the

general functional $\mathscr{F}$, is automatically satisfied when $\mathscr{F} = \mathscr{S}$ or $\mathscr{F} = \mathscr{I}$.

Assumption 2.2.4 *Functional $\mathscr{F} : \ell^\infty[0,1] \mapsto \mathbb{R}$ is Hadamard directionally differentiable at $\phi \in C[0,1]$ tangentially to the set $C[0,1]$, where the Hadamard directional derivative $\mathscr{F}'_\phi$ satisfies* (2.17).

2.2.2 Bootstrap

As shown in Fang & Santos (2014), the asymptotic distribution of $\sqrt{T_n}(\mathscr{F}(\hat{\phi}) - \mathscr{F}(\phi))$ can be obtained if $\mathscr{F}$ is Hadamard directionally differentiable, but this distribution could be nonstandard. We would need to approximate it using a bootstrap procedure of some sort. Let $\hat{\phi}^*$ denote a "bootstrapped version" of $\hat{\phi}$ which is defined as a function mapping the data $\{X_i^j\}_{i=1}^{n_j}$ for $j = 1,2$ and random weights $\{W_i^j\}_{i=1}^{n_j}$ that are independent of $\{X_i^j\}_{i=1}^{n_j}$ into $\mathbb{D}_{\mathscr{F}}$. This general bootstrap definition covers a large family of resampling schemes such as Bayesian, block, score, and weighted bootstraps. Further discussion can be found in Remark 3.2 in Fang & Santos (2014).

Specifically, in this paper we construct $\hat{\phi}^*$ in the following way:

1. Obtain the bootstrap sample $\{X_i^{j*}\}_{i=1}^{n_j}$ for $j = 1,2$ with replacement independently from $\{X_i^j\}_{i=1}^{n_j}$ for $j = 1,2$.

2. Calculate the following bootstrap objects:

   (i) Bootstrap CDF: $\hat{F}_j^*(x) = \frac{1}{n_j}\sum_{i=1}^{n_j} 1\{X_i^{j*} \le x\}$ for $x \in [0,\infty)$;

   (ii) Bootstrap quantile: $\hat{Q}_j^*(p) = \inf\{x : \hat{F}_j^*(x) \ge p\}$ for $p \in [0,1]$;

   (iii) Bootstrap LC: $\hat{L}_j^*(p) = \hat{\mu}_j^{*-1}\int_0^p \hat{Q}_j^*(t)\mathrm{d}t$ for $p \in [0,1]$, where $\hat{\mu}_j^*$ is the sample mean for bootstrap sample j.

3. Let $\hat{\phi}^* = \hat{L}_2^* - \hat{L}_1^*$.

Now we want to show the asymptotic distribution of $\sqrt{T_n}(\hat{\phi}^* - \hat{\phi})$. Notice that

$$\hat{L}_j^* = \mathscr{Z}(\hat{F}_j^*). \tag{2.22}$$

Also, notice that the bootstrap CDF can be written as

$$\hat{F}_j^* = \frac{1}{n_j} \sum_{i=1}^{n_j} W_i^j 1_{[X_i^j, \infty)}. \tag{2.23}$$

Lemma 2.2.3 *Under Assumptions 2.2.1 and 2.2.1,*

$$\sqrt{T_n}(\hat{\phi}^* - \hat{\phi}) \rightsquigarrow \sqrt{\lambda}\mathscr{L}_2 - \sqrt{1-\lambda}\mathscr{L}_1, \tag{2.24}$$

for almost every sequence $\{X_i^j\}_{i=1}^{n_j}$ *with* $j = 1,2$*. Also, it holds that*

(i) $\sqrt{T_n}\{\hat{\phi}^* - \hat{\phi}\}$ *is asymptotically measurable (jointly in* $\{X_i^j, W_i^j\}_{i=1}^{n_j}$ *with* $j = 1,2$*),*

(ii) $h(\sqrt{T_n}\{\hat{\phi}^* - \hat{\phi}\})$ *is a measurable function of* $\{W_i^j\}_{i=1}^{n_j}, j = 1,2$*, outer almost surely in* $\{X_i^j\}_{i=1}^{n_j}, j = 1,2$*, for any continuous and bounded* $h : \ell^\infty[0,1] \to \mathbb{R}$*.*

A natural approximation to the limiting distribution of $\sqrt{T_n}(\mathscr{F}(\hat{\phi}) - \mathscr{F}(\phi))$ is given by the bootstrap law of $\mathscr{F}'_\phi(\sqrt{T_n}(\hat{\phi}^* - \hat{\phi}))$. However, the exact form of $\mathscr{F}'_\phi$ is unknown because ϕ is unknown. We will approximate $\mathscr{F}'_\phi$ using an estimator $\hat{\mathscr{F}}'_\phi$ satisfying the following high level condition taken from Fang & Santos (2014, Ass. 3.3).

Assumption 2.2.5 $\hat{\mathscr{F}}'_\phi : \ell^\infty[0,1] \mapsto \mathbb{R}$ *is a function of* $\{X_i\}_{i=1}^n$ *satisfying for every compact set* $K \subset C[0,1]$*,* $K^\delta \equiv \{a \in \ell^\infty[0,1] : \inf_{b \in K} \|a - b\|_\infty < \delta\}$*, and every* $\varepsilon > 0$*, the property:*

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} P\Big(\sup_{h \in K^\delta} |\hat{\mathscr{F}}'_\phi(h) - \mathscr{F}'_\phi(h)| > \varepsilon \Big) = 0. \tag{2.25}$$

We show in the following subsection that the constructed $\hat{\mathscr{S}}'_\phi$ and $\hat{\mathscr{I}}'_\phi$ satisfy Assumption 2.2.5 when $\mathscr{F} = \mathscr{S}$ and $\mathscr{F} = \mathscr{I}$ respectively.

Consistency of the Estimators for the Hadamard Directional Derivatives of the Two Specific Functionals

As discussed in Example 2.3 in Fang & Santos (2014), a natural estimator for $\mathscr{S}'_\phi$ is given by

$$\hat{\mathscr{S}}'_n(h) = \sup_{p \in \hat{\Psi}_{[0,1]}(\phi)} h(p) \tag{2.26}$$

for $h \in \ell^\infty[0,1]$, where $\hat{\Psi}_{[0,1]}(\phi)$ is an estimator for the set $\Psi_{[0,1]}(\phi)$ in (2.18). To obtain such an estimator we construct

$$\hat{B}_n = \left\{ p \in [0,1], |\hat{\phi}(p)| \leq \tau_n \right\}, \tag{2.27}$$

where τ_n is a tuning parameter satisfying $\tau_n \to 0$ and $\sqrt{T_n}\tau_n \to \infty$ as $n \to \infty$. We will see that $\hat{B}_n$ provides a consistent estimate of

$$B(\phi) = \{ p \in [0,1], \phi(p) = 0 \}. \tag{2.28}$$

The set $B(\phi)$ is called the contact set of L_1 and L_2, and plays a similar role to the contact set in Linton *et al.* (2010).

Lemma 2.2.4 *Under H_0, $\Psi_{[0,1]}(\phi) = B(\phi)$ and $\hat{B}_n$ is a Hausdorff consistent estimator of $B(\phi)$, i.e. $d_H(B(\phi), \hat{B}_n) = o_P(1)$, where d_H is the Hausdorff metric.*

If we set $\hat{\Psi}_{[0,1]} = \hat{B}_n$ in (2.26) then Lemma B.3 in Fang & Santos (2014) implies that, when H_0 is satisfied, the resulting estimator $\hat{\mathscr{S}}'_n$ satisfies Assumption 2.2.5.

When $\mathscr{F} = \mathscr{I}$, we see from Example 2.5 in Fang & Santos (2014) that a natural estimator for $\mathscr{I}'_\phi$ is given by

$$\hat{\mathscr{I}}'_n(h) = \int_{\hat{B}_{+n}} h(p)\mathrm{d}p + \int_{\hat{B}_{0n}} \max(h(p), 0)\mathrm{d}p \tag{2.29}$$

for $h \in \ell^\infty([0,1])$, where $\hat{B}_{+n}$ and $\hat{B}_{0n}$ are estimators for $B_+(\phi)$ and $B_0(\phi)$. The sets $B(\phi)$ and

$B_0(\phi)$ are the same, but we use the latter notation here to emphasize the connection to $B_+(\phi)$. We estimate them by setting

$$\hat{B}_{+n} = \left\{ p \in [0,1] : \hat{\phi}(p) > \tau_n \right\} \tag{2.30}$$

and

$$\hat{B}_{0n} = \left\{ p \in [0,1] : |\hat{\phi}(p)| \leq \tau_n \right\}, \tag{2.31}$$

where again we require the tuning parameter τ_n to satisfy $\tau_n \to 0$ and $\sqrt{T_n}\tau_n \to \infty$ as $n \to \infty$.

Lemma 2.2.5 *$\mu(\hat{B}_{+n} \triangle B_+(\phi)) \to_p 0$ and $\mu(\hat{B}_{0n} \triangle B_0(\phi)) \to_p 0$, where μ is the Lebesgue measure and $A \triangle B$ denotes the symmetric difference between any sets A and B.*

From this lemma and Lemma B.3 in Fang & Santos (2014) it follows that $\hat{\mathscr{I}}'_n$ satisfies Assumption 2.2.5 above.

Remark 2.2.1 *When the null hypothesis is satisfied, $B_+(\phi) = \varnothing$. Consequently, if in place of* (2.30) *we define $\hat{B}_{+n} = \varnothing$, Lemma 2.2.5 continues to be valid under the null. In the simulations reported in Section 2.3 we define $\hat{B}_{+n}$ as in* (2.30)

Bootstrap-based Inference

Our bootstrap critical value $\hat{c}_{1-\alpha}$ is the $(1-\alpha)$-quantile of the bootstrap law of

$$\hat{\mathscr{F}}'_{\phi}(\sqrt{T}(\hat{\phi}^* - \hat{\phi})).$$

That is,

$$\hat{c}_{1-\alpha} = \inf\left\{ c : P\left(\hat{\mathscr{F}}'_{\phi}(\sqrt{T_n}(\hat{\phi}^* - \hat{\phi})) \leq c | \left\{X_i^1\right\}_{i=1}^{n_1}, \left\{X_i^2\right\}_{i=1}^{n_2} \right) \geq 1-\alpha \right\}. \tag{2.32}$$

The decision rule of the test is set to be

$$\text{Reject } H_0 \text{ if } \sqrt{T_n}\mathscr{F}(\hat{\phi}) > \hat{c}_{1-\alpha}. \tag{2.33}$$

Then for a general functional $\mathscr{F}$, we have the following theorem.

Theorem 2.2.1 *For functional $\mathscr{F} : \ell^\infty[0,1] \mapsto \mathbb{R}$, if Assumptions 2.2.1-2.2.5 hold, then under decision rule* (2.33)*,*

(i) if H_0 is true and the CDF of $\mathscr{F}'_\phi(\sqrt{\lambda}\mathscr{L}_2 - \sqrt{1-\lambda}\mathscr{L}_1)$ is continuous and strictly increasing at its $1-\alpha$ quantile $c_{1-\alpha}$, then

$$\hat{c}_{1-\alpha} \to_p c_{1-\alpha} \text{ and } \lim_{n\to\infty} P(\text{reject } H_0) = \alpha;$$

(ii) if H_0 is false, then

$$\lim_{n\to\infty} P(\text{reject } H_0) = 1.$$

For the specific functionals $\mathscr{S}$ and $\mathscr{I}$ we have the following corollary to Theorem 2.2.1.

Corollary 2.2.1 *If $\mathscr{F} = \mathscr{S}$ or $\mathscr{F} = \mathscr{I}$ and we estimate its Hadamard directional derivative as in* (2.26) *or* (2.29) *respectively, and if Assumptions 2.2.1 and 3.3.1 are satisfied, then under decision rule* (2.33)*,*

(i) if H_0 is true and the CDF of $\mathscr{F}'_\phi(\sqrt{\lambda}\mathscr{L}_2 - \sqrt{1-\lambda}\mathscr{L}_1)$ is continuous and strictly increasing at its $1-\alpha$ quantile $c_{1-\alpha}$, then

$$\hat{c}_{1-\alpha} \to_p c_{1-\alpha} \text{ and } \lim_{n\to\infty} P(\text{reject } H_0) = \alpha;$$

(ii) if H_0 is false, then

$$\lim_{n\to\infty} P(\textit{reject } H_0) = 1.$$

Theorem 2.2.1 and Corollary 2.2.1 both require the CDF of $\mathscr{F}'_\phi(\sqrt{\lambda}\mathscr{L}_2 - \sqrt{1-\lambda}\mathscr{L}_1)$ to be strictly increasing at its $1-\alpha$ quantile $c_{1-\alpha}$. In some cases this condition does not hold. For the functional $\mathscr{S}$, when $\Psi_{[0,1]}(\phi) = \{0,1\}$, e.g. L_2 is everywhere strictly below L_1 except at the endpoints 0 and 1, we have $\mathscr{S}'_\phi(\sqrt{\lambda}\mathscr{L}_2 - \sqrt{1-\lambda}\mathscr{L}_1) = 0$ a.s. For the functional $\mathscr{I}$, if $B_0(\phi)$ and $B_+(\phi)$ are Lebesgue measure 0, we have $\mathscr{I}'_\phi(\sqrt{\lambda}\mathscr{L}_2 - \sqrt{1-\lambda}\mathscr{L}_1) = 0$ a.s. In these cases, the test statistic and bootstrapped critical value will converge to zero and it is not clear how the rejection rate will behave asymptotically.

The estimated sets $\hat{B}_n$, $\hat{B}_{0n}$ and $\hat{B}_{+n}$ depend on the selection of the tuning parameter τ_n. If $\tau_n \geq 1$ then $\hat{B}_n$ and $\hat{B}_{0n}$ are equal to $[0,1]$ and $\hat{B}_{+n}$ is empty. In this case our test is the same as the test of Barrett *et al.* (2014). Reducing τ_n causes $\hat{B}_n$ and $\hat{B}_{0n}$ to get smaller and $\hat{B}_{+n}$ to get larger. This can improve power, at the risk of losing control of size if τ_n is chosen too small. A suitable balance needs to be achieved. We provide a simulation-based approach to choosing τ_n in Section 2.3.3.

2.3 Finite Sample Performance

2.3.1 Simulation Design

We ran a number of Monte Carlo simulations to investigate the finite sample size and power of our test and the test of Barrett *et al.* (2014). In each simulation we used sample sizes $n_j = 200$, $j = 1,2$, and nominal significance level $\alpha = 0.05$. We used a range of tuning parameter values for our contact set estimator: 0.01, 0.02, 0.04, 0.06, 0.08, 0.1. We used $R = 10000$ experimental replications, and employed the method of Giacomini *et al.* (2013) to expedite computation.

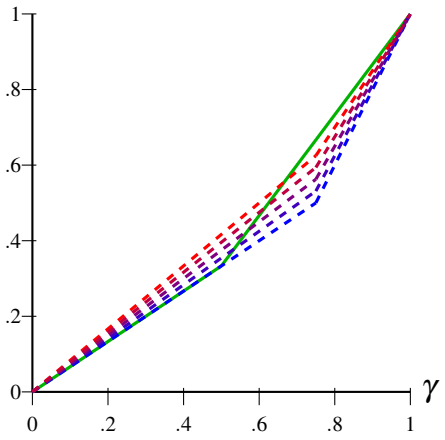


Figure 2.2. Lorenz Curves for X^1 (solid) and $X^2_{(\gamma)}$ (dashed), for Parameter Values $\gamma \in \{0, 0.25, 0.5, 0.75, 1\}$.

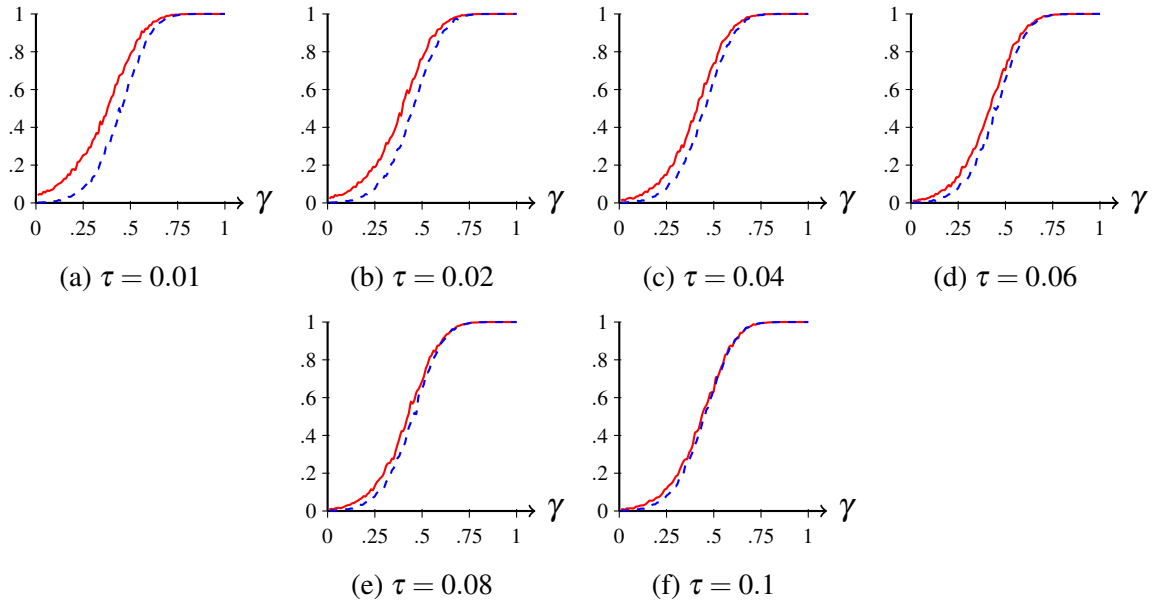
In each simulation the data $\{X^1_i\}_{i=1}^{n_1}$ were generated as independent copies of the random variable

$$X^1 = \begin{cases} 4 & \text{with probability } \frac{1}{2} \\ 8 & \text{with probability } \frac{1}{2}. \end{cases}$$

The data $\{X^2_j\}_{j=1}^{n_2}$ were generated as independent copies of the random variable

$$X^2_{(\gamma)} = \begin{cases} 4+\gamma & \text{with probability } \frac{3}{4} \\ 12-3\gamma & \text{with probability } \frac{1}{4}, \end{cases}$$

whose law is parametrized by $\gamma \in [0,4]$. The Lorenz curves corresponding to X^1 and $X^2_{(\gamma)}$ are displayed in Figure 2.2 for different values of γ. The Lorenz curve for X^1 is drawn with a solid line and has a kink at $p = 0.5$. The Lorenz curves for $X^2_{(\gamma)}$, $\gamma = 0, 0.25, 0.5, 0.75, 1$, are drawn with dashed lines and are kinked at $p = 0.75$. When $\gamma = 0$, the Lorenz curve for $X^2_{(\gamma)}$ is everywhere equal to or less than the Lorenz curve for X^1, so that the null hypothesis of Lorenz curve dominance is satisfied. When $\gamma > 0$ the null hypothesis is violated.



(a) $\tau = 0.01$ (b) $\tau = 0.02$ (c) $\tau = 0.04$ (d) $\tau = 0.06$

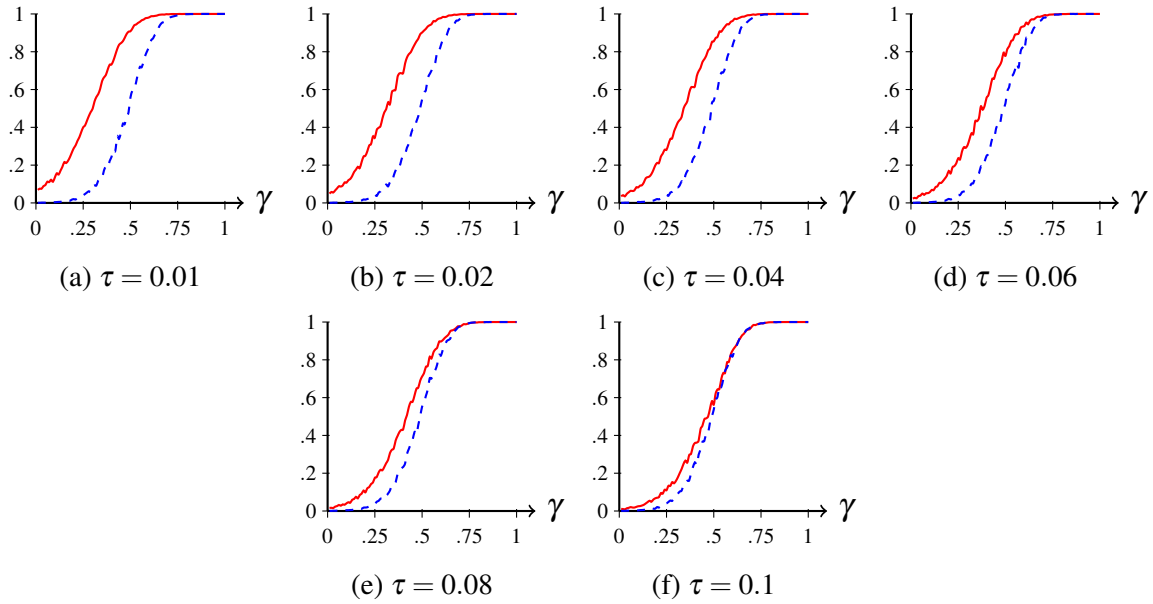
(e) $\tau = 0.08$ (f) $\tau = 0.1$

Figure 2.3. Rejection Rate Comparisons for $\mathscr{F} = \mathscr{S}$.

2.3.2 Simulation Results

Figures 2.3 and 2.4 display the simulated rejection rates for the test introduced in this paper and the test proposed by Barrett *et al.* (2014), for $\mathscr{F}$ equal to $\mathscr{S}$ and $\mathscr{I}$. The red curves represent the simulated rejection rates for our test as γ, which parametrizes the Lorenz curve in Figure 2.2, increases from zero to one. The blue curves represent the simulated rejection rates for the test of Barrett *et al.* (2014). It is apparent that the red curves are in all cases above the corresponding blue curves, reflecting the power improvement obtained using our modified bootstrap procedure. This is a consequence of Lemmas B.1.1 and B.1.2 in Appendix B.1, which assert that $\hat{\mathscr{S}}'_n(h) \leq \mathscr{S}(h)$ and $\hat{\mathscr{I}}'_n(h) \leq \mathscr{I}(h)$ for all $h \in \ell^\infty[0,1]$, implying that the critical value used for our test is equal to or less than the critical value used for the test of Barrett *et al.* (2014).

We also see in Figures 2.3 and 2.4 that as τ increases, the difference between the red and blue curves becomes smaller. This is because the estimates of the sets $\Psi_{[0,1]}(\phi), B_+(\phi), B_0(\phi)$ get larger as τ increases. So $\hat{\mathscr{S}}'_n(h)$ and $\hat{\mathscr{I}}'_n(h)$ get closer to $\mathscr{S}(h)$ and $\mathscr{I}(h)$, respectively.



(a) $\tau = 0.01$ (b) $\tau = 0.02$ (c) $\tau = 0.04$ (d) $\tau = 0.06$

(e) $\tau = 0.08$ (f) $\tau = 0.1$

Figure 2.4. Rejection Rate Comparisons for $\mathscr{F} = \mathscr{I}$.

2.3.3 Tuning Parameter Selection

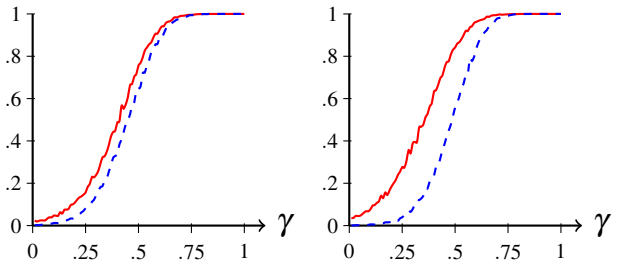
The power of our test increases as we reduce the tuning parameter τ_n, so in practice we would like to choose τ_n to be as small as possible while still controlling size. We suggest the following procedure.

(i) Choose a collection of candidate values for τ_n.

(ii) Resample with replacement from the data $\{X_i^1\}_{i=1}^{n_1}$ to create a bootstrap sample $\{X_i^{2B}\}_{i=1}^{n_2}$. Use the samples $\{X_i^1\}_{i=1}^{n_1}$ and $\{X_i^{2B}\}_{i=1}^{n_2}$ to test the null hypothesis of Lorenz dominance using the test of Barrett *et al.* (2014) and using our proposed test with each candidate value for τ_n. Record the outcome of each of these tests.

(iii) Repeat the previous steps many times. Compute the rejection rates of the different tests.

(iv) Choose the smallest candidate value for τ_n such that the rejection rate computed for our test is within ε of the rejection rate computed for the test of Barrett *et al.* (2014). Here, ε is a small tolerance parameter; we suggest $\varepsilon = 0.001$.



(a) $\mathscr{F} = \mathscr{S}, \tau = 0.03$ (a) $\mathscr{F} = \mathscr{I}, \tau = 0.045$

Figure 2.5. Power Curve Comparisons with Automatically Selected Tuning Parameters.

For extremely large values of τ_n our estimated contact set is $[0,1]$, and so our test is the same as the test of Barrett *et al.* (2014). So by including at least one very large candidate value of τ_n, in step (iv) there should always be at least one rejection rate within ε of the rejection rate for the test of Barrett *et al.* (2014).

The following intuition motivates our tuning parameter selection procedure. We know that when the two distributions F_1 and F_2 are equal, the test of Barrett *et al.* (2014) has limiting rejection rate equal to nominal size. So we would like our tuning parameter to deliver a similar rejection rate. By generating the bootstrap sample $\{X_i^{2B}\}_{i=1}^{n_2}$ by resampling from $\{X_i^1\}_{i=1}^{n_1}$, we approximate the setting where F_1 and F_2 are equal.

We applied our tuning parameter selection procedure to a sample of $n_1 = 200$ independent copies of X^1 as defined in Section 2.3.1, using a nominal size of 0.05. For both functionals $\mathscr{S}$ and $\mathscr{I}$, we picked τ_n from the grid $\tau_n = 0.01 + 0.005 \times k$ with $k = 0, \cdots, 18$. The rejection rate of the test of Barrett *et al.* (2014) was computed to be 0.0103 using $\mathscr{S}$ and 0.0198 using $\mathscr{I}$. The smallest tuning parameter values yielding rejection rates within $\varepsilon = 0.001$ of these rates were $\tau_n = 0.03$ using $\mathscr{S}$ and $\tau_n = 0.045$ using $\mathscr{I}$. We then repeated the simulations described in Sections 2.3.1 and 2.3.2 using these tuning parameter values. Figure 2.5 shows the power curves comparison for the functionals $\mathscr{S}$ and $\mathscr{I}$. We see that our procedure using the automatically selected tuning parameter values generates a large increase in power relative to the test of Barrett *et al.* (2014).

Chapter 2, in part is currently being prepared for submission for publication of the material. Sun, Zhenting; Beare, Brendan K. The dissertation author was the primary investigator and author of this material.

Chapter 3

High-Dimensional Semiparametric Models with Endogeneity

Abstract

When a model includes a large pool of regressors, endogeneity can arise incidentally and cause inconsistency of the estimators from a high-dimensional regression. In this paper, we propose a sieve focused GMM (SFGMM) estimator for general high-dimensional semiparametric conditional moment models in the presence of endogeneity. Under certain conditions, the SFGMM estimator has oracle consistency properties and converges at a desirable rate. We then establish the asymptotic normality of the plug-in SFGMM estimator for possibly irregular functionals. Simulation evidence illustrates the performance of the proposed estimator.

3.1 Introduction

In this paper, we consider high-dimensional semiparametric models in the presence of endogeneity. High dimensions in a variety of nonparametric and semiparametric models have been discussed in the literature, such as in Xie & Huang (2009), Ni *et al.* (2009), Chen *et al.* (2012), Peng & Huang (2011), and Zhu & Zhu (2009). As discussed in Fan & Liao (2014), as more and more explanatory variables are collected, the possibility that some of them end up being correlated with random noise increases. Fan & Liao (2014) propose a focused GMM estimator

that deals with both high dimensions and endogeneity in a general nonlinear parametric model. Many interesting models, such as linear models, logit models, and probit models, are examples of such a model. Under certain conditions, their FGMM estimator can be shown to have oracle properties. The present paper employs their focused GMM approach and constructs an oracle estimator for a general high-dimensional semiparametric model with possible endogeneity.

Nonparametric and semiparametric models with endogeneity but with low dimensions have attracted much attention.[1] When we introduce high dimensions into the model, the usual estimation and inference procedures might not work. We follow the basic setup of semiparametric conditional moment models in the literature and propose a new estimator that deals with high dimensions and endogeneity simultaneously.

Consider a high-dimensional semiparametric model with conditional moment restrictions:

$$E\left[\rho\left(Y, X^{(n)\prime}\theta_0^{(n)}, h_0\left(\delta\left(Y, X^{(n)\prime}\theta_0^{(n)}\right)\right)\right)|W^{(n)}\right] = 0, \tag{3.1}$$

where the dimension of $X^{(n)}$, say p, may increase as the sample size n goes to infinity. The dimension of Y, denoted by d_y, is fixed. We allow $(Y, X^{(n)})$ to include endogenous variables. $W^{(n)}$ is a vector of instrumental variables; ρ and δ are known smooth real-valued functions; and $\theta_0^{(n)}$ and h_0 are the finite-dimensional and infinite-dimensional parameters of interest, respectively. The possibly increasing dimension of $X^{(n)}$ as n goes to infinity captures the feature of high-dimensional models. All the superscripts (n) indicate that the dimensions would be increasing in n.

Model (3.1) is an extension of the classic semiparametric models in Ai & Chen (2003) and Chen & Pouzo (2009). It has been proved in Newey & Powell (2003), Ai & Chen (2003), Chen & Pouzo (2009), and Chen & Pouzo (2012) that the sieve minimum distance (SMD) estimator has good properties in general nonparametric and semiparametric models with regressors that have fixed dimensions. In this paper, we propose an estimator based on the sieve method that has

[1] See, for example, Newey & Powell (2003), Ai & Chen (2003), Chen & Pouzo (2009), and Chen & Pouzo (2012).

desirable oracle properties under additional high-dimensional assumptions. For simplicity, we will assume that there is only one infinite-dimensional parameter, h_0, in the model, and that there are no finite-dimensional parameters of interest other than $\theta_0^{(n)}$. It would be straightforward to include additional such parameters of interest in the model.

We assume that $h_0 \in \mathscr{H}$, where $\mathscr{H}$ is a function space for the infinite-dimensional parameter of interest, and that $\theta_0^{(n)} \in \Theta^{(n)}$, where $\Theta^{(n)} = \Theta^p$, $\Theta \subset \mathbb{R}$, and p is the number of regressors in $X^{(n)}$, which is increasing in n. Assume that for each fixed n, $\Theta^{(n)}$ is compact under the Euclidean norm $\|\cdot\|_E$. Let $\mathscr{A}^{(n)} = \Theta^{(n)} \times \mathscr{H}$, the parameter space for $\alpha^{(n)} = (\theta^{(n)}, h)$. We let $\{\mathscr{H}_k\}_k$ be a sequence of compact subsets of $\mathscr{H}$ under a strong norm $\|\cdot\|_s$ of $\mathscr{H}$ such that $\mathscr{H}_k \subset \mathscr{H}_{k+1}$. Let $\mathscr{A}_k^{(n)} = \Theta^{(n)} \times \mathscr{H}_k$, the sieve space for $\mathscr{A}^{(n)}$. For estimation, we assume that only a few of the parameters $\theta_{0j}^{(n)}$ are nonzero, that is, we partition $\theta_0^{(n)}$ into two parts as $\theta_0^{(n)} = (\theta_{0S}^{(n)\prime}, \theta_{0N}^{(n)\prime})'$ where $\theta_{0S}^{(n)}$ and $\theta_{0N}^{(n)}$ correspond to the important regressors $X_S^{(n)}$ and trivial regressors $X_N^{(n)}$, respectively, and $\theta_{0N}^{(n)} = 0$. Thus $X^{(n)\prime}\theta_0^{(n)} = (X_S^{(n)\prime}, X_N^{(n)\prime})(\theta_{0S}^{(n)\prime}, \theta_{0N}^{(n)\prime})'$, and we assume that the dimension of $\theta_{0S}^{(n)}$ is some s such that $s \leq n$ and s grows very slowly compared to n.

Throughout this paper, we denote the Euclidean norm by $\|\cdot\|_E$. Specifically, for every positive integer j and every j-dimensional vector x, we use $\|x\|_E = \sqrt{x_1^2 + x_2^2 + \cdots + x_j^2}$. Also, for every square matrix A, we let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the smallest and largest eigenvalues of A, and we use $\|A\|_E = \lambda_{\max}^{1/2}(A'A)$. For any (possibly random) positive sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, $a_n = O_p(b_n)$ means that $\lim_{c\to\infty} \limsup_n P(a_n/b_n > c) = 0$, and $a_n = o_p(b_n)$ means that for each $\varepsilon > 0$, $\lim_{n\to\infty} P(a_n/b_n > \varepsilon) = 0$.

3.2 Sieve Focused GMM Estimator

To simplify the notation, we omit the superscript (n) on parameters and random variables. Fan & Liao (2014) propose a focused GMM estimator for nonlinear parametric high-dimensional

models based on the general moment conditions

$$E\left[g(Y,X'\beta)|W\right]=0, \tag{3.2}$$

where g is a known smooth function and $\beta \in \mathbb{R}^p$ is the vector of finite-dimensional parameters of interest. The estimator proposed in Fan & Liao (2014) is a new method for dealing with high dimensions in general endogenous models. Model (3.1), which takes the nonparametric component into account, is a semiparametric extension of model (3.2). To illustrate these ideas, we introduce several specific examples of model (3.1).

Example 3.2.1 (label=PLM) *In the partially linear model (PLM),*

$$Y_1 = X'\theta + h(Y_2) + \varepsilon, \tag{3.3}$$

where $Y=(Y_1,Y_2)$ has a fixed dimension $d_y=2$, θ is a p-component column vector of regression coefficients associated with X, and h is an unknown function of Y_2. Ai & Chen (2003) and Chen & Pouzo (2009) discussed this model under the assumptions of finite dimensions with possible endogeneity. If $p\to\infty$ as $n\to\infty$, (3.3) *becomes a high-dimensional semiparametric model. Xie & Huang (2009) propose a smoothly clipped absolute deviation (SCAD) penalized estimator and obtain oracle properties for it under certain conditions by assuming that both X and Y_2 are exogenous variables. We also allow X and Y_2 to be endogenous.*

Example 3.2.2 *Another classic semiparametric model is the single-index model (SIM):*

$$Y = h\left(X'\theta\right) + \varepsilon, \tag{3.4}$$

where Y is a scalar variable, X is a p-component column vector of covariates, $\theta=(\theta_1,...,\theta_p)'$ is the vector of finite-dimensional parameters of interest, and h is a smooth unknown function. Model (3.4) *is another specific example of Ai & Chen (2003) and Chen & Pouzo (2009) under*

the assumptions of finite dimensions and endogeneity. Peng & Huang (2011) studied this model under the assumptions of high dimensions and exogeneity. Model (3.1) allows X in model (3.4) *to be endogenous .*

Example 3.2.3 *As mentioned earlier, model* (3.1) *can easily be transformed to include more complicated cases. Consider the partially linear single index (PLSI) model*

$$Y = \eta\left(Z'\alpha\right) + X'\theta + \varepsilon, \tag{3.5}$$

where Z and X are q- and p-dimensional covariate vectors, respectively. Liang et al. *(2010) study the case where $q \to \infty$ and $p \to \infty$ as $n \to \infty$ by assuming that both Z and X are exogenous variables. Model* (3.5) *can be incorporated into model* (3.1)*, as we consider an additional high-dimensional component $Z'\alpha$ in model* (3.1)*.*

Because of the high-dimensional component, model (3.1) is different from the classic semiparametric models in, for example, Ai & Chen (2003) and Chen & Pouzo (2009). In this paper, we will extend the basic idea in Fan & Liao (2014) to deal with high dimensions in general semiparametric models when endogeneity arises. Let $(f_1, f_2, \ldots)$, $(h_1, h_2, \ldots)$, and $(g_1, g_2, \ldots)$ be three different sets of transformations of W, for example, B-splines, Fourier series, polynomials, or any other series basis. Write $F = (f_1(W), f_2(W), \ldots)'$, $H = (h_1(W), h_2(W), \ldots)'$, and $G = (g_1(W), g_2(W), \ldots)'$. Then we obtain the moment conditions under (3.1):

$$E\left[\rho\left(Y, X'\theta_0, h_0\left(\delta\left(Y, X'\theta_0\right)\right)\right)F\right] = 0,$$

$$E\left[\rho\left(Y, X'\theta_0, h_0\left(\delta\left(Y, X'\theta_0\right)\right)\right)H\right] = 0,$$

and

$$E\left[\rho\left(Y, X'\theta_0, h_0\left(\delta\left(Y, X'\theta_0\right)\right)\right)G\right] = 0.$$

We write $(Y, X, W) = Z$. In model (3.1), the original map $\rho = \rho(u_1, u_2, u_3)$, where u_1 is a d_y-

dimensional variable, and u_2 and u_3 are scalar variables. For simplicity, we will also write $\rho = \rho(Z, \alpha)$ for model (3.1), where $\alpha = (\theta, h)$. Let $V = (F', H', G')'$.

Let $\mathscr{H}_k$ be a linear finite-dimensional sieve space for h. Each $h \in \mathscr{H}_k$ takes the form $h = \sum_{j=1}^{k} b_j \varphi_j$, where $\{\varphi_j\}_{j=1}^{\infty}$ is a set of basis functions in $\mathscr{H}$. As mentioned earlier, $\mathscr{A}_k^{(n)} = \Theta^{(n)} \times \mathscr{H}_k$ is a sieve space for $\mathscr{A}^{(n)}$. Because we don't know the exact form of h_0, we use $\sum_{j=1}^{k} b_j \varphi_j$ to approximate it. k will be required to be increasing in n, and to increase at a particular rate.

The norm used to measure the distance between two parameters in the space $\mathscr{A}^{(n)}$ is defined below. We will obtain the consistency for the parameters of interest under this norm. Notice that δ is assumed to be a scalar function in this model, so for all $h \in \mathscr{H}$ we first define

$$\|h\|_s = \max_{m \leq 2} \sup_t \left| \frac{\partial^m h(t)}{\partial t^m} \right|. \tag{3.6}$$

Then we let

$$\|\alpha\|_s = \|\theta\|_E + \|h\|_s \tag{3.7}$$

for all $\alpha \in \mathscr{A}^{(n)}$.

In applications, the smoothness of the functions in $\mathscr{H}$ determines how well a sieve space can approximate $\mathscr{H}$. A typical space of smooth functions is the Hölder space $\Lambda^{\gamma}(\mathscr{X})$ of order $\gamma > 0$. For all $g \in \Lambda^{\gamma}(\mathscr{X})$, $g : \mathscr{X} \mapsto \mathbb{R}$, the first $\underline{\gamma}$ derivatives of g are bounded, and the $\underline{\gamma}$-th derivatives are Hölder continuous with exponent $\gamma - \underline{\gamma} \in (0, 1]$, where $\underline{\gamma}$ is the largest integer such that $\underline{\gamma} < \gamma$. In model (3.1), δ is a scalar function, so $\mathscr{X} \subset \mathbb{R}$ and if we consider the Hölder space with $\gamma > 2$, then all $h \in \mathscr{H}$ satisfy the condition that $\|h\|_s < C$ for some $C > 0$.

Define

$$J = diag\{1\{\theta_1 \neq 0\}\omega_{11}, \ldots, 1\{\theta_p \neq 0\}\omega_{p1}, 1\{\theta_1 \neq 0\}\omega_{12}, \ldots, 1\{\theta_p \neq 0\}\omega_{p2}, \omega_{13}, \ldots, \omega_{k3}\},$$

where $\left\{w_{j1}, w_{j2}\right\}_{j=1}^{p}$ and $\left\{w_{j3}\right\}_{j=1}^{k}$ are some constant weights, and define

$$V = (f_1(W), \ldots, f_p(W), h_1(W), \ldots, h_p(W), g_1(W), \ldots, g_k(W))'.$$

Let $J(\theta)$ be a diagonal matrix whose diagonal elements are those ω_{jt}'s with $j \in \{1, \ldots, p\}$, $t \in \{1,2\}$, and $\theta_j \neq 0$ and all ω_{j3}'s. Let $V(\theta)$ be a column vector with elements all from V except for those f_j's and h_j's with $\theta_j = 0$.

The sieve focused GMM (SFGMM) loss function is constructed to be

$$\begin{aligned}
\tilde{Q}(\alpha) =& \frac{1}{2}\sum_{j=1}^{p} 1\left\{\theta_j \neq 0\right\} \times \left\{ w_{j1}\left[\frac{1}{n}\sum_{i=1}^{n}\rho\left(Z_i,\alpha\right) f_j\left(W_i\right)\right]^2 + w_{j2}\left[\frac{1}{n}\sum_{i=1}^{n}\rho\left(Z_i,\alpha\right) h_j\left(W_i\right)\right]^2 \right\} \\
&+ \frac{1}{2}\sum_{j=1}^{k}\left\{ w_{j3}\left[\frac{1}{n}\sum_{i=1}^{n}\rho\left(Z_i,\alpha\right) g_j\left(W_i\right)\right]^2 \right\} \\
=& \frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n}\rho\left(Z_i,\alpha\right) V_i\right]' J \left[\frac{1}{n}\sum_{i=1}^{n}\rho\left(Z_i,\alpha\right) V_i\right] \\
=& \frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n}\rho\left(Z_i,\alpha\right) V_i(\theta)\right]' J(\theta) \left[\frac{1}{n}\sum_{i=1}^{n}\rho\left(Z_i,\alpha\right) V_i(\theta)\right],
\end{aligned} \tag{3.8}$$

where $\alpha \in \mathscr{A}_k^{(n)}$. Let $b = (b_1, \ldots, b_k)'$.

In the sieve space, the parameters to be estimated are $\gamma = (\theta', b')'$. The SFGMM loss function consists of two parts. The first part (associated with F and H), that is, the first summation in (3.8), is for the parametric parameters θ, which is similar to that in Fan & Liao (2014). The second part (associated with G), that is, the second summation in (3.8), is for the sieve parameters b, that is, the coefficients of the basis functions which are used to approximate h.

Then the SFGMM sample criterion function is

$$\hat{Q}(\alpha) = \tilde{Q}(\alpha) + \sum_{j=1}^{p} P_n\left(\left|\theta_j\right|\right), \tag{3.9}$$

where $\{P_n\}$ is a sequence of penalty functions for θ. The SFGMM estimator is a local minimizer of the criterion function (3.9), which achieves variable selection.

The indicator function in the first part of the loss function (3.8) is included for the purpose of reducing dimensions and preventing the accumulation of estimation errors, so that the oracle consistency for the estimator is achievable. For details and examples of this, see Fan & Liao (2014). Fan & Liao (2014) also explained why two sets of IV's for θ are included. This over-identification setup rules out extreme cases in which most of the coefficients can be set to be zeros, which minimizes the criterion function, but they are far from the true values. The second summation in (3.8) is important not only because it shows the sample moment conditions for the sieve parameters but also because it helps rule out extreme cases in which all the parametric parameters are equal to 0. If we omit the second part, the criterion function becomes

$$\hat{Q}(\alpha) = \frac{1}{2}\sum_{j=1}^{p} 1\{\theta_j \neq 0\}\left\{ w_{j1}\left[\frac{1}{n}\sum_{i=1}^{n}\rho(Z_i,\alpha) f_j(W_i)\right]^2 + w_{j2}\left[\frac{1}{n}\sum_{i=1}^{n}\rho(Z_i,\alpha) h_j(W_i)\right]^2\right\} + \sum_{j=1}^{p} P_n(|\theta_j|). \tag{3.10}$$

It is obvious that $\theta = 0$ always minimizes the criterion function (3.10), but 0 is not likely to be the true value of θ in most cases. The second part, for the nonparametric component parameters in the loss function (3.8), is an adjustment for this extreme case. If we let all the parametric coefficients to be equal to 0 in the loss function (3.8), the first part will be 0, but the second part will become large.

3.3 Oracle Consistency and Convergence Rates

In order to establish the consistency under $\|\cdot\|_s$, we first impose several basic conditions.

Assumption 3.3.1 *The data* $\{Y_i, X_i, W_i\}_{i=1}^n$ *is i.i.d..*

Assumption 3.3.2 *The true parameter* α_0 *is uniquely identified by model (3.1).*

Assumption 3.3.3 *The parameter spaces satisfy the following conditions:*

(i) $\mathscr{A}_k^{(n)} = \Theta^{(n)} \times \mathscr{H}_k$ *is compact under* $\|\cdot\|_s$*;*

(ii) Let $\mathscr{A}_k^{(s)} = \Theta^{(s)} \times \mathscr{H}_k$*, where* $\Theta^{(s)}$ *is the space for* θ_S*. For each n, there exists* $\Pi_n \alpha_{0S} \in \mathscr{A}_k^{(s)}$ *such that* $\Pi_n \alpha_{0S} = \left(\theta_{0S}, \sum_{j=1}^k b_0 \varphi_j\right)$ *and* $\|\Pi_n \alpha_{0S} - \alpha_{0S}\|_s = O(\delta_{0S}) = o(1)$ *for some* $\delta_{0S} > 0$*.*

Define $d_n = \frac{1}{2} \min\left\{\left|\theta_{0j}\right| : \theta_{0j} \neq 0, j = 1, ..., p\right\}$. d_n represents the strength of the signals.

Assumption 3.3.4 *There is a penalty function* $P_n(t) : [0, \infty) \to \infty$ *such that*

(i) $P_n(0) = 0$*;*

(ii) $P_n(t)$ *is concave and nondecreasing on* $[0, \infty)$*, and it has a continuous derivative* $P_n'(t)$ *when* $t > 0$*;*

(iii) $\sqrt{s} P_n'(d_n) = o(1/\sqrt{n})$.

By Assumption 3.3.1, this paper focuses on independent data. Assumption 3.3.2 is the standard identification condition. Assumption 3.3.3(i) requires the sieve space to be compact under the norm $\|\cdot\|_s$. Assumption 3.3.3(ii) is a condition of the sieve space restricted to the important finite-dimensional parameters. $\Pi_n \alpha_{0S}$ is the projection of α_{0S} onto the sieve space, and the distance between $\Pi_n \alpha_{0S}$ and α_{0S} decreases at the rate δ_{0S}.

Assumption 3.3.4 defines the penalty function, which is similar to the concave penalty function in Fan & Li (2001). This condition is standard, and it can be shown that with properly chosen tuning parameters, the L_p penalty (for $p \leq 1$), hard-thresholding Antoniadis (1996), SCAD (Fan & Li, 2001), and MCP (Zhang, 2010) all satisfy these conditions. Fan & Li (2001) show that a folded concave penalty is needed for an estimator to achieve three important oracle properties: unbiasedness, sparsity, and continuity. In this paper, we employ the same conditions for the penalty function.

To obtain the oracle consistency of the SFGMM estimator, we now introduce the concept of functional differentiability. Define the first pathwise derivative of a functional $F : \mathscr{A}^{(n)} \mapsto \mathbb{R}$ in the direction $[\Delta\alpha_1]$ evaluated at α_0 by

$$\frac{\partial F(\alpha_0)}{\partial \alpha}[\Delta\alpha_1] = \frac{\partial F(\alpha_0 + t\Delta\alpha_1)}{\partial t}|_{t=0}. \tag{3.11}$$

Then the second pathwise derivative of F in the direction $[\Delta\alpha_1, \Delta\alpha_2]$ evaluated at α_0 is given by

$$\frac{\partial^2 F(\alpha_0)}{\partial \alpha^2}[\Delta\alpha_1, \Delta\alpha_2] = \frac{\partial \frac{\partial F(\alpha_0 + t\Delta\alpha_2)}{\partial \alpha}[\Delta\alpha_1]}{\partial t}|_{t=0}. \tag{3.12}$$

In this way, we can define the K-th pathwise derivative of F the direction $[\Delta\alpha_1, ..., \Delta\alpha_K]$ evaluated at α_0. If $\Delta\alpha_1 = ... = \Delta\alpha_K = \Delta\alpha$, then

$$\frac{\partial^K F(\alpha_0)}{\partial \alpha^K}[\Delta\alpha_1, \Delta\alpha_2, ..., \Delta\alpha_K] = \frac{\partial^K F(\alpha_0)}{\partial \alpha^K}[\Delta\alpha]^K = \frac{\partial^K F(\alpha_0 + t\Delta\alpha)}{\partial t^K}|_{t=0}. \tag{3.13}$$

Clearly, the pathwise functional derivative is essentially an extension of simple function derivatives. With the definitions above, we define the functional Taylor expansion.

Definition 3.3.1 *Suppose the functional $F : \mathscr{A}^{(n)} \mapsto \mathbb{R}$ has a $(K+1)$-st pathwise derivative in the direction $[\Delta\alpha]^{k+1}$. Then the functional Taylor expansion of F is*

$$\begin{aligned} F(\alpha_0 + \Delta\alpha) =& F(\alpha_0) + \frac{\partial F(\alpha_0)}{\partial \alpha}[\Delta\alpha] + \frac{1}{2!}\frac{\partial^2 F(\alpha_0)}{\partial \alpha^2}[\Delta\alpha]^2 \\ &+ ... + \frac{1}{K!}\frac{\partial F^K(\alpha_0)}{\partial \alpha^K}[\Delta\alpha]^K + \frac{1}{(K+1)!}\frac{\partial^{K+1} F(\alpha_0 + \tau\Delta\alpha)}{\partial \alpha^{K+1}}[\Delta\alpha]^{K+1}, \end{aligned} \tag{3.14}$$

where $\tau \in (0,1)$.

The functional Taylor expansion is an extension of the univariate Taylor expansion. If we let $f(t) = F(\alpha_0 + t\Delta\alpha)$ and expand $f(1)$ around $t = 0$, we get the functional Taylor expansion. Conditional on the dataset $Z = (Y, X, W)$, $\tilde{Q}(\alpha)$ is a functional of $\alpha \in \mathscr{A}^{(n)}$.

Now, we introduce conditions under which the oracle consistency of the SFGMM estimator can be obtained for model (3.1).

Assumption 3.3.5 *The residual function $\rho(Z,\alpha)$ satisfies the following:*

(i) The first and second path derivatives of $\rho(Z,\alpha)$ exist for all $\alpha \in \mathscr{A}^{(n)}$, for almost surely all Z;

(ii) As a map in model (3.1)*, $\rho : \mathbb{R}^{d_y+2} \mapsto \mathbb{R}$ and $\rho = \rho(u_1,u_2,u_3)$ with*

$$\sup_{u_1,u_2,u_3}\left|\frac{\partial\rho(u_1,u_2,u_3)}{\partial u_2}\right|<\infty,\ \sup_{u_1,u_2,u_3}\left|\frac{\partial\rho(u_1,u_2,u_3)}{\partial u_3}\right|<\infty,\ \sup_{u_1,u_2,u_3}\left|\frac{\partial^2\rho(u_1,u_2,u_3)}{\partial u_2^2}\right|<\infty,$$
$$\sup_{u_1,u_2,u_3}\left|\frac{\partial^2\rho(u_1,u_2,u_3)}{\partial u_3^2}\right|<\infty,\ \sup_{u_1,u_2,u_3}\left|\frac{\partial^2\rho(u_1,u_2,u_3)}{\partial u_2\partial u_3}\right|<\infty,$$

where u_1 is a d_y-dimension vector and u_2, u_3 are scalars.

Assumption 3.3.5 requires that the residual function $\rho(Z,\alpha)$ be second-order pathwise differentiable with respect to α, and that $\rho(u_1,u_2,u_3)$ have bounded first- and second-order partial derivatives with respect to u_2, u_3. This condition can easily be satisfied by many nonparametric and semiparametric models.

Example 3.3.1 (continues=PLM) *In the PLM, $Y_1 = X'\theta + h(Y_2) + \varepsilon$ and $\rho(Z,\alpha) = Y_1 - X'\theta - h(Y_2)$, where $Z=(Y,X)$ and $Y=(Y_1,Y_2)$. Then*

$$\begin{aligned}\frac{\partial\rho(Z,\alpha)}{\partial\alpha}[\Delta\alpha_1] &= \frac{\partial\rho(Z,\alpha+t\Delta\alpha_1)}{\partial t}|_{t=0} = \frac{\partial\left[Y_1 - X'(\theta+t\Delta\theta_1) - (h+t\Delta h_1)(Y_2)\right]}{\partial t}|_{t=0}\\ &= -X'\Delta\theta_1 - \Delta h_1(Y_2) \qquad (3.15)\end{aligned}$$

and

$$\frac{\partial^2\rho(Z,\alpha)}{\partial\alpha^2}[\Delta\alpha_1,\Delta\alpha_2] = 0. \qquad (3.16)$$

In this case, $\rho(u_1,u_2,u_3) = u_{11} - u_2 - u_3$, where $u_1 = (u_{11},u_{12})$. It is obvious that Assumption 3.3.5 holds in this example.

Example 3.3.2 (continues=SIM) *In the SIM, $Y = h(X'\theta) + \varepsilon$, and $\rho(Z,\alpha) = Y - h(X'\theta)$. Then if $\Delta\alpha_1 = (\Delta\theta_1, \Delta h_1)$ and $\Delta\alpha_2 = (\Delta\theta_2, \Delta h_2)$, where Δh_1 and Δh_2 are both second-order differentiable,*

$$\begin{aligned}\frac{\partial \rho(Z,\alpha)}{\partial \alpha}[\Delta\alpha_1] &= \frac{\partial \rho(Z,\alpha + t\Delta\alpha_1)}{\partial t}|_{t=0} = \frac{\partial [Y - (h + t\Delta h_1)(X'(\theta + t\Delta\theta_1))]}{\partial t}|_{t=0} \\ &= -h'\left(X'\theta\right)X'\Delta\theta_1 - \Delta h_1\left(X'\theta\right) \end{aligned} \tag{3.17}$$

and

$$\begin{aligned}\frac{\partial \rho^2(Z,\alpha)}{\partial \alpha^2}[\Delta\alpha_1, \Delta\alpha_2] &= \frac{\partial \left[-(h + t\Delta h_2)'(X'(\theta + t\Delta\theta_2))X'\Delta\theta_1 - \Delta h_1(X'(\theta + t\Delta\theta_2))\right]}{\partial t}|_{t=0} \\ &= -h''\left(X'\theta\right)X'\Delta\theta_1 X'\Delta\theta_2 - \Delta h_2'\left(X'\theta\right)X'\Delta\theta_1 - \Delta h_1'\left(X'\theta\right)X'\Delta\theta_2. \end{aligned} \tag{3.18}$$

In this case, $\rho(u_1,u_2,u_3) = u_1 - u_3$, so Assumption 3.3.5 holds.

Let $S = \left\{j \leq p : \theta_{0j} \neq 0\right\}$, the set of indexes of nonzero coefficients in the high dimensional component. Let $\rho_S(Z,\alpha_S) = \rho(Z,((\theta_S',0')',h))$, $\tilde{Q}_S(\alpha_S) = \tilde{Q}\left(\left(\left(\theta_S',0'\right)',h\right)\right)$, and $\hat{Q}_S(\alpha_S) = \hat{Q}\left(\left(\left(\theta_S',0'\right)',h\right)\right)$, where $\alpha_S = (\theta_S,h)$ and $\alpha = \left(\left(\theta_S',\theta_N'\right)',h\right)$. For simplicity, we will also write $\alpha = (\alpha_S,\theta_N)$ according to the context. Notice that S cannot be identified at the outset. We first explore the properties of the estimator for the nonzero finite-dimensional parameters and the infinite-dimensional parameters, pretending that S is known. Then under certain conditions, we show that S need not be identified but the whole estimator $\hat{\alpha}$ (i.e., for the important and unimportant finite-dimensional parameters and the infinite-dimensional parameter) can automatically achieve oracle properties. Notice that if Assumption 3.3.5 is satisfied by ρ, then $\rho_S(Z,\alpha_S)$ satisfies Assumption 3.3.5(i) with respect to α_S. Let J_S denote $J((\theta_S',0')')$, let V_S denote $V((\theta_S',0')')$, and let $\gamma_S = (\theta_S',b')'$.

Assumption 3.3.6 *Suppose the following inequalities hold:*

(i) *For all n,* $\max_{m\leq 2}\sup_t \sqrt{\sum_{j=1}^{k}\left(\partial^m \varphi_j(t)/\partial t^m\right)^2} = \delta_{n\varphi} < \infty$ *for some* $\delta_{n\varphi} > 0$*;*

(ii) $\lambda_{\max}\left\{E\left[(\partial \rho_S(Z_i,\Pi_n\alpha_{0S})/\partial\gamma_S)V_{Si}'\right]E\left[(\partial \rho_S(Z_i,\Pi_n\alpha_{0S})/\partial\gamma_S)V_{Si}'\right]'\right\} > C_\lambda(n)$ *for some* $C_\lambda(n) > 0$ *such that* $\lim_{n\to\infty} C_\lambda(n)^{-2}\delta_{n\varphi}^2 s\left(2s^2+sk\right)(2s+k)\sqrt{\log p/n} = 0$*;*

(iii) J_S *is a deterministic diagonal matrix based on the sample criterion function* (3.8)*, where* $\lambda_{\min}(J_S) > C_{JL}$ *and* $\lambda_{\max}(J_S) < C_{JU}$ *for some* $C_{JL}, C_{JU} > 0$*.*

Assumption 3.3.6(i) is a condition on the basis functions. An example of this is the Fourier basis. The condition similar to Assumption 3.3.6(ii) can be found in Bradic *et al.* (2011), Fan & Lv (2011), and Fan & Liao (2014). In this paper, we relax this condition for nonparametric and semiparametric models, that is, we allow $C_\lambda(n)$ to decay as $n\to\infty$, and Assumption 3.3.6(ii) requires that it not decay too fast. Assumption 3.3.6(iii) is a standard condition on J_S which can easily be satisfied by construction. Remember V_S and X_S are column vectors and we use V_{Sl} and X_{Sm} to denote the l-th and the m-th element of the vectors, respectively.

Assumption 3.3.7 *Suppose the following conditions hold:*

(i) $E\left[|V_{Sl}|^2\right] \leq M_V^2$ *for all* l *and some* $M_V > 0$*;* $E\left[|V_{Sl}X_{Sm}|^2\right] \leq M_{VX}^2$ *for all* l,m *and some* $M_{VX} > 0$*;* $E\left[|V_{Sl}X_{Sm}X_{St}|^2\right] \leq M_{VXX}^2$ *for all* l,m,t *and some* $M_{VXX} > 0$*;*

(ii) $C_\lambda(n)^{-1}\delta_{n\varphi}^2\sqrt{\frac{1}{n}(2s^2+sk)(2s+k)\log p} = o(d_n)$*;*

(iii) *Let* $C_\lambda(n)$ *satisfies 3.3.6(ii), then*

$$\|\alpha_{0S}-\Pi_n\alpha_{0S}\|_s = O_p\left(\min\left(\sqrt{\frac{(2s+k)\log p}{n(2s^2+sk)}}, \frac{\delta_{n\varphi}^2\sqrt{\frac{1}{n}(2s^2+sk)(2s+k)\log p}}{C_\lambda(n)}\right)\right);$$

(iv) $C_\lambda(n)^{-1}\delta_{n\varphi}^2\left(2s^2+sk\right)(2s+k)\sqrt{\frac{1}{n}\log p} = o\left(P_n'(0^+)\right)$.

Assumption 3.3.7(i) shows some moment conditions on X and V. It requires that the moments be uniformly bounded as more and more important regressors enter the model.

Assumption 3.3.7(ii) requires the signal to be sufficiently strong, that is,

$$d_n \gg \left(\delta_{n\varphi}^2 \sqrt{(2s^2+sk)(2s+k)\log p/n} \right).$$

Assumption 3.3.7(iii) demonstrates the rate of convergence of the projection of α_{0S}. This is a high-level assumption. Assumption 3.3.7(iv) shows the relationship among s, k, p, $\delta_{n\varphi}$, n, and the penalty function at the origin. When the SCAD penalty function is used with a tuning parameter λ_n, $P_n'(0^+) = \lambda_n$. Then assumption 3.3.7(iv) implies that λ_n cannot decrease too fast.

Theorem 3.3.1 *Under Assumptions 3.3.1-3.3.7, with probability approaching one, there is a local minimizer* $\hat{\alpha} = ((\hat{\theta}_S', 0')', \hat{h})$ *of* $\hat{Q}(\alpha)$ *in the sieve space* $\mathscr{A}_k^{(n)}$ *such that* $\|\hat{\alpha} - \alpha_0\|_s = O_p(\delta_{n\alpha})$, *where* $\delta_{n\alpha} = O(C_\lambda (n)^{-1} \delta_{n\varphi}^2 \sqrt{n^{-1}(2s^2+sk)(2s+k)\log p})$.

The basic idea of the existence of such a local minimizer is that we first find a neighborhood of the sieve projection $\Pi_n \alpha_{0S}$ in which the sample criterion function evaluated on the boundary is larger than that same function evaluated at $\Pi_n \alpha_{0S}$; then by continuity of the criterion function on the sieve space, we know that there must be a local minimizer of the criterion function inside this neighborhood. As $n \to \infty$, the neighborhood shrinks by construction and the minimizer gets closer to the projection $\Pi_n \alpha_{0S}$. Under the assumption that $\Pi_n \alpha_{0S}$ approaches the true value α_{0S} fast enough, the local minimizer converges to α_{0S} at a certain rate.

Theorem 3.3.1 shows the existence of a consistent estimator for α_0 on the sieve space which achieves variable selection. In addition, it provides a desirable convergence rate under which we can obtain the asymptotic normality of a plug-in SFGMM estimator.

3.4 Asymptotic Normality of Plug-in SFGMM Estimator

In this paper, we are interested in the asymptotic distribution of a plug-in SFGMM estimator $f(\hat{\alpha}_S)$ for $f(\alpha_{0S})$, where $f : \mathscr{A}^{(s)} \mapsto \mathbb{R}$ is a known measurable map. For example, $f(\alpha_S) = \eta' \theta_S$ for some known vector η with dimension s, which is similar to that in Fan & Liao

(2014), and $f(\alpha_S) = h(\bar{\delta})$ when we care about the behavior of h_0 at a specific point $\bar{\delta}$. We don't consider the unimportant regressors here, since by Theorem 3.3.1 we can identify S. Chen *et al.* (2014) provide a method to obtain the asymptotic normality of $f(\hat{\alpha}_S)$ in the finite-dimensional case, while in this paper we extend this method to obtain the limiting distribution under the assumptions of high dimensions.

By Theorem 3.3.1, we have $\|\hat{\alpha}_S - \alpha_{0S}\|_s = O_p(\delta_{n\alpha})$, where $\delta_{n\alpha}$ is the convergence rate, and $\hat{\alpha}_S$ is a local minimizer in $\mathscr{A}_k^{(s)}$. Next, we construct a subset of $\mathscr{A}_k^{(s)}$, which we denote by $\mathscr{B}_n^S(\tau)$, in which $\hat{\alpha}_S$ is a global minimizer. Specifically,

$$\mathscr{B}_0^S = \left\{ \alpha_S \in \mathscr{A}^{(s)} : \|\alpha_S - \alpha_{0S}\|_s \leq \delta_{n\alpha} \right\}$$

and

$$\mathscr{B}_n^S(\tau) = \left\{ \alpha_S \in \mathscr{A}_k^{(s)} : \|\theta_S - \theta_{0S}\|_E \leq \tau_1(n)\, e_{1n}, \|b - b_0\|_E \leq \tau_2(n)\, e_{2n} \right\},$$

where $\tau = (\tau_1(n), \tau_2(n))$ and $\tau_1(n)$, $\tau_2(n)$, e_{1n}, and e_{2n} are obtained as in the proof of Theorem 3.3.1. Then it can be shown that $\mathscr{B}_n^S(\tau) \subseteq \mathscr{B}_0^S \cap \mathscr{A}_k^s$.

Suppose that for all $\alpha_S \in \mathscr{B}_0^S$, $\tilde{Q}_S(\alpha_S) - \tilde{Q}_S(\alpha_{0S})$ can be well approximated by a score process $\Delta(Z, \alpha_{0S})[\alpha_S - \alpha_{0S}]$ such that $\Delta(Z, \alpha_{0S})[\alpha_S - \alpha_{0S}]$ is linear in $\alpha_S - \alpha_{0S}$. When $\tilde{Q}_S(\alpha_S)$ is pathwise differentiable at α_{0S} in the direction $[\alpha_S - \alpha_{0S}]$ for almost surely all Z and the pathwise derivative is linear in $\alpha_S - \alpha_{0S}$, we let

$$\Delta(Z, \alpha_{0S})[\alpha_S - \alpha_{0S}] = \frac{\partial \tilde{Q}_S(\alpha_{0S} + \tau(\alpha_S - \alpha_{0S}))}{\partial \tau}|_{\tau=0}. \tag{3.19}$$

Suppose that for all $\alpha_S \in \mathscr{B}_0^S$, $\partial E[\rho_S(Z, \alpha_{0S} + \tau(\alpha_S - \alpha_{0S}))V]/\partial\tau$ exists in a neighborhood of 0 and $\partial E[\rho_S(Z, \alpha_{0S} + \tau(\alpha_S - \alpha_{0S}))V]/\partial\tau$ is a linear functional of $\alpha_S - \alpha_{0S}$. Notice

that by definition,

$$\frac{\partial E\left[\rho_S\left(Z,\alpha_{0S}\right)V\right]}{\partial\alpha_S}\left[\alpha_S-\alpha_{0S}\right]=\lim_{\tau\to 0}\frac{E\left[\rho_S\left(Z,\alpha_{0S}+\tau\left(\alpha_S-\alpha_{0S}\right)\right)V\right]-E\left[\rho_S\left(Z,\alpha_{0S}\right)V\right]}{\tau}. \tag{3.20}$$

Then for any $\alpha_{S1},\alpha_{S2}\in\mathscr{B}_0^S$, define an inner product

$$\begin{aligned}\left\langle\alpha_{S1}-\alpha_{0S},\alpha_{S2}-\alpha_{0S}\right\rangle=&\left(\frac{\partial E\left[\rho_S\left(Z_i,\alpha_{0S}\right)V_{Si}'\right]}{\partial\alpha_S}\left[\alpha_{S1}-\alpha_{0S}\right]\right)J_S\\&\cdot\left(\frac{\partial E\left[\rho_S\left(Z_i,\alpha_{0S}\right)V_{Si}\right]}{\partial\alpha_S}\left[\alpha_{S2}-\alpha_{0S}\right]\right),\end{aligned} \tag{3.21}$$

and the corresponding norm of $\alpha_S\in\mathscr{B}_0^S$:

$$\left\|\alpha_S-\alpha_{0S}\right\|^2=\left(\frac{\partial E\left[\rho_S\left(Z_i,\alpha_{0S}\right)V_{Si}'\right]}{\partial\alpha_S}\left[\alpha_S-\alpha_{0S}\right]\right)J_S\left(\frac{\partial E\left[\rho_S\left(Z_i,\alpha_{0S}\right)V_{Si}\right]}{\partial\alpha_S}\left[\alpha_S-\alpha_{0S}\right]\right). \tag{3.22}$$

We say that $\alpha_S=\alpha_{0S}$ if $\left\|\alpha_S-\alpha_{0S}\right\|^2=0$, which means that the parameters are defined, in the sense of an equivalent class, according to the metric $\|\cdot\|$.

Under the regularity conditions,

$$\frac{\partial E\left[\rho_S\left(Z_i,\alpha_{0S}\right)V_{Si}\right]}{\partial\alpha_S}\left[\alpha_S-\alpha_{0S}\right]=E\left[\frac{\partial\rho_S\left(Z_i,\alpha_{0S}\right)}{\partial\alpha_S}\left[\alpha_S-\alpha_{0S}\right]V_{Si}\right]$$

for all n, that is, E and $\partial/\partial\alpha_S$ can be interchanged. Sufficient conditions for this interchange condition can be found in Chen *et al.* (2014), and we simply assume that this condition holds for all n.

Let $\mathscr{V}^S=clsp\left(\mathscr{B}_0^S\right)-\{\alpha_{0S}\}$, where $clsp\left(\mathscr{B}_0^S\right)$ denotes the closed linear span of $\mathscr{B}_0^S$ under $\|\cdot\|$. Then $\mathscr{V}^S$ is a Hilbert space under $\langle\cdot,\cdot\rangle$.

Define

$$\alpha_{0S,n} \in \arg \min_{\alpha_S \in clsp\left(\mathscr{B}_n^S(\tau)\right)} \|\alpha_S - \alpha_{0S}\|, \tag{3.23}$$

and let $\mathscr{V}_n^S(\tau) = clsp\left(\mathscr{B}_n^S(\tau)\right) - \left\{\alpha_{0S,n}\right\}$. Then for each n, $\mathscr{V}_n^S(\tau)$ is a Hilbert space under $\langle\cdot,\cdot\rangle$, and by definition $\langle \alpha_{0S,n} - \alpha_{0S}, v_n\rangle = 0$ for all $v_n \in \mathscr{V}_n^S(\tau)$.

For all $v \in \mathscr{V}^S$, define $\partial f(\alpha_{0S})/\partial \alpha_S[v]$ to be the pathwise derivative of the functional f at α_{0S} in the direction $v = \alpha_S - \alpha_{0S} \in \mathscr{V}^S$:

$$\frac{\partial f(\alpha_{0S})}{\partial \alpha_S}[v] = \frac{\partial f(\alpha_{0S} + \tau v)}{\partial \tau}|_{\tau=0}. \tag{3.24}$$

In what follows, we suppose that $\partial f(\alpha_{0S})/\partial \alpha_S[\cdot]$ is a linear functional on $\mathscr{V}^S$ and also on $\mathscr{V}_n^S(\tau)$. Since $\mathscr{V}_n^S(\tau)$ is a finite-dimensional Hilbert space and any linear functional on a finite-dimensional Hilbert space is bounded, by the Riesz representation theorem there is a $v_n^* \in \mathscr{V}_n^S(\tau)$ such that

$$\frac{\partial f(\alpha_{0S})}{\partial \alpha_S}[v_n] = \langle v_n^*, v_n\rangle \text{ for all } v_n \in \mathscr{V}_n^S(\tau) \tag{3.25}$$

and

$$\frac{\partial f(\alpha_{0S})}{\partial \alpha_S}[v_n^*] = \|v_n^*\|^2 = \sup_{v_n \in \mathscr{V}_n^S(\tau), v_n \neq 0} \left|\frac{\partial f(\alpha_{0S})}{\partial \alpha_S}[v_n]\right|^2 / \|v_n\|^2. \tag{3.26}$$

Details on how to find the closed form of v_n^* can be found in Chen *et al.* (2014). For completeness, we show briefly how to find the representer for each n in (3.25).

By definition, the sieve Riesz representer $v_n^* = \left(v_{\theta,n}^*, v_{h,n}^*\right) = \left(v_{\theta,n}^*, \sum_{j=1}^k b_j^* \varphi_j\right) \in \mathscr{V}_n^S(\tau)$

solves the optimization problem

$$\frac{\partial f(\alpha_{0S})}{\partial \alpha_S}[v_n^*] = \|v_n^*\|^2 = \sup_{v=(v_\theta,v_h)\in\mathscr{V}_n^S(\tau),v\neq 0} \frac{\left|\frac{\partial f(\alpha_{0S})}{\partial \theta'}v_\theta + \frac{\partial f(\alpha_{0S})}{\partial h}[v_h]\right|^2}{\langle v,v\rangle}$$
$$= \sup_{\gamma=\left(v_\theta',b'\right)',b\neq 0} \frac{\gamma' F_k F_k' \gamma}{\gamma' R_k \gamma}, \tag{3.27}$$

where $F_k = \left(\frac{\partial f(\alpha_{0S})}{\partial \theta_S'}, \frac{\partial f(\alpha_{0S})}{\partial h}[\varphi_1], ..., \frac{\partial f(\alpha_{0S})}{\partial h}[\varphi_k]\right)'$ is an $(s+k)$-component vector and R_k is an $(s+k)\times(s+k)$ positive definite matrix such that $\gamma' R_k \gamma = \langle v,v\rangle$ for all $v \in \mathscr{V}_n^S(\tau)$.

Let $\Phi_n = (\varphi_1, \varphi_2, ..., \varphi_k)'$, and define

$$R_k = \begin{pmatrix} I_{11} & I_{n,12} \\ I_{n,21} & I_{n,22} \end{pmatrix} \text{ and } R_k^{-1} = \begin{pmatrix} I_n^{11} & I_n^{12} \\ I_n^{21} & I_n^{22} \end{pmatrix}, \tag{3.28}$$

where $I_{11} = E\left[-\frac{\partial^2 \hat{Q}_S(\alpha_{0S})}{\partial \theta_S \partial \theta_S'}\right]$, $I_{n,12} = E\left[-\frac{\partial^2 \hat{Q}_S(\alpha_{0S})}{\partial \theta_S \partial h}[\varphi_1], ..., -\frac{\partial^2 \hat{Q}_S(\alpha_{0S})}{\partial \theta_S \partial h}[\varphi_k]\right]$, $I_{n,21} = I_{n,12}'$, and

$$I_{n,22} = \begin{pmatrix} \frac{\partial^2 \hat{Q}_S(\alpha_{0S})}{\partial h \partial h}[\varphi_1,\varphi_1] & \frac{\partial^2 \hat{Q}_S(\alpha_{0S})}{\partial h \partial h}[\varphi_1,\varphi_2] & \cdots & \frac{\partial^2 \hat{Q}_S(\alpha_{0S})}{\partial h \partial h}[\varphi_1,\varphi_k] \\ \frac{\partial^2 \hat{Q}_S(\alpha_{0S})}{\partial h \partial h}[\varphi_2,\varphi_1] & \frac{\partial^2 \hat{Q}_S(\alpha_{0S})}{\partial h \partial h}[\varphi_2,\varphi_2] & \cdots & \frac{\partial^2 \hat{Q}_S(\alpha_{0S})}{\partial h \partial h}[\varphi_2,\varphi_k] \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \hat{Q}_S(\alpha_{0S})}{\partial h \partial h}[\varphi_k,\varphi_1] & \frac{\partial^2 \hat{Q}_S(\alpha_{0S})}{\partial h \partial h}[\varphi_k,\varphi_2] & \cdots & \frac{\partial^2 \hat{Q}_S(\alpha_{0S})}{\partial h \partial h}[\varphi_k,\varphi_k] \end{pmatrix},$$

under the regularity conditions. Then the γ_n^* which solves the optimization problem (3.27) is given by

$$\gamma_n^* = \left(v_{\theta,n}^{*\prime}, b_n^{*\prime}\right)' = R_k^{-1} F_k. \tag{3.29}$$

The sieve Riesz representer is $v_n^* = \left(v_{\theta,n}^*, \Phi_n' b_n^*\right) \in \mathscr{V}_n^S(\tau)$. Also, $\|v_n^*\|^2 = \gamma_n^{*\prime} R_k \gamma_n^*$, which is finite for all n but it is possible that $\|v_n^*\| \to \infty$ as $n \to \infty$, in which case the functional $f(\alpha_{0S})$ is called an irregular functional. Chen *et al.* (2014) discuss the irregular case where $\|v_n^*\| \to \infty$,

while in this paper the high dimensions could be another reason for the blow-up of $\|v_n^*\|$.

To obtain the asymptotic normality of $\sqrt{n}(f(\hat{\alpha}_S) - f(\alpha_0))$, we introduce the following conditions.

Assumption 3.4.1 $s\sqrt{n(2s+k)(2s^2+sk)}\delta_{n\alpha}^2 = o(1)$.

Assumption 3.4.2 *The map f satisfies the following conditions:*

(i) $v \mapsto \partial f(\alpha_{0S})/\partial \alpha_S[v]$ *is a linear functional from* $\mathscr{V}^S$ *to* $\mathbb{R}$*;*

(ii) $\sup_{\alpha_S \in \mathscr{B}_n^S(\tau)} |f(\alpha_S) - f(\alpha_{0S}) - \partial f(\alpha_{0S})/\partial \alpha_S[\alpha_S - \alpha_{0S}]|/\|v_n^*\| = o\left(n^{-1/2}\right)$*;*

(iii) $\left|\partial f(\alpha_{0S})/\partial \alpha_S\left[\alpha_{0S,n} - \alpha_{0S}\right]\right|/\|v_n^*\| = \left(n^{-1/2}\right)$.

Assumption 3.4.1 gives the convergence rate for obtaining the asymptotic normality. In the low-dimensional case, this condition is equivalent to $k/n = o(1)$. In the high-dimensional case, Assumption 3.4.1 is a high-level condition, since s is increasing in n and $\delta_{n\alpha}$ is decreasing slower than $1/\sqrt{n}$. Assumption 3.4.2 shows local properties of the functional $f(\alpha_{0S})$. In Assumption 3.4.2(i), the linearity of $\partial f(\alpha_{0S})/\partial \alpha_S$ guarantees the existence of the Riesz representer defined earlier. Assumption 3.4.2(ii), which controls the linear approximation error of the possibly nonlinear functional f, is automatically satisfied if f is a linear functional. Assumption 3.4.2(iii) controls the bias which is due to the finite-dimensional sieve approximation $\alpha_{0S,n}$ of α_{0S}.

For simplicity, we assume that

$$\Delta(Z, \alpha_{0S})[\alpha_S - \alpha_{0S}] = \frac{\partial \tilde{Q}_S(\alpha_{0S} + \tau(\alpha_S - \alpha_{0S}))}{\partial \tau}|_{\tau=0}.$$

For example, in the PLM and the SIM, it is easy to show that the first pathwise derivative of $\tilde{Q}_S$ is linear. Also, we define

$$\begin{aligned}\|v_n^*\|_{sd}^2 = &\left(E\left[\frac{\partial \rho_S(Z, \alpha_{0S})}{\partial \alpha_S}[v_n^*]V_S'\right]\right)J_S \\ &\cdot [Var(\rho_S(Z, \alpha_{0S})V_S)]J_S\left(E\left[\frac{\partial \rho_S(Z, \alpha_{0S})}{\partial \alpha_S}[v_n^*]V_S\right]\right),\end{aligned} \tag{3.30}$$

the sample variance of v_n^* for model (3.1). Next, we introduce an assumption on $\|v_n^*\|_{sd}^2$.

Assumption 3.4.3 $\|v_n^*\| / \|v_n^*\|_{sd} = O(1)$.

Given the definition of v_n^*, we have that $\|v_n^*\| > 0$ and it is nondecreasing in $\dim\left(\mathscr{V}_n^S(\tau)\right)$ and nondecreasing in n. Since s increases as $n \to \infty$, it is possible that $\|v_n^*\| / \|v_n^*\|_{sd} = o(1)$. Assumption 3.4.3 rules out this possibility.

Define $u_n^* = v_n^* / \|v_n^*\|_{sd}$. Then $u_n^* = O(1)$, and by the linearity of Δ and the central limit theorem,

$$\begin{aligned}\sqrt{n}\Delta(Z, \alpha_{0S})[u_n^*] &= \sqrt{n}\frac{\partial \tilde{Q}_S(\alpha_{0S})}{\partial \alpha_S}[u_n^*] \\ &= \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \rho_S(Z_i, \alpha_{0S})}{\partial \alpha_S}[u_n^*] V_{Si}'\right] J_S \left[\frac{1}{\sqrt{n}}\sum_{i=1}^n \rho_S(Z_i, \alpha_{0S}) V_{Si}\right] \to_L N(0,1).\end{aligned} \tag{3.31}$$

We establish the asymptotic normality in the next theorem.

Theorem 3.4.1 *Under assumptions 3.3.1–3.3.7 and 3.4.1–3.4.3,*

$$\sqrt{n}\left(f(\hat{\alpha}_S) - f(\alpha_{0S})\right) / \|v_n^*\|_{sd} = \sqrt{n}\Delta(Z, \alpha_{0S})[u_n^*] + o_p(1) \to_L N(0,1), \tag{3.32}$$

where $\hat{\alpha}$ is the SFGMM estimator obtained in Theorem 3.3.1.

Theorem 3.4.1 shows that although the dimension of the regressors grows as n increases, under certain conditions the asymptotic normality of the plug-in estimator still holds.

3.4.1 Consistent Estimate of $J(\theta)$

Up until this point, we have assumed that $J(\theta)$ is deterministic, given that we know how to choose the weights $\{w_{j1}, w_{j2}\}_{j=1}^p$ and $\{w_{j3}\}_{j=1}^k$. But in applications, that is not always the

case. For example, if we want to standardize the scale, we can use

$$w_{j1} = 1/Var\left(f_j(W)\right), w_{j2} = 1/Var\left(h_j(W)\right), w_{j3} = 1/Var\left(g_j(W)\right). \tag{3.33}$$

Since we don't know the distribution of W, a consistent estimate of $J(\theta)$ is needed. In this example, we can use $\hat{w}_{j1} = 1/\widehat{Var}\left(f_j(W)\right), \hat{w}_{j2} = 1/\widehat{Var}\left(h_j(W)\right)$, and $\hat{w}_{j3} = 1/\widehat{Var}\left(g_j(W)\right)$ to construct $\hat{J}(\theta)$.

With the estimated version of $J(\theta)$, we obtain a new loss function, with $J(\theta)$ replaced by $\hat{J}(\theta)$, that is,

$$\begin{aligned}\tilde{Q}^w(\alpha) =& \frac{1}{2}\sum_{j=1}^{p} 1\left\{\theta_j \neq 0\right\}\left\{\hat{w}_{j1}\left[\frac{1}{n}\sum_{i=1}^{n}\rho(Z_i,\alpha) f_j(W_i)\right]^2 + \hat{w}_{j2}\left[\frac{1}{n}\sum_{i=1}^{n}\rho(Z_i,\alpha) h_j(W_i)\right]^2\right\} \\ &+ \frac{1}{2}\sum_{j=1}^{k}\left\{\hat{w}_{j3}\left[\frac{1}{n}\sum_{i=1}^{n}\rho(Z_i,\alpha) g_j(W_i)\right]^2\right\} \\ =& \frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n}\rho(Z_i,\alpha) V_i(\theta)\right]' \hat{J}(\theta)\left[\frac{1}{n}\sum_{i=1}^{n}\rho(Z_i,\alpha) V_i(\theta)\right].\end{aligned} \tag{3.34}$$

Then we let $\hat{\alpha}^w = \left(\left(\hat{\theta}_S^{w\prime}, 0'\right)', \hat{h}^w\right)$ be a local minimizer of $\hat{Q}^w(\alpha)$, where $\hat{Q}^w(\alpha) = \tilde{Q}^w(\alpha) + \sum_{j=1}^{p_n} P_n\left(\left|\theta_j\right|\right)$.

Assumption 3.4.4 *$\hat{J}(\theta)$ is a uniformly consistent estimate of $J(\theta)$ such that*

$$\sup_{\theta\in\Theta^{(n)}} \left\|\hat{J}(\theta) - J(\theta)\right\|_E = o_p(1).$$

Because of the special structure of $\hat{J}(\theta)$ and $J(\theta)$, to verify Assumption 3.4.4, it suffices to show that $\hat{J}$ is a consistent estimate of J, where

$$J = diag\left\{w_{11}, ..., w_{p1}, w_{12}, ..., w_{p2}, w_{13}, ..., w_{k3}\right\} \tag{3.35}$$

and

$$\hat{J} = diag\left\{\hat{w}_{11}, ..., \hat{w}_{p1}, \hat{w}_{12}, ..., \hat{w}_{p2}, \hat{w}_{13}, ..., \hat{w}_{k3}\right\}. \tag{3.36}$$

Proposition 3.4.1 *Under Assumptions 3.3.1–3.3.7 and 3.4.1-3.4.4,*

$$\sqrt{n}\left(f\left(\hat{\alpha}_S^w\right) - f\left(\alpha_{0S}\right)\right) = \sqrt{n}\Delta\left(Z, \alpha_{0S}\right)\left[u_n^*\right] + o_p\left(1\right) \to_L N\left(0,1\right). \tag{3.37}$$

The proof is similar to that of Theorem 3.4.1 under Assumption 3.4.4, so we omit it.

3.5 Implementation

As explained in Fan & Liao (2014), inclusion of an indicator function in (3.8) leads to dimension reduction, but it also yields a non-smooth loss function; as a result, minimizing $\hat{Q}\left(\alpha\right)$ is generally NP-hard. Horowitz (1992) and Bondell & Reich (2012) propose a smoothing technique which is employed by Fan & Liao (2014). We also apply this method to approximate the indicator function by a smooth kernel $K : \mathbb{R} \to \mathbb{R}$ such that

(i) $0 \leq K\left(t\right) < M$ for some finite M and all $t \geq 0$;

(ii) $K\left(0\right) = 0$ and $\lim_{|t|\to\infty} K\left(t\right) = 1$;

(iii) $\limsup_{|t|\to\infty} \left|K'\left(t\right)t\right| = 0$, and $\limsup_{|t|\to\infty} \left|K''\left(t\right)t^2\right| < \infty$.

One example is

$$K\left(t\right) = \frac{F\left(t\right) - F\left(0\right)}{1 - F\left(0\right)},$$

where $F\left(t\right)$ is a twice-differentiable cumulative distribution function. Given a predetermined small number r_n, the indicator function is approximated by $K\left(\theta_j^2/r_n\right)$. Then the smoothed

SFGMM criterion function is given by

$$\hat{Q}_K(\alpha) = \tilde{Q}_K(\alpha) + \sum_{j=1}^{p} P_n\left(\left|\theta_j\right|\right), \tag{3.38}$$

where $\tilde{Q}_K$ is the continuous approximation of $\tilde{Q}(\alpha)$ with the indicator function replaced by K. As $r_n \downarrow 0$, $K\left(\theta_j^2/r_n\right)$ converges to $1\left\{\theta_j \neq 0\right\}$, and $\tilde{Q}_K$ is a smoothed approximation of $\tilde{Q}$.

With the kernel approximation, we next employ the iterative coordinate algorithm to minimize the smoothed SFGMM criterion function. This algorithm, which has been used in Fu (1998), Daubechies *et al.* (2004), Fan & Lv (2011), and Fan & Liao (2014), etc, goes as follows:

1 Obtain an initial value for $(\theta', b')'$, for example by least-square estimation.

2 Keep the other coordinates fixed at their values while minimizing the sample criterion function by choosing values for one coordinate of $(\theta', b')'$.

3 Iterate step 2 for each coordinate until the difference between the old sample criterion function and the updated one converges to 0.

3.6 Simulation Evidence

In the simulation, we generated data by using a semiparametric model.

Specifically, we simulated from the partially linear model

$$Y_1 = X'\theta_0 + \frac{1}{\exp\left(X'\theta_0 + Y_2\right)^2} + \varepsilon \tag{3.39}$$

with $p = 50$ or 100, and $s = 5$, and with X_1, X_2, X_3, X_4, and X_5 being the important regressors. Let $(\theta_{01}, \theta_{02}, \theta_{03}, \theta_{04}, \theta_{05}) = (5, -4, 7, -2, 1.5)$, $\theta_{06} =, \cdots, = \theta_{0p} = 0$, and $h_0(t) = e^{-t}$. We used Fourier basis functions to approximate $h_0 = 1/\exp(\delta)$ with $k = 5$.

Y_2 was set to be an $N(0,1)$ random variable. For each j, X_j was classified as either exogenous (in which case it's denoted by X_j^x) or endogenous (in which case it's denoted by

X_j^e). Let $F=(F_1,...,F_p)'$, $H=(H_1,...,H_p)'$, and $G=(G_1,...,G_k)'$ be the transformations of a three-dimensional instrumental variable $W=(W_1,W_2,W_3)'\sim N(0,I_3)$, where

$$F_j(W)=\sqrt{2}\{\sin(j\pi W_1)+\sin(j\pi W_2)+\sin(j\pi W_3)\},1\leq j\leq p;$$
$$H_j(W)=\sqrt{2}\{\cos(j\pi W_1)+\cos(j\pi W_2)+\cos(j\pi W_3)\},1\leq j\leq p;$$
$$G_j(W)=\sqrt{2}\{\sin((p+j)\pi W_1)+\sin((p+j)\pi W_2)+\sin((p+j)\pi W_3)\},1\leq j\leq k.$$

X_j^x and X_j^e were generated as follows:

$$X_j^x=F_j+H_j+u_j, X_j^e=\left(F_j+H_j+1\right)(3\varepsilon+1),$$

where $\left\{\varepsilon,u_1,...,u_p\right\}$ are independent $N(0,1)$. There were a total of $m=10$ or $m=50$ endogenous variables, which would be specified later for two different cases.

The dataset contains $n=100$ i.i.d. tuples (Y,X,F,H,G). We used SCAD penalty functions with different predetermined tuning parameters λ.

For the smoothing indicator function in the loss function, we used logistic cumulative distribution function with $r_n=0.1$, that is,

$$F(t)=\frac{\exp(t)}{1+\exp(t)}, K\left(\frac{\theta_j^2}{r_n}\right)=2F\left(\frac{\theta_j^2}{r_n}\right)-1.$$

There were 100 replications for each simulation. Four performance measures were used to evaluate the properties of the SFGMM estimator. The first one is the mean standard error for the important regressors, MSE_S: the average of $\left\|\hat{\theta}_S-\theta_{0S}\right\|_E$ over the 100 replications. The second one is the mean standard error for the unimportant regressors, MSE_N: the average of $\left\|\hat{\theta}_N-\theta_{0N}\right\|_E$ over the 100 replications. The third one is the average number of correctly selected nonzero coefficients, the true posstive (TP). And the last one is the average number of incorrectly selected coefficients, the false positive (FP).

3.6.1 Endogeneity in Both Important and Unimportant Regressors

The m endogenous variables are $(X_1, X_2, X_3, X_6, ..., X_{m+2})'$ with $m = 10$ or $m = 50$, and the other variables are exogenous. Thus three of the important regressors $(X_1, X_2, X_3)'$ are endogenous, and two of them, $(X_4, X_5)'$, are exogenous.

As we see in Table 3.1, when λ increases, which shows more power on variable selection, FP is decreasing in both cases ($p = 50$ and $p = 100$), and so is TP. Also, TP deviates very little from 5, and FP is not large compared to p. The MSE_S is increasing, and MSE_N is decreasing, which is consistent with the changes in TP and FP, and they both remain at low levels. It is worth noting that as p increases, none of the four measurements gets worse by much, which shows the power of the SFGMM estimator on variable selection under high dimensions.

Table 3.1. Endogeneity in Both Important and Unimportant Regressors.

	$p = 50$,	$m = 10$		$p = 100$,	$m = 50$	
	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$
MSE_S	0.2338	0.2961	0.3832	0.1836	0.2880	0.3715
MSE_N	0.1055	0.0594	0.0519	0.0554	0.0527	0.0475
TP	4.9400	4.8900	4.8400	4.9400	4.8700	4.8100
FP	4.7100	3.0900	2.3000	6.7500	4.7300	3.9000

3.6.2 Endogeneity in Only Unimportant Regressors

In this case, $m = 10$ or $m = 50$, and all the endogeneity lies in the unimportant regressors. Table 3.2 shows that all the measurements perform better than when there is endogeneity in both kinds of regressors. As λ increases, FP decreases to a very small number and TP stays at 5. MSE_N decreases, while MSE_S doesn't increase by much, and both of them stay at low levels. As p increases, none of the four measurements changes by much, which shows the reliability of the SFGMM estimator on variable selection in high dimensions.

The reason why the results shown in Table 3.2 are better than those in Table 3.1 is that all the endogeneity comes from the unimportant regressors, so it doesn't affect the estimation in a serious way.

Table 3.2. Endogeneity in Only Unimportant Regressors.

	$p = 50,\ m = 10$			$p = 100,\ m = 50$		
	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$
MSE_S	0.0943	0.1006	0.1074	0.0845	0.0934	0.1002
MSE_N	0.0646	0.0487	0.0380	0.0431	0.0373	0.0319
TP	5	5	5	5	5	5
FP	4.4700	2.9800	2.1000	7.0700	4.6700	3.2700

Chapter 3, in part is currently being prepared for submission for publication of the material. Sun, Zhenting. The dissertation author was the primary investigator and author of this material.

Appendix A

Proofs for Chapter 1

A.1 Some Useful Lemmas

Lemma A.1.1 *Let $\mathbb{D}_n(\omega) \subset \mathbb{D}$ for all $\omega \in \Omega$. Let $\mathbb{F}_n(\omega) = \{f : f : \mathbb{D}_n(\omega) \to \mathbb{E}\}$. Let $g_n(\omega) \in \mathbb{F}_n(\omega)$ such that, for almost surely $\omega \in \Omega$, if $x_n \to x$ with $x_n \in \mathbb{D}_n(\omega)$ and $x \in \mathbb{D}_0$, then $g_n(\omega)(x_n) \to g(x)$, where $\mathbb{D}_0 \subset \mathbb{D}$ and $g : \mathbb{D}_0 \to \mathbb{E}$. Let $X_n(\omega)$ be maps with values in $\mathbb{D}_n(\omega)$, let X be Borel measurable and separable, and take values in $\mathbb{D}_0$. Then*

$$X_n \rightsquigarrow X \text{ implies that } g_n(X_n) \rightsquigarrow g(X).$$

Proof of Lemma A.1.1.

The proof is an extension of that of Theorem 1.11.1 in van der Vaart & Wellner (1996). Let F be a closed set in $\mathbb{E}$. Then almost surely

$$\cap_n^{\infty} \overline{\cup_{m=n}^{\infty} g_m^{-1}(F)} \subset g^{-1}(F) \cup (\mathbb{D} - \mathbb{D}_0).$$

For every fixed k, by Portmanteau Theorem,

$$\limsup \mathbb{P}^*(g_n(X_n) \in F) \leq \limsup \mathbb{P}^*\left(X_n \in \overline{\cup_{m=k}^{\infty} g_m^{-1}(F)}\right) \leq \mathbb{P}\left(X \in \overline{\cup_{m=k}^{\infty} g_m^{-1}(F)}\right),$$

where $\mathbb{P}^*$ denotes the outer probability. If we let $k \to \infty$,

$$\begin{aligned}
&\mathbb{P}\left(X \in \overline{\cup_{m=k}^{\infty} g_m^{-1}(F)}\right) \to \mathbb{P}\left(X \in \cap_{k=1}^{\infty} \overline{\cup_{m=k}^{\infty} g_m^{-1}(F)}\right) \\
=&\mathbb{P}\left(X \in \cap_{k=1}^{\infty} \overline{\cup_{m=k}^{\infty} g_m^{-1}(F)}, \cap_{n}^{\infty} \overline{\cup_{m=k}^{\infty} g_m^{-1}(F)} \subset g^{-1}(F) \cup (\mathbb{D} - \mathbb{D}_0)\right) \leq \mathbb{P}\left(X \in g^{-1}(F)\right).
\end{aligned}$$

■

Lemma A.1.2 *Let $\mathbb{D}$ and $\mathbb{E}$ be metrizable topological vector spaces and r_n constants with $r_n \to \infty$. Let $\hat{\phi}_n : \Omega \to \mathbb{D}_{\mathscr{F}} \subset \mathbb{D}$ be a random variable which takes values in $\mathbb{D}_{\mathscr{F}} \subset \mathbb{D}$. Let $\mathscr{F} : \mathbb{D}_{\mathscr{F}} \to \mathbb{E}$ satisfy that for almost surely $\omega \in \Omega$,*

$$r_n(\mathscr{F}(\hat{\phi}_n + r_n^{-1} h_n) - \mathscr{F}(\hat{\phi}_n)) \to \mathscr{F}'_{\phi}(h),$$

for every converging sequence h_n with $\hat{\phi}_n + r_n^{-1} h_n \in \mathbb{D}_{\mathscr{F}}$ for all n and $h_n \to h \in \mathbb{D}_0 \subset \mathbb{D}$ and some arbitrary map $\mathscr{F}'_{\phi}$ on $\mathbb{D}_0$. If $X_n : \Omega \to \mathbb{D}_{\mathscr{F}}$ are maps with $r_n(X_n - \hat{\phi}_n) \rightsquigarrow X$, where X is Borel measurable and separable, and takes its values in $\mathbb{D}_0$, then $r_n(\mathscr{F}(X_n) - \mathscr{F}(\hat{\phi}_n)) \rightsquigarrow \mathscr{F}'_{\phi}(X)$. Moreover, if $\mathscr{F}'_{\phi}$ is continuous on the whole of $\mathbb{D}$, then $r_n(\mathscr{F}(X_n) - \mathscr{F}(\hat{\phi}_n)) - \mathscr{F}'_{\phi}(r_n(X_n - \hat{\phi}_n))$ converges to zero in outer probability.

Proof of Lemma A.1.2.

The proof is an extension of that of Theorem 3.9.5 in van der Vaart & Wellner (1996). We define a map $g_n(h) = r_n(\mathscr{F}(\hat{\phi}_n + r_n^{-1} h) - \mathscr{F}(\hat{\phi}_n))$. For any $\omega \in \Omega$, g_n is defined on the domain $\mathbb{D}_n(\omega) = \{h : \hat{\phi}_n(\omega) + r_n^{-1} h \in \mathbb{D}_{\mathscr{F}}\}$ and by assumptions $g_n(h_n) \to \mathscr{F}'_{\phi}(h)$ a.s. for all $h_n \to h \in \mathbb{D}_0$. Then by Lemma A.1.1,

$$g_n(r_n(X_n - \hat{\phi}_n)) \rightsquigarrow \mathscr{F}'_{\phi}(X).$$

Now suppose $\mathscr{F}'_{\phi}$ is continuous on the whole of $\mathbb{D}$ and we let $f_n(h_1, h_2) = (g_n(h_1), \mathscr{F}'_{\phi}(h_2))$ such that for any $\omega \in \Omega$, $(h_1, h_2) \in \mathbb{D}_n(\omega) = \{(h_1, h_2) : \hat{\phi}_n(\omega) + r_n^{-1} h_1 \in \mathbb{D}_{\mathscr{F}}, h_2 \in \mathbb{D}\}$. By

Lemma A.1.1 again,

$$\begin{pmatrix} r_n(\mathscr{F}(X_n)-\mathscr{F}(\hat{\phi}_n)) \\ \mathscr{F}'_{\phi}(r_n(X_n-\hat{\phi}_n)) \end{pmatrix} = \begin{pmatrix} g_n \\ \mathscr{F}'_{\phi} \end{pmatrix} (r_n(X_n-\hat{\phi}_n)) \rightsquigarrow \begin{pmatrix} \mathscr{F}'_{\phi} \\ \mathscr{F}'_{\phi} \end{pmatrix} (X).$$

Then by continuous mappting theorem,

$$r_n(\mathscr{F}(X_n)-\mathscr{F}(\hat{\phi}_n)) - \mathscr{F}'_{\phi}(r_n(X_n-\hat{\phi}_n)) \rightsquigarrow 0.$$

■

Lemma A.1.3 *$\mathscr{H}$ is complete in $L^2(R)$ under norm $\|\cdot\|_{L^2(R)}$.*

Proof of Lemma A.1.3.

Suppose there is a Cauchy sequence $\{h_n\} \subset \mathscr{H}$ such that $\|h_n - h_m\|_{L^2(R)} \to 0$ as $n,m \to \infty$, then there is $h \in L^2(R)$ such that $\|h_n - h\|_{L^2(R)} \to 0$ because $L^2(R)$ is complete. Remember by (1.7), $h_n = (-1)^{d_n} \cdot 1_{B_n \times \{d_n\}}$, where B_n is a closed interval in $\mathbb{R}$ and $d_n \in \{0,1\}$.

(1) If for every $N > 0$, there are $n,m > N$ such that $d_n \neq d_m$, then

$$\begin{aligned} \|h_n - h_m\|^2_{L^2(R)} &= \int \left| (-1)^{d_n} \cdot 1_{B_n \times \{d_n\}} - (-1)^{d_m} \cdot 1_{B_m \times \{d_m\}} \right|^2 dR \\ &= \int 1_{B_n \times \{d_n\}} dR + \int 1_{B_m \times \{d_m\}} dR \to 0. \end{aligned}$$

This implies $\int 1_{B_n \times \{d_n\}} dR \to 0$ as $n \to \infty$. Thus we can find $B = \{a\}$ for some $a \in \mathbb{R}$ such that $\mathbb{P}(Y \in B) = 0$ and $h_n \to 1_{B \times \{0\}} \in \mathscr{H}$.

(2) There are $d, N_d > 0$ such that for all $m,n > N_d$, $d_n = d_m = d$, then

$$\begin{aligned} \|h_n - h_m\|^2_{L^2(R)} &= \int \left| 1_{B_n \times \{d\}} - 1_{B_m \times \{d\}} \right|^2 dR \\ &= \int 1_{B_n \setminus B_m \times \{d\}} dR + \int 1_{B_m \setminus B_n \times \{d\}} dR \to 0. \end{aligned}$$

It is possible that $\int 1_{B_n \times \{d\}} dR \to 0$ and then we can find $B = \{a\}$ for some $a \in \mathbb{R}$ such

that $\mathbb{P}(Y \in B) = 0$ and $h_n \to 1_{B\times\{0\}} \in \mathscr{H}$.

Now let $h_n = (-1)^d \cdot 1_{B_n\times\{d\}}$. Suppose there is $\varepsilon > 0$ such that for all $N_\varepsilon > 0$, there is $n > \max\{N_\varepsilon, N_d\}$ such that $\|h_n\|_{L^2(R)} > \varepsilon$. For all $\delta_1 \ll \varepsilon$, $\exists N_1 > N_d$ such that $\|h_n - h_m\|_{L^2(R)} < \delta_1$ for all $m, n > N_1$. And we can find $n_1 > N_1$ such that $\|h_{n_1}\|_{L^2(R)} > \varepsilon$ and $\|h_n - h_{n_1}\|_{L^2(R)} < \delta_1$ for all $n > n_1$. Now let $0 < \delta_2 < \delta_1$, there is $N_2 > N_1$ such that $\|h_n - h_m\|_{L^2(R)} < \delta_2$ for all $m, n > N_2$. Then we can find $n_2 > N_2$ such that $\|h_{n_2}\|_{L^2(R)} > \varepsilon$ and $\|h_n - h_{n_2}\|_{L^2(R)} < \delta_2$ for all $n > n_2$. In this way, we can find a sequence $\{B_{n_k}\}_k$ with $\left\|(-1)^d \cdot 1_{B_{n_k}\times\{d\}}\right\|_{L^2(R)} > \varepsilon$ and $\|h_n - h_{n_k}\|_{L^2(R)} < \delta_k$ for all $n > n_k$ with $\delta_k \downarrow 0$. We let $B^\infty = \overline{\cup_{j=1}^\infty \cap_{k=j}^\infty B_{n_k}}$. Because B_{n_k} is always a closed interval, B^∞ is a closed interval. Now we have

$$\left\|h_{n_k} - (-1)^d \cdot 1_{B^\infty\times\{d\}}\right\|_{L^2(R)} \to 0 \text{ as } k \to \infty,$$

because $\|h_n - h_{n_k}\|_{L^2(R)} < \delta_k$ for all $n > n_k$.

Last, we have

$$\left\|h_n - (-1)^d \cdot 1_{B^\infty\times\{d\}}\right\|_{L^2(R)} \le \|h_n - h_{n_k}\|_{L^2(R)} + \left\|h_{n_k} - (-1)^d \cdot 1_{B^\infty\times\{d\}}\right\|_{L^2(R)} \to 0.$$

Clearly $(-1)^d \cdot 1_{B^\infty\times\{d\}} \in \mathscr{H}$. ■

Lemma A.1.4 *$\mathscr{H}$ is a VC class with VC-index $V(\mathscr{H}) = 3$.*

Proof of Lemma A.1.4.

All the functions $h \in \mathscr{H}$ take the form $h = -1_{B\times\{1\}}$ or $h = 1_{B\times\{0\}}$, where B is a closed interval. If $h = -1_{B\times\{1\}}$, the subgraph of h is

$$C_{1B} = \left\{(y, w, t) \subset \mathscr{Y} \times \{0,1\} \times \mathbb{R} : t < -1_{B\times\{1\}}(y, w)\right\}.$$

If $h = 1_{B\times\{0\}}$, the subgraph of h is

$$C_{0B} = \left\{(y,w,t) \subset \mathscr{Y} \times \{0,1\} \times \mathbb{R} : t < 1_{B\times\{0\}}(y,w)\right\}.$$

We let $\mathscr{C} = \{C_{1B}, C_{0B} : B \text{ is a closed interval in } \mathbb{R}\}$.

Suppose any two different points $a_1 = (y_1, w_1, t_1), a_2 = (y_2, w_2, t_2) \in \mathscr{Y} \times \{0,1\} \times \mathbb{R}$, with $y_1 < y_2$, $w_1 = w_2 = 0$ and $0 \leq t_1, t_2 < 1$. Then there is a point $\bar{y} \in (y_1, y_2)$. Let $B_1 = [y_1, \bar{y}]$, $B_2 = [\bar{y}, y_2]$ and $B_3 = [y_1, y_2]$. Now we have

$$\{a_1\} = C_{0B_1} \cap \{a_1, a_2\}, \{a_2\} = C_{0B_2} \cap \{a_1, a_2\}, \{a_1, a_2\} = C_{0B_3} \cap \{a_1, a_2\}.$$

Suppose any three different points $a_1 = (y_1, w_1, t_1)$, $a_2 = (y_2, w_2, t_2)$, $a_3 = (y_3, w_3, t_3)$ $\in \mathscr{Y} \times \{0,1\} \times \mathbb{R}$ and we have a set $\{a_1, a_2, a_3\}$. Without loss of generality, suppose $t_1 \leq t_2 \leq t_3 < 1$.

(1) Suppose $t_1 \geq 0$. In this case, it should hold that $w_1 = w_2 = w_3 = 0$ so that $\{a_i\}$ can be picked up for each i. Without loss of generality, suppose $y_1 \leq y_2 \leq y_3$. If we want to pick out $\{a_1, a_3\}$, we need to find a closed interval B such that $y_1, y_3 \in B$ and then $a_1, a_3 \in C_{0B}$. However, $a_2 \in C_{0B}$.

(2) Suppose $t_1 < 0$, $t_2 \geq 0$. Then $w_2 = w_3 = 0$ in order to pick out $\{a_i\}$ for each $i = 2, 3$ by using C_{0B} for some closed interval B. But in this case, $\mathscr{C}$ cannot only pick out $\{a_2\}, \{a_3\}$ or $\{a_2, a_3\}$, since for every closed interval B, $a_1 \in C_{0B}$.

(3) Suppose $t_1, t_2 < 0$, $t_3 \geq 0$. Then we need $w_3 = 0$ in order to pick out $\{a_3\}$ by using C_{0B} for some closed interval B. In this case, $\mathscr{C}$ cannot only pick out a_3, since for every closed interval B, $a_1, a_2 \in C_{0B}$.

For $t_1, t_2, t_3 < 0$, for every closed interval B, $a_1, a_2, a_3 \in C_{0B}$. If we want $\mathscr{C}$ to pick out $\{a_k\}$ or $\{a_k, a_{k'}\}$ for some different $k, k' = 1, 2, 3$, we need to use C_{1B}. If $w_k = 0$, then for every B, $a_k \in C_{1B}$. Thus we consider $w_1 = w_2 = w_3 = 1$.

(4) Suppose $-1 \leq t_1, t_2, t_3 < 0$. Now without loss of generality, we assume $y_1 \leq y_2 \leq y_3$. But now if we want to pick out $\{a_2\}$, we need to find a closed interval B such that $y_1, y_3 \in B$ but $y_2 \notin B$. Not possible.

(5) Suppose $t_k < -1$ for some $k = 1, 2, 3$. In this case, $a_k \in C_{1B}$ for all closed B.

Therefore, with the discussion of all cases above, we conclude that $\mathscr{H}$ is a VC class with VC-index $V(\mathscr{H}) = 3$. ■

Lemma A.1.5 *For any probability measure R on $\mathscr{Y} \times \{0,1\}$, $\mathscr{H}$ is compact under $\|\cdot\|_{L^r(R)}$ for $r \in \mathbb{N}$.*

Proof of Lemma A.1.5.

Let $N(\varepsilon, \mathscr{H}, L^r(R))$ denote the covering number for $\mathscr{H}$ with all $\varepsilon > 0$.

Since $\mathscr{H}$ is a VC class by Lemma A.1.4, with envelope function $F = 1$ and $r > 1$, by Theorem 2.6.7 in van der Vaart & Wellner (1996), for any probability measure R,

$$N\left(\varepsilon \|F\|_{L^r(R)}, \mathscr{H}, L^r(R)\right) \leq KV(\mathscr{H})(16e)^{V(\mathscr{H})}(1/\varepsilon)^{r(V(\mathscr{H})-1)},$$

for a universal constant K and $0 < \varepsilon < 1$. Then by Lemma A.1.3, $\mathscr{H}$ is compact in $\mathscr{H}$. ■

Lemma A.1.6 *For any probability measure R in $\mathscr{P}$, $\mathscr{H}$ is a R-Donsker.*

Proof of Lemma A.1.6.

For every $\delta > 0$ and $R \in \mathscr{P}$, define $\mathscr{H}_{\delta,R} = \left\{h - g : h, g \in \mathscr{H}, \|h - g\|_{L^2(R)} < \delta\right\}$ and $\mathscr{H}_\infty^2 = \left\{(h-g)^2 : h, g \in \mathscr{H}\right\}$. First we show that $\mathscr{H}_{\delta,R}$ is R-measurable for all $R \in \mathscr{P}$. Similarly to the construction of $\mathscr{H}$ in (1.7), we construct another function space by

$$\mathscr{H}_q = \left\{h = (-1)^d \cdot 1_{B \times \{d\}} : B = [a,b] \text{ with rational numbers } a, b \text{ and } d \in \{0,1\}\right\}.$$

Let $\mathbb{Q}$ denote the set of all rational numbers. Since $\mathbb{Q}$ is countable and therefore the set of ordered

pairs of elements in $\mathbb{Q}$ is countable, $\mathscr{H}_q$ is countable. Now define

$$\mathscr{H}_{q\delta,R} = \left\{ h - g : h, g \in \mathscr{H}_q, \|h-g\|_{L^2(R)} < \delta \right\}.$$

Clearly, $\mathscr{H}_{q\delta,R}$ is a countable subset of $\mathscr{H}_{\delta,R}$. For any $h \in \mathscr{H}$, there is a sequence $h_m \in \mathscr{H}_q$ such that $h_m(x) - h(x)$ for every x because $\mathbb{Q}$ is dense in $\mathbb{R}$. For example, if $h = 1_{\left[\sqrt{2},\sqrt{3}\right]}$, we can use $h_m = 1_{[a_m,b_m]}$ to approximate h such that $a_m \uparrow \sqrt{2}$, $b_m \downarrow \sqrt{3}$ and $a_m, b_m \in \mathbb{Q}$. Also, for all $\delta > 0$, if $h, g \in \mathscr{H}_q$ and $\|h-g\|_{L^2(R)} < \delta$, $\|h_m - g_m\|_{L^2(R)} < \delta$ for large m. By Example 2.3.4 in van der Vaart & Wellner (1996), $\mathscr{H}_{\delta,R}$ is R-measurable and this is true for all $\delta > 0$. Similarly, $\mathscr{H}_\infty^2$ is R-measurable.

By the construction of $\mathscr{H}$, clearly, $F = 1$ is a measurable envelope function with

$$\int F^2 dR < \infty.$$

By Lemma A.1.4, $\mathscr{H}$ is a VC class with VC-index $V(\mathscr{H}) = 3$. By Theorem 2.6.7 in van der Vaart & Wellner (1996), for every probability measure R, the covering number for every integer r satisfies

$$N(\varepsilon, \mathscr{H}, L^r(R)) \leq K \cdot 2 \cdot (16e)^3 \left(\frac{1}{\varepsilon}\right)^{2r}$$

for a universal constant K and $0 < \varepsilon < 1$. Also, let $\mathscr{Q}$ denote the set of finitely discrete probability measures, for all $H \in \mathscr{Q}$, when $\varepsilon \geq 2$,

$$N\left(\varepsilon \|F\|_{L^2(H)}, \mathscr{H}, L^2(H)\right) = N\left(\varepsilon, \mathscr{H}, L^2(H)\right) = 1,$$

which implies

$$
\begin{aligned}
&\int_0^{\infty} \sup_{H\in\mathscr{Q}} \sqrt{\log N\left(\varepsilon \|F\|_{L^2(H)}, \mathscr{H}, L^2(H)\right)} d\varepsilon \\
\leq &\int_0^{2} \sup_{H\in\mathscr{Q}} \sqrt{\log N\left(\varepsilon \|F\|_{L^2(H)}, \mathscr{H}, L^2(H)\right)} d\varepsilon \\
\leq &\int_1^{2} \sup_{H\in\mathscr{Q}} \sqrt{\log N\left(\varepsilon \|F\|_{L^2(H)}, \mathscr{H}, L^2(H)\right)} d\varepsilon + \int_0^{1} \sup_{H\in\mathscr{Q}} \sqrt{\log N\left(\varepsilon \|F\|_{L^2(H)}, \mathscr{H}, L^2(H)\right)} d\varepsilon \\
\leq &C_1 + \int_0^{1} \sqrt{\log\left\{K\cdot 2\cdot (16e)^3 \left(\frac{1}{\varepsilon}\right)^4\right\}} d\varepsilon < \infty,
\end{aligned}
$$

where C_1 is a large positive number and the third inequality follows from Theorem 2.6.7 in van der Vaart & Wellner (1996). The result follows from Theorem 2.5.2 in van der Vaart & Wellner (1996). ■

Assumption A.1.1 *Suppose it holds that:*

(i) Probability measures in $\mathscr{P}$ are nondegenerate and have a common dominating measure μ for the coordinate, where μ is the Lebesgue measure, a point mass measure with finite support points, or their mixture. The density functions $p = \frac{dR}{d\mu}$ are bounded uniformly over $\mathscr{P}$, that is, there is $M < \infty$ such that for all $R \in \mathscr{P}$, $p(y,d) \leq M$ for μ-almost every $y \in \mathscr{Y}$ and $d = 0, 1$.

(ii) The set $\mathscr{P}$ is uniformly tight, i.e. for any $\varepsilon > 0$, there is a compact set $K \subset \mathscr{Y} \times \{0,1\}$ such that

$$\sup_{R\in\mathscr{P}} R(K^c) < \varepsilon.$$

Lemma A.1.7 *Suppose Assumption A.1.1 holds. Let $\{R^m \in \mathscr{P} : m = 1, 2, \cdots\}$ be a sequence of probability measures that converges weakly to $R \in \mathscr{P}$. Let R_m^m denote the empirical measure of a iid sample $\{X_{mi}\}_{i=1}^m$ from distribution R^m with $R_m^m = m^{-1}\sum_{i=1}^m \delta_{X_i}$, where δ_{X_i} is the Dirac*

measure at the observation X_i. Construct the $\mathscr{H}$-indexed empirical process $G_{P^k,m}$ by

$$G_{m,R^m} = \sqrt{m}\left(R_m^m - R^m\right),$$

that is, for all $h \in \mathscr{H}$,

$$G_{m,R^m}(h) = \sqrt{m}\left(R_m^m - R^m\right)(h) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \left(h(X_i) - \int h dR^m \right).$$

Then G_{m,R^m} converges weakly to the R-Brownian bridge G_R.

Proof of Lemma A.1.7.

For every $\delta > 0$ and $R \in \mathscr{P}$, define $\mathscr{H}_{\delta,P} = \left\{ h - g : h, g \in \mathscr{H}, \|h-g\|_{L^2(P)} < \delta \right\}$ and $\mathscr{H}_\infty^2 = \left\{ (h-g)^2 : h, g \in \mathscr{H} \right\}$. Similarly to the proof of Lemma A.1.6, we can show that $\mathscr{H}_{\delta,R}$ is R-measurable and $\mathscr{H}_\infty^2$ is R-measurable. By the construction of $\mathscr{H}$, clearly $F = 1$ is a measurable envelope function with

$$\sup_{R \in \mathscr{P}} \int F^2 1\{F > M\} dR \to 0, \text{ as } M \to \infty.$$

Also, similarly to Lemma A.1.6, it holds that

$$\int_0^\infty \sup_{H \in \mathscr{Q}} \sqrt{\log N\left(\varepsilon \|F\|_{L^2(H)}, \mathscr{H}, L^2(H)\right)} d\varepsilon < \infty,$$

where $\mathscr{Q}$ denotes the set of finitely discrete probability measures. Now by Theorem 2.8.3 in van der Vaart & Wellner (1996), $\mathscr{H}$ is Donsker and pre-Gaussian uniformly in $R \in \mathscr{P}$.

For every $R \in \mathscr{P}$, we define a semimetric ρ_R by

$$\rho_R(h_1, h_2) = \|(h_1 - h_2) - R(h_1 - h_2)\|_{L^2(R)} \tag{A.1}$$

for all $h_1, h_2 \in \mathscr{H}$. Then by Lemma B.2 in Kitagawa (2015), under Assumption A.1.1,

$$\begin{aligned}\sup_{h,g\in\mathscr{H}} \left|\rho_{R^m}(h,g) - \rho_R(h,g)\right|^2 &\leq \sup_{h,g\in\mathscr{H}} \left|\rho^2_{R^m}(h,g) - \rho^2_R(h,g)\right| \\ &\leq \sup_{A\in\mathscr{B}(\mathscr{Y}\times\{0,1\})} \left|\left(R^m - R\right)(A)\right| \to 0,\end{aligned}$$

where $\mathscr{B}(\mathscr{Y}\times\{0,1\})$ is the Borel σ-algebra on $\mathscr{Y}\times\{0,1\}$. Also, for all $\varepsilon > 0$,

$$\limsup_{m\to\infty} \int F \cdot 1\left\{F \geq \varepsilon\sqrt{m}\right\} dR^m = 0.$$

By Lemma 2.8.7 in van der Vaart & Wellner (1996), $G_{m,R^m} \rightsquigarrow \mathbb{G}_R$, where $\mathbb{G}_R$ is the R-Brownian bridge. ■

Lemma A.1.8 *$\mathscr{H}$ is Glivenko-Cantelli uniformly in $R \in \mathscr{P}$.*

Proof of Lemma A.1.8.

Similarly to the proof of Lemma A.1.6, $\mathscr{H}$ is R-measurable for every $R \in \mathscr{P}$. And with $F = 1$ being an envelope function of $\mathscr{H}$,

$$\lim_{M\to\infty} \sup_{R\in\mathscr{P}} \int F \cdot 1\{F > M\} dR = 0.$$

By Lemma A.1.4, $\mathscr{H}$ is a VC class with VC-index $V(\mathscr{H}) = 3$. By Theorem 2.6.7 in van der Vaart & Wellner (1996), for every $r \geq 1$ and every probability measure H,

$$N\left(\varepsilon \|F\|_{L^r(H)}, \mathscr{H}, L^r(H)\right) \leq K \cdot 2 \cdot (16e)^3 \left(\frac{1}{\varepsilon}\right)^{2r}$$

for a universal constant K and $0 < \varepsilon < 1$. Then

$$\sup_{H\in\mathscr{Q}_n} \log N\left(\varepsilon \|F\|_{L^r(H)}, \mathscr{H}, L^r(H)\right) = o(n),$$

where $\mathscr{Q}_n$ is the collection of all possible realizations of empirical measures of n observations. By Theorem 2.8.1 in van der Vaart & Wellner (1996), $\mathscr{H}$ is Glivenko-Cantelli uniformly in $R \in \mathscr{P}$. ■

Definition A.1.1 *Let $\mathbb{D}$ and $\mathbb{E}$ be Banach spaces, and $\mathscr{F} : \mathbb{D}_{\mathscr{F}} \subset \mathbb{D} \to \mathbb{E}$. The map $\mathscr{F}$ is said to be Hadamard directionally differentiable at $\phi \in \mathbb{D}_{\mathscr{F}}$ tangentially to a set $\mathbb{D}_0 \subset \mathbb{D}$, if there is a continuous map $\mathscr{F}'_\phi : \mathbb{D}_0 \to \mathbb{E}$ such that*

$$\lim_{n\to\infty} \left\| \frac{\mathscr{F}(\phi + t_n \psi_n) - \mathscr{F}(\phi)}{t_n} - \mathscr{F}'_\phi(\psi) \right\|_{\mathbb{E}} = 0,$$

for all sequences $\{\psi_n\} \subset \mathbb{D}$ and $\{t_n\} \subset \mathbb{R}_+$ such that $t_n \downarrow 0$, $\psi_n \to \psi \in \mathbb{D}_0$ as $n \to \infty$ and $\phi + t_n \psi_n \in \mathbb{D}_{\mathscr{F}}$ for all n.

Lemma A.1.9 *Let $R = P/2 + Q/2$. The map defined in* (2.4) *is Hadamard directionally differentiable at ϕ defined by* (1.8) *tangentially to $C(\mathscr{H})$ and $\mathscr{S}'_\phi : C(\mathscr{H}) \to \mathbb{R}$ satisfies*

$$\mathscr{S}'_\phi(\psi) = \sup_{h \in \Psi_{\mathscr{H}}} \psi(h), \psi \in C(\mathscr{H}), \tag{A.2}$$

where $\Psi_{\mathscr{H}} = \arg\max_{h \in \mathscr{H}} \phi(h)$.

Proof of Lemma A.1.9.

By Lemma A.1.5, $\mathscr{H}$ is compact under $\|\cdot\|_{L^2(R)}$. First, we show ϕ defined in (1.8) is in $C(\mathscr{H})$. For all $h_1, h_2 \in \mathscr{H}$,

$$\|h_1 - h_2\|^2_{L^2(R)} = \int |h_1 - h_2|^2 d\left(\frac{1}{2}P + \frac{1}{2}Q\right) = \frac{1}{2}\int |h_1 - h_2|^2 dP + \frac{1}{2}\int |h_1 - h_2|^2 dQ.$$

Also,

$$
\begin{aligned}
&|\phi(h_1)-\phi(h_2)| \\
&=\left|\left(E\left[h_1\left(Y^1,D^1\right)\right]-E\left[h_1\left(Y^0,D^0\right)\right]\right)-\left(E\left[h_2\left(Y^1,D^1\right)\right]-E\left[h_2\left(Y^0,D^0\right)\right]\right)\right| \\
&\leq \int |h_1-h_2|\,dP+\int |h_1-h_2|\,dQ \leq \left(\int |h_1-h_2|^2\,dP\right)^{1/2}+\left(\int |h_1-h_2|^2\,dQ\right)^{1/2}.
\end{aligned}
$$

It is clear that when $\|h_1-h_2\|^2_{L^2(R)} \to 0$, $|\phi(h_1)-\phi(h_2)| \to 0$. This implies $\phi \in C(\mathscr{H})$. Then the result follows Lemma B.1 in Fang & Santos (2014). ■

Remark A.1.1 *The map defined in* (2.4) *is a supremum over all* $h \in \mathscr{H}$. *We will not use the Hadamard directional derivative of* $\mathscr{S}$ *directly, but it provides an idea for us to obtain the asymptotic distribution of the test statistic and apply the bootstrap method for Hadamard directional differentiable maps.*

Lemma A.1.10 $N(\varepsilon, \mathscr{H}_K \times \mathscr{G}, \rho_P) = O\left(1/\varepsilon^4\right)$ *as* $\varepsilon \to 0$.

Proof of Lemma A.1.10.

Since $\mathscr{H}$ is a VC class by Lemma A.1.4, with envelope function $F=1$ and $r>1$, by Theorem 2.6.7 in van der Vaart & Wellner (1996), we have for every probability measure R,

$$
N\left(\varepsilon \|F\|_{L^r(R)}, \mathscr{H}, L^r(R)\right) \leq K3(16e)^3(1/\varepsilon)^{2r}
$$

for a universal constant K and $0<\varepsilon<1$. It is not hard to see that $F=1$ is also an envelope function of $\mathscr{H}_K$ and for any $r \geq 1$,

$$
N(\varepsilon, \mathscr{H}_K, L^r(P)) = N(\varepsilon, \mathscr{H}, L^r(P_{YD})),
$$

Where P_{YD} is the probability measure on $\mathbb{R}^2$ for (Y,D) induced by P.

By the construction of $\mathscr{H}_K \times \mathscr{G}$ and metric ρ_P

$$N(\varepsilon, \mathscr{H}_K \times \mathscr{G}, \rho_P) \leq \max\left\{N\left(\frac{\varepsilon}{3}, \mathscr{H}_K, L^2(P)\right), N\left(\frac{\varepsilon}{3}, \mathscr{G}_K, L^2(P)\right)\right\}.$$

By definition, $\mathscr{G}_K = \left\{1_{\mathbb{R}\times\{0,1\}\times\{z_k\}} : k = 1, 2, \ldots, K\right\}$. It is easy to show that the $\mathscr{I}_K = \left\{1_{\{z_k\}}\right\}_k$ is a VC class of functions with VC-index equal to 2. So by Theorem 2.6.7 in van der Vaart & Wellner (1996), for every ε and every $r \geq 1$, and for every probability measure R,

$$N(\varepsilon, \mathscr{G}_K, L^r(R)) = N(\varepsilon, \mathscr{I}_K, L^r(R_Z)) \leq C(1/\varepsilon)^r,$$

for some constant $C > 0$, where R_Z is the probability measure on $\mathbb{R}$ for Z induced by R. This implies

$$N(\varepsilon, \mathscr{H}_K \times \mathscr{G}, \rho_P) = O\left(\frac{1}{\varepsilon^4}\right) \text{ as } \varepsilon \to 0.$$

■

Lemma A.1.11 *$\mathscr{H}_K \times \mathscr{G}$ is complete under ρ_P.*

Proof of Lemma A.1.11.

Similarly to the proof of Lemma A.1.3, it can be shown that $\mathscr{H}_K$ and $\mathscr{G}_K$ are both complete under $\|\cdot\|_{L^2(P)}$, which implies $\mathscr{H}_K \times \mathscr{G}$ is complete under ρ_P. ■

Lemma A.1.12 *$\mathscr{H}_K$ and $\mathscr{G}_K$ are Glivenko-Cantelli uniformly in $R \in \mathscr{P}_3$.*

Proof of Lemma A.1.12.

Similarly to the proof of Lemma A.1.6, with $\mathscr{G}$ being a countable set, $\mathscr{H}_K$ and $\mathscr{G}$ are R-measurable for every $R \in \mathscr{P}_3$. And with $F = 1$ being an envelope function of $\mathscr{H}_K$ and $\mathscr{G}$,

$$\lim_{M\to\infty} \sup_{R\in\mathscr{P}_3} \int F \cdot 1\{F > M\} \, dR = 0.$$

By Lemma A.1.4, $\mathscr{H}$ is a VC class with VC-index $V(\mathscr{H}) = 3$. By Theorem 2.6.7 in van der Vaart & Wellner (1996), with $r \geq 1$, we have for every probability measure R,

$$N\left(\varepsilon \|F\|_{L^r(R)}, \mathscr{H}, L^r(R)\right) \leq K \cdot 2 \cdot (16e)^3 \left(\frac{1}{\varepsilon}\right)^{2r}$$

for a universal constant K and $0 < \varepsilon < 1$. As we have shown in Lemma A.1.10 for all $\varepsilon > 0$,

$$N(\varepsilon, \mathscr{H}_K, L^r(R)) = N(\varepsilon, \mathscr{H}, L^r(R_{YD}))$$

for every probability measure R and induced probability measure R_{YD} for (Y, D). Then

$$\sup_{H \in \mathscr{Q}_n} \log N\left(\varepsilon \|F\|_{L^r(H)}, \mathscr{H}_K, L^r(H)\right) = o(n),$$

where $\mathscr{Q}_n$ is the collection of all possible realizations of empirical measures of n observations. By Theorem 2.8.1 in van der Vaart & Wellner (1996), $\mathscr{H}_K$ is Glivenko-Cantelli uniformly in $R \in \mathscr{P}_3$.

Also, as shown in Lemma A.1.10, for every $\varepsilon > 0$ and every $r \in \mathbb{N}$,

$$N(\varepsilon, \mathscr{G}_K, L^r(R)) \leq C(1/\varepsilon)^r,$$

which implies

$$\sup_{H \in \mathscr{Q}_n} \log N\left(\varepsilon \|F\|_{L^r(H)}, \mathscr{G}_K, L^r(H)\right) = o(n),$$

where $\mathscr{Q}_n$ is the collection of all possible realizations of empirical measures of n observations. By Theorem 2.8.1 in van der Vaart & Wellner (1996), $\mathscr{G}_K$ is Glivenko-Cantelli uniformly in $R \in \mathscr{P}_3$. ■

A.2 Results in Sections 1.2 and 1.3

Proof of Lemma 1.2.1.

Suppose $d_{\max}$ exists. Under Assumption 1.2.2, we can define $Y_d = Y_{dz_1} = Y_{dz_2} = \cdots = Y_{dz_K}$ for all $d \in \mathscr{D}_J$. For all $k \leq K-1$, define

$$P_k(B, d_{\max}) = \mathbb{P}(Y \in B, D = d_{\max} | Z = z_k) = \mathbb{P}(Y_{d_{\max}} \in B, D_{z_k} = d_{\max}),$$
$$P_{k+1}(B, d_{\max}) = \mathbb{P}(Y \in B, D = d_{\max} | Z = z_{k+1}) = \mathbb{P}\left(Y_{d_{\max}} \in B, D_{z_{k+1}} = d_{\max}\right).$$

Also, under Assumption 1.2.2(iii),

$$\begin{aligned}\mathbb{P}(Y_{d_{\max}} \in B, D_{z_k} = d_{\max}) &= \sum_j \mathbb{P}\left(Y_{d_{\max}} \in B, D_{z_k} = d_{\max}, D_{z_{k+1}} = d_j\right) \\ &= \mathbb{P}\left(Y_{d_{\max}} \in B, D_{z_k} = d_{\max}, D_{z_{k+1}} = d_{\max}\right)\end{aligned}$$

and

$$\mathbb{P}\left(Y_{d_{\max}} \in B, D_{z_{k+1}} = d_{\max}\right) = \sum_j \mathbb{P}\left(Y_{d_{\max}} \in B, D_{z_k} = d_j, D_{z_{k+1}} = d_{\max}\right).$$

Thus it holds that

$$\begin{aligned}&\mathbb{P}\left(Y_{d_{\max}} \in B, D_{z_{k+1}} = d_{\max}\right) - \mathbb{P}(Y_{d_{\max}} \in B, D_{z_k} = d_{\max}) \\ =&\mathbb{P}\left(Y_{d_{\max}} \in B, D_{z_k} \neq d_{\max}, D_{z_{k+1}} = d_{\max}\right) \geq 0.\end{aligned}$$

Suppose $d_{\min}$ exists. Similarly, under Assumption 1.2.2,

$$P_k(B, d_{\min}) = \mathbb{P}(Y \in B, D = d_{\min} | Z = z_k) = \mathbb{P}\left(Y_{d_{\min}} \in B, D_{z_k} = d_{\min}\right),$$
$$P_{k+1}(B, d_{\min}) = \mathbb{P}(Y \in B, D = d_{\min} | Z = z_{k+1}) = \mathbb{P}\left(Y_{d_{\min}} \in B, D_{z_{k+1}} = d_{\min}\right).$$

Also, under Assumption 1.2.2(iii),

$$\mathbb{P}\left(Y_{d_{\min}} \in B, D_{z_k} = d_{\min}\right) = \sum_j \mathbb{P}\left(Y_{d_{\min}} \in B, D_{z_k} = d_{\min}, D_{z_{k+1}} = d_j\right)$$

and

$$\begin{aligned}\mathbb{P}\left(Y_{d_{\min}} \in B, D_{z_{k+1}} = d_{\min}\right) &= \sum_j \mathbb{P}\left(Y_{d_{\min}} \in B, D_{z_k} = d_j, D_{z_{k+1}} = d_{\min}\right) \\ &= \mathbb{P}\left(Y_{d_{\min}} \in B, D_{z_k} = d_{\min}, D_{z_{k+1}} = d_{\min}\right).\end{aligned}$$

Thus it holds that

$$\begin{aligned}&\mathbb{P}\left(Y_{d_{\min}} \in B, D_{z_k} = d_{\min}\right) - \mathbb{P}\left(Y_{d_{\min}} \in B, D_{z_{k+1}} = d_{\min}\right) \\ =&\mathbb{P}\left(Y_{d_{\min}} \in B, D_{z_k} = d_{\min}, D_{z_{k+1}} \neq d_{\min}\right) \geq 0.\end{aligned}$$

■

Proof of Theorem 1.3.1.

Define $G_m, H_n : \mathscr{H} \to \mathbb{R}$ by

$$\begin{aligned}G_m(h) &= \frac{\sqrt{T_N}}{m} \sum_{i=1}^{m} \left\{h\left(Y_i^1, D_i^1\right) - E\left[h\left(Y^1, D^1\right)\right]\right\}, \\ H_n(h) &= \frac{\sqrt{T_N}}{n} \sum_{i=1}^{n} \left\{h\left(Y_i^0, D_i^0\right) - E\left[h\left(Y^0, D^0\right)\right]\right\},\end{aligned}$$

for every $h \in \mathscr{H}$. Then

$$G_m(h) - H_n(h) = \sqrt{T_N}(\hat{\phi}(h) - \phi(h)).$$

By Lemma A.1.6, $G_m \rightsquigarrow \sqrt{1-\lambda}\,\mathbb{G}_P$ and $H_n \rightsquigarrow \sqrt{\lambda}\,\mathbb{G}_Q$. Notice that

$$\begin{aligned}
G_m(h) =& \frac{\sqrt{T_N}}{m}\sum_{i=1}^{m}\left\{h\left(Y_i^1, D_i^1\right) - E\left[h\left(Y^1, D^1\right)\right]\right\} \\
=& \sqrt{\frac{n}{N}}\left(\frac{1}{\sqrt{m}} - \frac{1}{\sqrt{\lambda N}}\right)\sum_{i=1}^{m}\left\{h\left(Y_i^1, D_i^1\right) - E\left[h\left(Y^1, D^1\right)\right]\right\} \\
&+ \left(\sqrt{\frac{n}{N}} - \sqrt{(1-\lambda)}\right)\frac{1}{\sqrt{\lambda N}}\sum_{i=1}^{m}\left\{h\left(Y_i^1, D_i^1\right) - E\left[h\left(Y^1, D^1\right)\right]\right\} \\
&+ \sqrt{(1-\lambda)}\frac{1}{\sqrt{\lambda N}}\sum_{i=1}^{m}\left\{h\left(Y_i^1, D_i^1\right) - E\left[h\left(Y^1, D^1\right)\right]\right\},
\end{aligned}$$

and it is not hard to show that as $N \to \infty$,

$$\begin{aligned}
\sqrt{\frac{n}{N}}\left(\frac{1}{\sqrt{m}} - \frac{1}{\sqrt{\lambda N}}\right)\sum_{i=1}^{m}\left\{h\left(Y_i^1, D_i^1\right) - E\left[h\left(Y^1, D^1\right)\right]\right\} &\to_p 0, \\
\left(\sqrt{\frac{n}{N}} - \sqrt{(1-\lambda)}\right)\frac{1}{\sqrt{\lambda N}}\sum_{i=1}^{m}\left\{h\left(Y_i^1, D_i^1\right) - E\left[h\left(Y^1, D^1\right)\right]\right\} &\to_p 0.
\end{aligned}$$

Similarly,

$$\begin{aligned}
H_n(h) =& \frac{\sqrt{T_N}}{n}\sum_{i=1}^{n}\left\{h\left(Y_i^0, D_i^0\right) - E\left[h\left(Y^0, D^0\right)\right]\right\} \\
=& \sqrt{\frac{m}{N}}\left(\frac{1}{\sqrt{n}} - \frac{1}{\sqrt{(1-\lambda)N}}\right)\sum_{i=1}^{n}\left\{h\left(Y_i^0, D_i^0\right) - E\left[h\left(Y^0, D^0\right)\right]\right\} \\
&+ \left(\sqrt{\frac{m}{N}} - \sqrt{\lambda}\right)\frac{1}{\sqrt{(1-\lambda)N}}\sum_{i=1}^{n}\left\{h\left(Y_i^0, D_i^0\right) - E\left[h\left(Y^0, D^0\right)\right]\right\} \\
&+ \sqrt{\lambda}\frac{1}{\sqrt{(1-\lambda)N}}\sum_{i=1}^{n}\left\{h\left(Y_i^0, D_i^0\right) - E\left[h\left(Y^0, D^0\right)\right]\right\}
\end{aligned}$$

and

$$\begin{aligned}
\sqrt{\frac{m}{N}}\left(\frac{1}{\sqrt{n}} - \frac{1}{\sqrt{(1-\lambda)N}}\right)\sum_{i=1}^{n}\left\{h\left(Y_i^0, D_i^0\right) - E\left[h\left(Y^0, D^0\right)\right]\right\} &\to_p 0, \\
\left(\sqrt{\frac{m}{N}} - \sqrt{\lambda}\right)\frac{1}{\sqrt{(1-\lambda)N}}\sum_{i=1}^{n}\left\{h\left(Y_i^0, D_i^0\right) - E\left[h\left(Y^0, D^0\right)\right]\right\} &\to_p 0.
\end{aligned}$$

Under Assumption 1.3.1, by Example 1.4.6 and Theorem 1.3.6 in van der Vaart & Wellner (1996),

$$\sqrt{(1-\lambda)}\frac{1}{\sqrt{\lambda N}}\sum_{i=1}^{m}\left\{h\left(Y_i^1,D_i^1\right)-E\left[h\left(Y^1,D^1\right)\right]\right\}$$
$$-\sqrt{\lambda}\frac{1}{\sqrt{(1-\lambda)N}}\sum_{i=1}^{n}\left\{h\left(Y_i^0,D_i^0\right)-E\left[h\left(Y^0,D^0\right)\right]\right\}\rightsquigarrow\sqrt{1-\lambda}\mathbb{G}_P-\sqrt{\lambda}\mathbb{G}_Q,$$

where $\mathbb{G}_P$ and $\mathbb{G}_Q$ are Gaussian processes. Thus, we have

$$\sqrt{T_N}\left(\hat{\phi}-\phi\right)\rightsquigarrow\sqrt{1-\lambda}\mathbb{G}_P-\sqrt{\lambda}\mathbb{G}_Q.$$

Together with Lemma A.1.8, we obtain the marginal weak convergence of $\hat{\phi}/(\xi\vee\hat{\sigma}_N)$.

Next we want to show $\sqrt{T_N}(\hat{\phi}-\phi)/(\xi\vee\hat{\sigma}_N)\rightsquigarrow(\sqrt{1-\lambda}\mathbb{G}_P-\sqrt{\lambda}\mathbb{G}_Q)/(\xi\vee\sigma)$. By Theorems 1.5.4 and 1.5.7 in van der Vaart & Wellner (1996), it suffices to show the marginal convergence, which has been obtained above, plus $\mathscr{H}$ being totally bounded and the sequence $\sqrt{T_N}(\hat{\phi}-\phi)/(\xi\vee\hat{\sigma}_N)$ being asymptotically uniformly equicontinuous, both with respect to some semimetric ρ.

For all $h\in\mathscr{H}$,

$$\begin{aligned}\sqrt{T_N}\frac{\left(\hat{\phi}-\phi\right)(h)}{\xi\vee\hat{\sigma}_N}&=\sqrt{T_N}\frac{\left(P_m(h)-Q_n(h)-P(h)+Q(h)\right)}{\xi\vee\hat{\sigma}_N}\\&=\frac{\sqrt{T_N}\left[P_m(h)-P(h)\right]-\sqrt{T_N}\left[Q_n(h)-Q(h)\right]}{\xi\vee\hat{\sigma}_N}.\end{aligned}$$

Since we have shown that $G_m\rightsquigarrow\sqrt{1-\lambda}\mathbb{G}_P$ and $H_n\rightsquigarrow\sqrt{\lambda}\mathbb{G}_Q$, as illustrated in Section 2.8.2 in van der Vaart & Wellner (1996), $\mathscr{H}$ is totally bounded under semimetrics ρ_P and ρ_Q which are defined in (A.1) for P and Q, and for all $\varepsilon,\eta>0$, there are $\delta_P,\delta_Q>0$ such that

$$\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho_P(h,g)<\delta_P}\left|G_m(h)-G_m(g)\right|>\varepsilon\right)<\frac{\eta}{2}$$

and

$$\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho_Q(h,g)<\delta_Q}|H_n(h)-H_n(g)|>\varepsilon\right)<\frac{\eta}{2},$$

where $\mathbb{P}^*$ is the outer probability. Thus there is $\delta=\min\{\delta_P,\delta_Q\}$ such that

$$\begin{aligned}&\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho_P(h,g)<\delta}|G_m(h)-G_m(g)|>\varepsilon\right)\\&+\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho_Q(h,g)<\delta}|H_n(h)-H_n(g)|>\varepsilon\right)<\eta.\end{aligned}\tag{A.3}$$

Define a new metric associated with probability R by

$$\rho'_R(h_1,h_2)=\|(h_1-h_2)\|_{L^2(R)}.$$

It is not hard to show that $\rho'_R(h_1,h_2)\geq\rho_R(h_1,h_2)$ for all $h_1,h_2\in\mathscr{H}$. Then we define another new metric $\rho=\sqrt{\rho_P'^2+\rho_Q'^2}$ on $\mathscr{H}$.

By (A.3), for all $\varepsilon,\eta>0$, there is $\delta>0$ such that

$$\begin{aligned}&\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho_P(h,g)<\delta}|G_m(h)-G_m(g)|>\frac{\varepsilon\xi}{4}\right)\\&+\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho_Q(h,g)<\delta}|H_n(h)-H_n(g)|>\frac{\varepsilon\xi}{4}\right)<\eta/2.\end{aligned}\tag{A.4}$$

On the other hand,

$$\begin{aligned}
&\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho(h,g)<\delta}\left|\sqrt{T_N}\left(\hat{\phi}-\phi\right)(g)\left(\frac{1}{\xi\vee\hat{\sigma}_N(h)}-\frac{1}{\xi\vee\hat{\sigma}_N(g)}\right)\right|>\frac{\varepsilon}{2}\right)\\
\leq&\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho(h,g)<\delta}\left|\sqrt{T_N}\left(\hat{\phi}-\phi\right)(g)\right|\left|\xi\vee\hat{\sigma}_N(h)-\xi\vee\hat{\sigma}_N(g)\right|>\frac{\varepsilon}{2}\xi^2\right)\\
\leq&\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho(h,g)<\delta}\left|\sqrt{T_N}\left(\hat{\phi}-\phi\right)(g)\right|\left|\hat{\sigma}_N(h)-\hat{\sigma}_N(g)\right|>\frac{\varepsilon}{2}\xi^2\right)\\
\leq&\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho(h,g)<\delta}\left|\sqrt{T_N}\left(\hat{\phi}-\phi\right)(g)\right|\left|\sigma_N(h)-\sigma_N(g)\right|+o_p(1)>\frac{\varepsilon}{2}\xi^2\right)\\
\leq&\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho(h,g)<\delta}\left|\sqrt{T_N}\left(\hat{\phi}-\phi\right)(g)\right|\left|\sigma_N(h)-\sigma_N(g)\right|>\frac{\varepsilon}{4}\xi^2\right).
\end{aligned}$$

By definition,

$$\begin{aligned}
\left|\sigma_N(h)-\sigma_N(g)\right|^2&\leq\left|\sigma_N(h)-\sigma_N(g)\right|\left|\sigma_N(h)+\sigma_N(g)\right|=\left|\sigma_N^2(h)-\sigma_N^2(g)\right|\\
&=\left|\begin{array}{c}(1-\lambda)\left|P(h)\right|\left(1-\left|P(h)\right|\right)+\lambda\left|Q(h)\right|\left(1-\left|Q(h)\right|\right)\\-(1-\lambda)\left|P(g)\right|\left(1-\left|P(g)\right|\right)+\lambda\left|Q(g)\right|\left(1-\left|Q(g)\right|\right)\end{array}\right|\\
&\leq 3(1-\lambda)P\left(\left|h-g\right|^2\right)+3\lambda Q\left(\left|h-g\right|^2\right).
\end{aligned}$$

Now it is easy to show that

$$\left|\sigma_N(h)-\sigma_N(g)\right|\leq\sqrt{3\left(\rho_P'^2(h,g)+\rho_Q'^2(h,g)\right)}=\sqrt{3}\rho(h,g).$$

Therefore,

$$\begin{aligned}
&\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho(h,g)<\delta}\left|\sqrt{T_N}\left(\hat{\phi}-\phi\right)(g)\right|\left|\sigma_N(h)-\sigma_N(g)\right|>\frac{\varepsilon}{2}\xi^2\right)\\
\leq&\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{g\in\mathscr{H}}\left|\sqrt{T_N}\left(\hat{\phi}-\phi\right)(g)\right|\sqrt{3}\delta>\frac{\varepsilon}{4}\xi^2\right)=o(\delta).
\end{aligned}$$

We let δ be smaller than that in (A.4) such that

$$\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{g\in\mathscr{H}}\left|\sqrt{T_N}\left(\hat{\phi}-\phi\right)(g)\right|\sqrt{3\delta}>\frac{\varepsilon}{4}\xi^2\right)<\eta/2. \tag{A.5}$$

Combining (A.4) and (A.5) gives us that there is $\delta>0$ such that

$$\begin{aligned}
&\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho(h,g)<\delta}\left|\begin{array}{c}\sqrt{T_N}\frac{\left(\hat{\phi}-\phi\right)(h)}{\xi\vee\hat{\sigma}_N(h)}-\sqrt{T_N}\frac{\left(\hat{\phi}-\phi\right)(g)}{\xi\vee\hat{\sigma}_N(h)}\\+\sqrt{T_N}\frac{\left(\hat{\phi}-\phi\right)(g)}{\xi\vee\hat{\sigma}_N(h)}-\sqrt{T_N}\frac{\left(\hat{\phi}-\phi\right)(g)}{\xi\vee\hat{\sigma}_N(g)}\end{array}\right|>\varepsilon\right)\\
\leq&\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho(h,g)<\delta}\left|\left[G_m(h)-G_m(g)\right]-\left[H_n(h)-H_n(g)\right]\right|>\frac{\varepsilon\xi}{2}\right)\\
&+\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho(h,g)<\delta}\left|\sqrt{T_N}\frac{\left(\hat{\phi}-\phi\right)(g)}{\xi\vee\hat{\sigma}_N(h)}-\sqrt{T_N}\frac{\left(\hat{\phi}-\phi\right)(g)}{\xi\vee\hat{\sigma}_N(g)}\right|>\frac{\varepsilon}{2}\right)\\
\leq&\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho_P(h,g)<\delta}\left|G_m(h)-G_m(g)\right|>\frac{\varepsilon\xi}{4}\right)\\
&+\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho_Q(h,g)<\delta}\left|H_n(h)-H_n(g)\right|>\frac{\varepsilon\xi}{4}\right)\\
&+\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho(h,g)<\delta}\left|\sqrt{T_N}\left(\hat{\phi}-\phi\right)(g)\left(\frac{1}{\xi\vee\hat{\sigma}_N(h)}-\frac{1}{\xi\vee\hat{\sigma}_N(g)}\right)\right|>\frac{\varepsilon}{2}\right)<\eta.
\end{aligned}$$

This implies that $\sqrt{T_N}(\hat{\phi}-\phi)/(\xi\vee\hat{\sigma}_N)$ is asymptotically uniformly equicontinuous.

Notice that for all $h,g\in\mathscr{H}$,

$$\rho^2(h,g)=2\int|h-g|^2\,d\left(\frac{P+Q}{2}\right),$$

where $P/2+Q/2$ is a probability measure. By Lemma A.1.5, $\mathscr{H}$ is totally bounded under $\|\cdot\|_{L^2\left(\frac{P+Q}{2}\right)}$. Then the total boundedness of $\mathscr{H}$ under ρ follows from that for all $\varepsilon>0$,

$$N(\varepsilon,\mathscr{H},\rho)\leq N\left(\varepsilon/\sqrt{2},\mathscr{H},\|\cdot\|_{L^2\left(\frac{P+Q}{2}\right)}\right)<\infty.$$

Now by Theorems 1.5.4 and 1.5.7 in van der Vaart & Wellner (1996),

$$\sqrt{T_N}\frac{\left(\hat{\phi}-\phi\right)}{\xi\vee\hat{\sigma}_N}\rightsquigarrow\frac{\sqrt{1-\lambda}\mathbb{G}_P-\sqrt{\lambda}\mathbb{G}_Q}{\xi\vee\sigma}.$$

Let $\varphi_N=\phi/(\xi\vee\hat{\sigma}_N)$. Given any sequence $r_N\to\infty$, define

$$\mathbb{D}_N(\omega)=\left\{\varphi\in\ell^\infty\left(\bar{\mathscr{H}}\right):\varphi_N(\omega)+r_N^{-1}\varphi\in\ell^\infty\left(\bar{\mathscr{H}}\right)\right\}$$

for all $\omega\in\Omega$. Define

$$g_N(\omega)(\psi)=r_N\left(S\left(\varphi_N(\omega)+r_N^{-1}\psi\right)-S(\varphi_N(\omega))\right)$$

for all $\omega\in\Omega$, $\psi\in\mathbb{D}_N(\omega)$. We know $\varphi_N\to\phi/(\xi\vee\sigma)=\varphi_0$ a.s..

Let $\Omega_0=\{\omega\in\Omega:\varphi_N(\omega)\to\varphi_0\}$ and $\mathbb{P}(\Omega_0)=1$. Now we want to show that there is some g, for all $\omega\in\Omega_0$, $g_N(\omega)(\psi_N)\to g(\psi)$ for all $\psi_N\in\mathbb{D}_N(\omega)$ such that $\psi_N\to\psi$ for some $\psi\in C(\mathscr{H})$. We extend the proof of Lemma B.1 in Fang & Santos (2014) to show this result.

Given $\left\{Y_i^1,D_i^1\right\}_{i=1}^m$ and $\left\{Y_i^0,D_i^0\right\}_{i=1}^n$, P_m and Q_n have finitely many possible values on $\mathscr{H}$ respectively. Suppose totally there are J_N pairs of possible values for P_m and Q_n. Under lemma A.1.5, $\mathscr{H}$ is compact. Since ϕ is continuous on $\mathscr{H}$, for all $\psi\in C(\mathscr{H})$, $\sup_{h\in\mathscr{H}}(\varphi_N+t_N\psi)(h)<\infty$.

We define a correspondence $\bar{\Psi}_{\mathscr{H}}:C(\mathscr{H})\twoheadrightarrow\mathscr{H}$ by

$$\bar{\Psi}_{\mathscr{H}}(\psi)=\{h\in\mathscr{H}:\psi(h)=\mathscr{S}(\psi)\}$$

for all $\psi\in C(\mathscr{H})$. By Theorem 17.31 in Aliprantis & Border (2006), $\bar{\Psi}_{\mathscr{H}}(\varphi_0)$ is a nonempty compact set. We now extend the domain of $\bar{\Psi}_{\mathscr{H}}$ to $\ell^\infty(\mathscr{H})$ such that $\bar{\Psi}_{\mathscr{H}}:\ell^\infty(\mathscr{H})\twoheadrightarrow\mathscr{H}$:

$$\bar{\Psi}_{\mathscr{H}}(\psi)=\{h\in\mathscr{H}:\psi(h)=\mathscr{S}(\psi)\}$$

for all $\psi \in \ell^{\infty}(\mathscr{H})$. Now we want to show $\bar{\Psi}_{\mathscr{H}}$ is upper hemicontinuous at φ_0. (See Definition 17.2 of upper hemicontinuity in Aliprantis & Border (2006)) Suppose there is a sequence $\{\psi_n, h_n\}$ such that $h_n \in \bar{\Psi}_{\mathscr{H}}(\psi_n)$ and $\psi_n \to \varphi_0$. It is easy to show that

$$|\mathscr{S}(\psi_n) - \mathscr{S}(\varphi_0)| \leq \|\psi_n - \varphi_0\|_{\infty} \to 0,$$

which implies $\psi_n(h_n) \to \mathscr{S}(\varphi_0)$. Suppose h_n has no limit in $\bar{\Psi}_{\mathscr{H}}(\varphi_0)$. This implies for each $h \in \bar{\Psi}_{\mathscr{H}}(\varphi_0)$, there is an open neighborhood V_h and n_h such that $h_n \notin V_h$ when $n \geq n_h$. Because we have shown $\bar{\Psi}_{\mathscr{H}}(\varphi_0)$ is compact in $\mathscr{H}$, there is a finite open cover such that $\bar{\Psi}_{\mathscr{H}}(\varphi_0) \subset V = V_{h_1} \cup \cdots \cup V_{h_M}$. Let $n_0 = \max_{m \leq M} n_{h_m}$. Thus if $n > n_0$, then $h_n \notin V$ and therefore $h_n \notin \bar{\Psi}_{\mathscr{H}}(\varphi_0)$. Since $\mathscr{H}$ is compact and V^c is closed in $\mathscr{H}$, V^c is compact. Then

$$\sup_{h \in V^c} \varphi_0(h) < \sup_{h \in \mathscr{H}} \varphi_0(h) = \sup_{h \in \bar{\Psi}_{\mathscr{H}}(\varphi_0)} \varphi_0(h).$$

We let $\delta = \sup_{h \in \mathscr{H}} \varphi_0(h) - \sup_{h \in V^c} \varphi_0(h)$. Remember

$$\psi_n(h_n) = \sup_{h \in \mathscr{H}} \psi_n(h) = \sup_{h \in V^c} \psi_n(h).$$

Thus,

$$\left| \psi_n(h_n) - \sup_{h \in V^c} \varphi_0(h) \right| \leq \|\psi_n - \varphi_0\|_{\infty} \to 0.$$

For all n that is large enough,

$$\psi_n(h_n) \leq \sup_{h \in V^c} \varphi_0(h) + \frac{\delta}{2} < \sup_{h \in \mathscr{H}} \varphi_0(h).$$

This contradicts $\psi_n(h_n) \to \mathscr{S}(\varphi_0)$. Thus, there is $h \in \bar{\Psi}_{\mathscr{H}}(\varphi_0)$ such that $h_n \to h$. Then by Theorem 17.20 in Aliprantis & Border (2006), $\bar{\Psi}_{\mathscr{H}}$ is upper hemicontinuous at φ_0.

It is easy to show that under H_0,

$$\bar{\Psi}_{\mathscr{H}}(\varphi_N) = \bar{\Psi}_{\mathscr{H}}(\varphi_0) = \bar{\Psi}_{\mathscr{H}}(\phi).$$

Since $\varphi_N + t_N\psi$ is not continuous in $\mathscr{H}$, $\bar{\Psi}_{\mathscr{H}}(\varphi_N + t_N\psi)$ may be empty. As we have shown earlier, P_m and Q_n can have at most J_N pairs of possible values on $\mathscr{H}$. We can construct a modified version of φ_N, denoted by φ'_N such that φ'_N is upper semicontinuous and

(i) $\sup_{h\in\mathscr{H}} \varphi_N(h) = \sup_{h\in\mathscr{H}} \varphi'_N(h)$,

(ii) $\sup_{h\in\mathscr{H}} (\varphi_N + t_N\psi)(h) = \sup_{h\in\mathscr{H}} (\varphi'_N + t_N\psi)(h)$,

(iii) $\varphi'_N + t_N\psi \to \varphi_0$, as $N \to \infty$.

Specifically, we can set the value φ'_N at discontinuities to the largest limit value at that point. In this case, $\bar{\Psi}_{\mathscr{H}}(\varphi'_N + t_N\psi) \neq \varnothing$, because $\varphi_N + t_N\psi$ is upper semicontinuous and $\mathscr{H}$ is compact.

Let $t_N = r_N^{-1}$. It is easy to show that

$$\left| \sup_{h\in\mathscr{H}} \{\varphi_N(h) + t_N\psi_N(h)\} - \sup_{h\in\mathscr{H}} \{\varphi_N(h) + t_N\psi(h)\} \right| \leq t_N \|\psi_N - \psi\|_\infty = o(t_N).$$

Since $\varphi'_N(\omega) + t_N\psi$ converges to φ_0 and $\bar{\Psi}_{\mathscr{H}}$ is upper hemicontinuous at φ_0, there is a sequence δ_N such that

$$\bar{\Psi}_{\mathscr{H}}\left(\varphi'_N + t_N\psi\right) \subset \bar{\Psi}_{\mathscr{H}}(\varphi_0)^{\delta_N},$$

where $\bar{\Psi}_{\mathscr{H}}(\varphi_0)^{\delta_N} = \left\{h \in \mathscr{H} : \inf_{h'\in\bar{\Psi}_{\mathscr{H}}(\varphi_0)} \|h - h'\|_{L^2(R)} \leq \delta_N\right\}$ and $R = P/2 + Q/2$. Remember that under H_0,

$$\bar{\Psi}_{\mathscr{H}}(\varphi_0) = \bar{\Psi}_{\mathscr{H}}(\varphi_N) \text{ and } \sup_{h\in\bar{\Psi}_{\mathscr{H}}(\varphi_0)} \varphi_N(h) = \sup_{h\in\mathscr{H}} \varphi_N(h) = 0.$$

Thus, we have that under H_0,

$$\begin{aligned}
&\left|\sup_{h\in\mathscr{H}}\left\{\varphi_N(h)+t_N\psi(h)\right\}-\sup_{h\in\bar{\Psi}_{\mathscr{H}}(\varphi_0)}\left\{\varphi_N(h)+t_N\psi(h)\right\}\right| \\
=&\sup_{h\in\bar{\Psi}_{\mathscr{H}}\left(\varphi_N'+t_N\psi\right)}\left\{\varphi_N'(h)+t_N\psi(h)\right\}-\sup_{h\in\bar{\Psi}_{\mathscr{H}}(\varphi_0)}\left\{\varphi_N(h)+t_N\psi(h)\right\} \\
\leq&\sup_{h\in\bar{\Psi}_{\mathscr{H}}(\varphi_0)^{\delta_N}}\left\{\varphi_N'(h)+t_N\psi(h)\right\}-\sup_{h\in\bar{\Psi}_{\mathscr{H}}(\varphi_0)}t_N\psi(h) \\
\leq&\sup_{h_1,h_2\in\mathscr{H},\|h_1-h_2\|_{L^2(R)}\leq\delta_n}t_N\left|\psi(h_1)-\psi(h_2)\right|=o(t_N).
\end{aligned}$$

Finally, put all things together and we have

$$\begin{aligned}
&\left|\sup_{h\in\mathscr{H}}\left\{\varphi_N(h)+t_N\psi_N(h)\right\}-\sup_{h\in\mathscr{H}}\varphi_N(h)-t_N\sup_{h\in\bar{\Psi}_{\mathscr{H}}(\varphi_0)}\psi(h)\right| \\
\leq&\left|\sup_{h\in\bar{\Psi}_{\mathscr{H}}(\varphi_0)}\left\{\varphi_N(h)+t_N\psi(h)\right\}-t_N\sup_{h\in\bar{\Psi}_{\mathscr{H}}(\varphi_0)}\psi(h)\right|+o(t_N)=o(t_N).
\end{aligned}$$

This implies that $g_N(\omega)(\psi_N)\to\sup_{h\in\Psi_{\mathscr{H}}(\varphi_0)}\psi(h)$.

By Lemma A.1.2,

$$\sqrt{T_N}\left(\mathscr{S}\left(\frac{\hat{\phi}}{\xi\vee\hat{\sigma}_N}\right)-\mathscr{S}\left(\frac{\phi}{\xi\vee\hat{\sigma}_N}\right)\right)\rightsquigarrow\mathscr{S}_{\bar{\Psi}_{\mathscr{H}}(\varphi_0)}\left(\frac{\sqrt{1-\lambda}\mathbb{G}_P-\sqrt{\lambda}\mathbb{G}_Q}{\xi\vee\sigma}\right).$$

Notice that under H_0, $\mathscr{S}_{\bar{\Psi}_{\mathscr{H}}(\varphi_0)}=\mathscr{S}_{\phi}'$, where $\mathscr{S}_{\phi}'$ is obtained in Lemma A.1.9. As defined in the context, under H_0,

$$\Psi_{\mathscr{H}}=\{h\in\mathscr{H}:\phi(h)=0\}=\bar{\Psi}_{\mathscr{H}}(\phi)=\bar{\Psi}_{\mathscr{H}}(\varphi_0),$$

which gives

$$\mathscr{S}_{\bar{\Psi}_{\mathscr{H}}(\varphi_0)}\left(\frac{\sqrt{1-\lambda}\mathbb{G}_P-\sqrt{\lambda}\mathbb{G}_Q}{\xi\vee\sigma}\right)=\mathscr{S}_{\Psi_{\mathscr{H}}}\left(\frac{\sqrt{1-\lambda}\mathbb{G}_P-\sqrt{\lambda}\mathbb{G}_Q}{\xi\vee\sigma}\right).$$

Now we want to show that under H_0, if restricted on $\Psi_{\mathscr{H}}$, $\sqrt{1-\lambda}\mathbb{G}_P - \sqrt{\lambda}\mathbb{G}_Q \stackrel{L}{=} \mathbb{G}_H$.

First, note that for all $h \in \Psi_{\mathscr{H}}$, $\phi(h) = 0$, that is, $P(h) = Q(h) = H(h)$. Also,

$$\begin{aligned} &Cov\left(\frac{\sqrt{1-\lambda}\mathbb{G}_P - \sqrt{\lambda}\mathbb{G}_Q}{\xi \vee \sigma}(h), \frac{\sqrt{1-\lambda}\mathbb{G}_P - \sqrt{\lambda}\mathbb{G}_Q}{\xi \vee \sigma}(g)\right) \\ =&\frac{(1-\lambda)[P(hg) - P(h)P(g)] + \lambda[Q(hg) - Q(h)Q(g)]}{(\xi \vee \sigma(h)) \cdot (\xi \vee \sigma(g))}. \end{aligned}$$

Suppose $h, g \in \Psi_{\mathscr{H}}$ with $h = (-1)^{d_h} \cdot 1_{B_h \times \{d_h\}}$ and $g = (-1)^{d_g} \cdot 1_{B_g \times \{d_g\}}$.

(i) If $d_h \neq d_g$, then $hg = 0$ and thus $P(hg) = Q(hg) = H(hg)$.

(ii) If $d_h = d_g$, then $hg = 1_{B_h \cap B_g \times \{d_h\}}$ and thus $P(hg) = Q(hg) = H(hg)$.

Therefore, we now have

$$\begin{aligned} &Cov\left(\frac{\sqrt{1-\lambda}\mathbb{G}_P - \sqrt{\lambda}\mathbb{G}_Q}{\xi \vee \sigma}(h), \frac{\sqrt{1-\lambda}\mathbb{G}_P - \sqrt{\lambda}\mathbb{G}_Q}{\xi \vee \sigma}(g)\right) \\ =&\frac{H(hg) - H(h)H(g)}{(\xi \vee \sigma(h)) \cdot (\xi \vee \sigma(g))} = Cov\left(\frac{\mathbb{G}_H}{\xi \vee \sigma}(h), \frac{\mathbb{G}_H}{\xi \vee \sigma}(g)\right). \end{aligned}$$

Equivalence of the covariance kernels implies equivalence of the probability laws of the mean zero Gaussian processes, thus

$$\mathscr{S}_{\Psi_{\mathscr{H}}}\left(\frac{\sqrt{1-\lambda}\mathbb{G}_P - \sqrt{\lambda}\mathbb{G}_Q}{\xi \vee \sigma}\right) \stackrel{L}{=} \mathscr{S}_{\Psi_{\mathscr{H}}}\left(\frac{\mathbb{G}_H}{\xi \vee \sigma}\right).$$

■

Given the probability measure $R = 1/2 \cdot P + 1/2 \cdot Q$, we define

$$\overrightarrow{d}_H(A_1, A_2) = \sup_{a \in A_1} \inf_{b \in A_2} \|a - b\|_{L^2(R)}$$

and

$$d_H(A_1, A_2) = \max\left\{\overrightarrow{d}_H(A_1, A_2), \overrightarrow{d}_H(A_2, A_1)\right\},$$

for all set $A_1, A_2 \subset \mathscr{H}$. The following lemma concludes that $\hat{\Psi}_{\mathscr{H}}$ in (1.16) is valid in the sense that $\hat{\mathscr{S}}_N$ satisfies Assumption 3.3 in Fang & Santos (2014).

Lemma A.2.1 *Under Assumptions 1.3.1 and 1.3.2, if H_0 is true, $d_H(\hat{\Psi}_{\mathscr{H}}, \Psi_{\mathscr{H}}) \to_p 0$.*

Proof of Lemma A.2.1.

First, for all $\varepsilon > 0$,

$$\begin{aligned}\lim_{N\to\infty} \mathbb{P}\left(\overrightarrow{d}_H\left(\Psi_{\mathscr{H}}, \hat{\Psi}_{\mathscr{H}}\right) > \varepsilon\right) &\leq \lim_{N\to\infty} \mathbb{P}\left(\Psi_{\mathscr{H}} \backslash \hat{\Psi}_{\mathscr{H}} \neq \varnothing\right) \\ &\leq \lim_{N\to\infty} \mathbb{P}\left(\sup_{h\in\Psi_{\mathscr{H}}\backslash\hat{\Psi}_{\mathscr{H}}} \left|\hat{\phi}(h) - \phi(h)\right| > \tau_N\right) \\ &\leq \lim_{N\to\infty} \mathbb{P}\left(\sup_{h\in\mathscr{H}} \sqrt{T_N}\left|\hat{\phi}(h) - \phi(h)\right| > \sqrt{T_N}\tau_N\right).\end{aligned}$$

By Theorem 1.3.1, $\sqrt{T_N}\left(\hat{\phi} - \phi\right) \rightsquigarrow \sqrt{1-\lambda}\mathbb{G}_P - \sqrt{\lambda}\mathbb{G}_Q$. Then by Theorem 1.3.6 in van der Vaart & Wellner (1996),

$$\sup_{h\in\mathscr{H}} \sqrt{T_N}\left|\hat{\phi}(h) - \phi(h)\right| \rightsquigarrow \sup_{h\in\mathscr{H}} \left|\sqrt{1-\lambda}\mathbb{G}_P(h) - \sqrt{\lambda}\mathbb{G}_Q(h)\right|.$$

If $\sqrt{T_N}\tau_N \to \infty$, then $\lim_{n\to\infty} \mathbb{P}\left(\overrightarrow{d}_H\left(\Psi_{\mathscr{H}}, \hat{\Psi}_{\mathscr{H}}\right) > \varepsilon\right) = 0$.

Next, consider $\overrightarrow{d}_H\left(\hat{\Psi}_{\mathscr{H}}, \Psi_{\mathscr{H}}\right)$. Define $d(h, \Psi_{\mathscr{H}}) = \inf_{g\in\Psi_{\mathscr{H}}} \|h - g\|_{L^2(R)}$ for all $h \in \mathscr{H}$. For each $\varepsilon > 0$, let $D_\varepsilon = \{h \in \mathscr{H} : d(h, \Psi_{\mathscr{H}}) \geq \varepsilon\}$. In Lemma A.1.9, we have shown $\phi \in C(\mathscr{H})$ with $\mathscr{H}$ being compact under norm $\|\cdot\|_{L^2(R)}$. Suppose there is $\{h_n\} \subset D_\varepsilon$ and $h_n \to h$ for some $h \in \mathscr{H}$, then

$$\begin{aligned}d(h, \Psi_{\mathscr{H}}) &= \inf_{g\in\Psi_{\mathscr{H}}} \|h - g\|_{L^2(R)} = \inf_{g\in\Psi_{\mathscr{H}}} \|h - h_n + h_n - g\|_{L^2(R)} \\ &\geq \inf_{g\in\Psi_{\mathscr{H}}} \|h_n - g\|_{L^2(R)} - \|h - h_n\|_{L^2(R)} \geq \varepsilon - \|h - h_n\|_{L^2(R)},\end{aligned}$$

which is true for all n. Letting $n \to \infty$ gives us $d(h, \Psi_{\mathscr{H}}) \geq \varepsilon$. This implies D_ε is closed in $\mathscr{H}$ which is compact, thus D_ε is compact. If $D_\varepsilon \neq \varnothing$, then $\exists \delta_\varepsilon > 0$ such that $\inf_{h\in D_\varepsilon} |\phi(h)| > \delta_\varepsilon$.

Then we have

$$\lim_{N\to\infty}\mathbb{P}\left(\overrightarrow{d}_H\left(\hat{\Psi}_{\mathscr{H}},\Psi_{\mathscr{H}}\right)>\varepsilon\right)=\lim_{N\to\infty}\mathbb{P}\left(\sup_{h\in\hat{\Psi}_{\mathscr{H}}}\inf_{g\in\Psi_{\mathscr{H}}}\|h-g\|_{L^2(R)}>\varepsilon\right)$$
$$\leq\lim_{N\to\infty}\mathbb{P}\left(\sup_{h\in\hat{\Psi}_{\mathscr{H}}\setminus\Psi_{\mathscr{H}}}|\phi(h)|>\delta_\varepsilon,\ \sup_{h\in\hat{\Psi}_{\mathscr{H}}\setminus\Psi_{\mathscr{H}}}\left|\hat{\phi}(h)\right|\leq\tau_N\right).$$

Here we define events:

$$A_N=\left\{\sup_{h\in\hat{\Psi}_{\mathscr{H}}\setminus\Psi_{\mathscr{H}}}|\phi(h)|-\frac{\delta_\varepsilon}{2}\leq\sup_{h\in\hat{\Psi}_{\mathscr{H}}\setminus\Psi_{\mathscr{H}}}\left|\hat{\phi}(h)\right|\leq\sup_{h\in\hat{\Psi}_{\mathscr{H}}\setminus\Psi_{\mathscr{H}}}|\phi(h)|+\frac{\delta_\varepsilon}{2}\right\}.$$

Now we have

$$\mathbb{P}\left(\sup_{h\in\mathscr{H}}\left|\hat{\phi}(h)-\phi(h)\right|\leq\frac{\delta_\varepsilon}{2}\right)\leq\mathbb{P}\left(\sup_{h\in\hat{\Psi}_{\mathscr{H}}\setminus\Psi_{\mathscr{H}}}\left|\hat{\phi}(h)-\phi(h)\right|\leq\frac{\delta_\varepsilon}{2}\right)$$
$$\leq\mathbb{P}\left(\left|\sup_{h\in\hat{\Psi}_{\mathscr{H}}\setminus\Psi_{\mathscr{H}}}\left|\hat{\phi}(h)\right|-\sup_{h\in\hat{\Psi}_{\mathscr{H}}\setminus\Psi_{\mathscr{H}}}|\phi(h)|\right|\leq\frac{\delta_\varepsilon}{2}\right)=\mathbb{P}(A_N).$$

By Lemma A.1.8, $\lim_{n\to\infty}\mathbb{P}(A_N)=1$. Thus,

$$\lim_{N\to\infty}\mathbb{P}\left(\overrightarrow{d}_H\left(\hat{\Psi}_{\mathscr{H}},\Psi_{\mathscr{H}}\right)>\varepsilon\right)$$
$$\leq\lim_{N\to\infty}\mathbb{P}\left(\sup_{h\in\hat{\Psi}_{\mathscr{H}}\setminus\Psi_{\mathscr{H}}}|\phi(h)|>\delta_\varepsilon,\ \sup_{h\in\hat{\Psi}_{\mathscr{H}}\setminus\Psi_{\mathscr{H}}}\left|\hat{\phi}(h)\right|\leq\tau_N,A_N\right)$$
$$\leq\lim_{N\to\infty}\mathbb{P}\left(\sup_{h\in\hat{\Psi}_{\mathscr{H}}\setminus\Psi_{\mathscr{H}}}\left|\hat{\phi}(h)\right|\geq\frac{\delta_\varepsilon}{2},\ \sup_{h\in\hat{\Psi}_{\mathscr{H}}\setminus\Psi_{\mathscr{H}}}\left|\hat{\phi}(h)\right|\leq\tau_N\right)=0\text{ as }\tau_N\downarrow 0.$$

■

Proof of Theorem 1.3.2.

(i) We first show that $\hat{c}_{1-\alpha}\to_p c_{1-\alpha}$, where $c_{1-\alpha}$ is defined by

$$c_{1-\alpha}=\inf\left\{c:\mathbb{P}\left(\mathscr{S}_{\Psi_{\mathscr{H}}}\left(\frac{\sqrt{1-\lambda}\mathbb{G}_P-\sqrt{\lambda}\mathbb{G}_Q}{\xi\vee\sigma}\right)\leq c\right)\geq 1-\alpha\right\}.$$

$C(\mathscr{H})$ is complete under $\|\cdot\|_\infty$. By Lemma A.1.6, $\mathscr{H}$ is Donsker for both P and Q. Also, by the construction of $\mathscr{H}$, $P\|h-Ph\|_{\mathscr{H}}^2<\infty$ and $Q\|h-Qh\|_{\mathscr{H}}^2<\infty$. We let

$$\begin{aligned}\hat{G}_{Pm}(h)=&\frac{1}{\sqrt{m}}\left[\sum_{i=1}^{m}h\left(Y_i^{1*},D_i^{1*}\right)-\sum_{i=1}^{m}h\left(Y_i^{1},D_i^{1}\right)\right]\\=&\frac{1}{\sqrt{\lambda N}}\left[\sum_{i=1}^{m}h\left(Y_i^{1*},D_i^{1*}\right)-\sum_{i=1}^{m}h\left(Y_i^{1},D_i^{1}\right)\right]\\&+\left(\frac{1}{\sqrt{m}}-\frac{1}{\sqrt{\lambda N}}\right)\left[\sum_{i=1}^{m}h\left(Y_i^{1*},D_i^{1*}\right)-\sum_{i=1}^{m}h\left(Y_i^{1},D_i^{1}\right)\right]=\hat{G}_{\lambda Pm}(h)+A_N\end{aligned}$$

and

$$\begin{aligned}\hat{G}_{Qn}(h)=&\frac{1}{\sqrt{n}}\left[\sum_{i=1}^{n}h\left(Y_i^{0*},D_i^{0*}\right)-\sum_{i=1}^{n}h\left(Y_i^{0},D_i^{0}\right)\right]\\=&\frac{1}{\sqrt{(1-\lambda)N}}\left[\sum_{i=1}^{n}h\left(Y_i^{0*},D_i^{0*}\right)-\sum_{i=1}^{n}h\left(Y_i^{0},D_i^{0}\right)\right]\\&+\left(\frac{1}{\sqrt{n}}-\frac{1}{\sqrt{(1-\lambda)N}}\right)\left[\sum_{i=1}^{n}h\left(Y_i^{0*},D_i^{0*}\right)-\sum_{i=1}^{n}h\left(Y_i^{0},D_i^{0}\right)\right]=\hat{G}_{\lambda Qn}(h)+B_N.\end{aligned}$$

where $A_n\to 0$ a.s. and $B_n\to 0$ a.s.. Then by Theorem 3.6.2 in van der Vaart & Wellner (1996), we have that

$$\sup_{f\in BL_1}\left|E\left[f\left(\hat{G}_{\lambda Pm}\right)|\left\{\left(Y_i^1,D_i^1\right)\right\}_{i=1}^m,\left\{\left(Y_i^0,D_i^0\right)\right\}_{i=1}^n\right]-E\left[f\left(\mathbb{G}_P\right)\right]\right|\to 0$$

and

$$\sup_{f\in BL_1}\left|E\left[f\left(\hat{G}_{\lambda Qn}\right)|\left\{\left(Y_i^1,D_i^1\right)\right\}_{i=1}^m,\left\{\left(Y_i^0,D_i^0\right)\right\}_{i=1}^n\right]-E\left[f\left(\mathbb{G}_Q\right)\right]\right|\to 0$$

for almost all sequence $\left\{\left(Y_i^1,D_i^1\right)\right\}_{i=1}^m,\left\{\left(Y_i^0,D_i^0\right)\right\}_{i=1}^n$, where $BL_1=BL_1\left(\ell^\infty(\mathscr{H})\right)$. Then because conditional on $\left\{\left(Y_i^1,D_i^1\right)\right\}_{i=1}^m,\left\{\left(Y_i^0,D_i^0\right)\right\}_{i=1}^n$, $\hat{G}_{\lambda Pm}$ is independent of $\hat{G}_{\lambda Qn}$, we have

given almost every sequence $\left\{\left(Y_i^1, D_i^1\right)\right\}_{i=1}^m, \left\{\left(Y_i^0, D_i^0\right)\right\}_{i=1}^n$,

$$\left(\hat{G}_{\lambda Pm}, \hat{G}_{\lambda Qn}\right) \rightsquigarrow \left(\mathbb{G}_P, \mathbb{G}_Q\right),$$

where $\mathbb{G}_P$ and $\mathbb{G}_Q$ are independent processes. This implies

$$\left(\hat{G}_{Pm}, \hat{G}_{Qn}\right) \rightsquigarrow \left(\mathbb{G}_P, \mathbb{G}_Q\right).$$

Then by the continuous mapping theorem, conditional on almost all sequences $\left\{\left(Y_i^1, D_i^1\right)\right\}_{i=1}^m$, $\left\{\left(Y_i^0, D_i^0\right)\right\}_{i=1}^n$, $\sqrt{T_N}\left(\hat{\phi}^* - \hat{\phi}\right) \rightsquigarrow \sqrt{1-\lambda}\mathbb{G}_P - \sqrt{\lambda}\mathbb{G}_Q$.

Given any $\left\{\left(Y_i^1, D_i^1\right)\right\}_{i=1}^m$, $\left\{\left(Y_i^0, D_i^0\right)\right\}_{i=1}^n$, $\left\|\hat{P}_m^* - P_m\right\|_\infty \to 0$ a.s. and $\left\|\hat{Q}_n^* - Q_n\right\|_\infty \to 0$ a.s. by Lemma A.1.4 and Glivenko-Cantelli Theorem. Then by Lemma A.1.8,

$$\left\|\hat{P}_m^* - P\right\|_\infty \to 0 \text{ a.s. } \left\|\hat{Q}_n^* - Q\right\|_\infty \to 0 \text{ a.s.}$$

which implies $\left\|\hat{\sigma}_N^* - \sigma\right\|_\infty \to 0$ a.s. Repeating the proof for asymptotic uniform equicontinuity in Theorem 1.3.1 gives us

$$\frac{\sqrt{T_N}\left(\hat{\phi}^* - \hat{\phi}\right)}{\xi \vee \hat{\sigma}_N^*} \rightsquigarrow \frac{\sqrt{1-\lambda}\mathbb{G}_P - \sqrt{\lambda}\mathbb{G}_Q}{\xi \vee \sigma}$$

for almost all sequences $\left\{\left(Y_i^1, D_i^1\right)\right\}_{i=1}^m, \left\{\left(Y_i^0, D_i^0\right)\right\}_{i=1}^n$. Or in another word,

$$\sup_{f \in BL_1} \left| \begin{array}{c} E\left[f\left(\frac{\sqrt{T_N}\left(\hat{\phi}^* - \hat{\phi}\right)}{\xi \vee \hat{\sigma}_N^*}\right) \mid \left\{\left(Y_i^1, D_i^1\right)\right\}_{i=1}^m, \left\{\left(Y_i^0, D_i^0\right)\right\}_{i=1}^n\right] \\ -E\left[f\left(\frac{\sqrt{1-\lambda}\mathbb{G}_P - \sqrt{\lambda}\mathbb{G}_Q}{\xi \vee \sigma}\right)\right] \end{array} \right| \to 0,$$

for almost all sequences $\left\{\left(Y_i^1, D_i^1\right)\right\}_{i=1}^m, \left\{\left(Y_i^0, D_i^0\right)\right\}_{i=1}^n$. Also, similarly to the proof of Theorem

2.9.7 in van der Vaart & Wellner (1996), we have

$$\sup_{f\in BL_1}\left|\begin{array}{c} E\left[f\left(\frac{\sqrt{T_N}\left(\hat{\phi}^*-\hat{\phi}\right)}{\xi\vee\hat{\sigma}_N^*}\right)^*|\left\{\left(Y_i^1,D_i^1\right)\right\}_{i=1}^m,\left\{\left(Y_i^0,D_i^0\right)\right\}_{i=1}^n\right] \\ -E\left[f\left(\frac{\sqrt{1-\lambda}\mathbb{G}_P-\sqrt{\lambda}\mathbb{G}_Q}{\xi\vee\sigma}\right)\right]\end{array}\right|\to 0$$

and

$$\sup_{f\in BL_1}\left|\begin{array}{c} E\left[f\left(\frac{\sqrt{T_N}\left(\hat{\phi}^*-\hat{\phi}\right)}{\xi\vee\hat{\sigma}_N^*}\right)|\left\{\left(Y_i^1,D_i^1\right)\right\}_{i=1}^m,\left\{\left(Y_i^0,D_i^0\right)\right\}_{i=1}^n\right] \\ -E\left[f\left(\frac{\sqrt{1-\lambda}\mathbb{G}_P-\sqrt{\lambda}\mathbb{G}_Q}{\xi\vee\sigma}\right)\right]^*\end{array}\right|\to 0$$

for almost all sequences $\left\{\left(Y_i^1,D_i^1\right)\right\}_{i=1}^m,\left\{\left(Y_i^0,D_i^0\right)\right\}_{i=1}^n$, where

$$f\left(\frac{\sqrt{T_N}\left(\hat{\phi}^*-\hat{\phi}\right)}{\xi\vee\hat{\sigma}_N^*}\right)^* \text{ and } f\left(\frac{\sqrt{T_N}\left(\hat{\phi}^*-\hat{\phi}\right)}{\xi\vee\hat{\sigma}_N^*}\right)_*$$

denote measurable majorants and minorants with respect to random weights used in bootstrap and $\left\{\left(Y_i^1,D_i^1\right)\right\}_{i=1}^m,\left\{\left(Y_i^0,D_i^0\right)\right\}_{i=1}^n$ jointly. This shows the asymptotic measurability of $(\xi\vee\hat{\sigma}_N^*)^{-1}\sqrt{T_N}(\hat{\phi}^*-\hat{\phi})$. Also, $f((\xi\vee\hat{\sigma}_N^*)^{-1}\sqrt{T_N}(\hat{\phi}^*-\hat{\phi}))$ is a measurable function of random weights used in constructing $\hat{\phi}^*$ for almost every sequence $\left\{\left(Y_i^1,D_i^1\right)\right\}_{i=1}^m,\left\{\left(Y_i^0,D_i^0\right)\right\}_{i=1}^n$, for each continuous and bounded f. All these imply that Assumptions 3.1 and 3.2 in Fang & Santos (2014) hold.

By Theorem 1.3.1, Assumption 2.2 in Fang & Santos (2014) holds. Moreover $\mathscr{S}_{\Psi_{\mathscr{H}}}=\mathscr{S}'_{\phi}$ satisfies Assumption 2.3(i) in Fang & Santos (2014) holds. By Lemma A.2.1 and Lemma B.3 in Fang & Santos (2014), Assumption 3.3 in Fang & Santos (2014) holds. If the CDF of $\mathscr{S}_{\Psi_{\mathscr{H}}}((\xi\vee\sigma)^{-1}(\sqrt{1-\lambda}\mathbb{G}_P-\sqrt{\lambda}\mathbb{G}_Q))$ is strictly increasing at its $1-\alpha$ quantile $c_{1-\alpha}$, by Theorem 3.3 and Corollary 3.2 in Fang & Santos (2014), $\hat{c}_{1-\alpha}\to_p c_{1-\alpha}$.

Thus, if H_0 is true and the CDF of $\mathscr{S}_{\Psi_{\mathscr{H}}}((\sqrt{1-\lambda}\mathbb{G}_P-\sqrt{\lambda}\mathbb{G}_Q)/(\xi\vee\sigma))$ is continuous at its $1-\alpha$ quantile $c_{1-\alpha}$, under decision rule,

$$\mathbb{P}\left(\sqrt{T_N}\mathscr{S}(\hat{\phi})>\hat{c}_{1-\alpha}\right)=1-\mathbb{P}\left(\sqrt{T_N}\mathscr{S}(\hat{\phi})-\hat{c}_{1-\alpha}+c_{1-\alpha}\leq c_{1-\alpha}\right)\to\alpha.$$

(ii) Suppose H_0 is false, that is $\sup_{h\in\mathscr{H}}\phi(h)>0$.

First consider $\mathscr{S}(\sqrt{T_N}(\hat{\phi}^*-\hat{\phi})/(\xi\vee\hat{\sigma}_N^*))$. Since we have shown that

$$\sqrt{T_N}(\hat{\phi}^*-\hat{\phi})/(\xi\vee\hat{\sigma}_N^*)\rightsquigarrow(\sqrt{1-\lambda}\mathbb{G}_P-\sqrt{\lambda}\mathbb{G}_Q)/(\xi\vee\sigma)$$

for almost all sequences $\left\{\left(Y_i^1,D_i^1\right)\right\}_{i=1}^m,\left\{\left(Y_i^0,D_i^0\right)\right\}_{i=1}^n$, by the continuous mapping theorem,

$$\mathscr{S}\left(\frac{\sqrt{T_N}\left(\hat{\phi}^*-\hat{\phi}\right)}{\xi\vee\hat{\sigma}_N^*}\right)\rightsquigarrow\mathscr{S}\left(\frac{\sqrt{1-\lambda}\mathbb{G}_P-\sqrt{\lambda}\mathbb{G}_Q}{\xi\vee\sigma}\right).$$

Construct

$$\hat{c}'_{1-\alpha}=\inf\left\{c:\mathbb{P}\left(\mathscr{S}\left(\frac{\sqrt{T_N}\left(\hat{\phi}^*-\hat{\phi}\right)}{\xi\vee\hat{\sigma}_N^*}\right)\leq c|\{(Y_i^1,D_i^1)\}_{i=1}^m,\{(Y_i^0,D_i^0)\}_{i=1}^n\right)\geq1-\alpha\right\}. \tag{A.6}$$

By Theorem 11.1 in Davydov *et al.* (1998), the CDF of $\mathscr{S}\left((\sqrt{1-\lambda}\mathbb{G}_P-\sqrt{\lambda}\mathbb{G}_Q)/(\xi\vee\sigma)\right)$ is strictly increasing and continuous everywhere except on a countable subset of its support. By the proof similar to that of Corollary 3.2 in Fang & Santos (2014), $\hat{c}'_{1-\alpha}\to_p c'_{1-\alpha}$, where $c'_{1-\alpha}$ is the $1-\alpha$ quantile of $\mathscr{S}\left((\sqrt{1-\lambda}\mathbb{G}_P-\sqrt{\lambda}\mathbb{G}_Q)/(\xi\vee\sigma)\right)$. By construction, $0\leq\hat{c}_{1-\alpha}\leq\hat{c}'_{1-\alpha}$. Thus, $\hat{c}_{1-\alpha}=O_p(1)$.

By Lemma A.1.8,

$$\mathbb{P}\left(\sqrt{T_N}\mathscr{S}\left(\hat{\phi}\right)>\hat{c}_{1-\alpha}\right)\to1.$$

■

A.3 Results in Section 1.4

Proof of Lemma 1.4.1.

Let $m_k = \sum_{i=1}^N 1_{\mathbb{R}\times\{0,1\}\times\{z_k\}}(Y_i, D_i, Z_i)$. Now we have

$$
\begin{aligned}
&\hat{\phi}_K(h,g) \\
=&\frac{1}{N}\sum_{i=1}^N \frac{(h\cdot g_2)(Y_i,D_i,Z_i)}{\frac{1}{N}\sum_{i=1}^N g_2(Y_i,D_i,Z_i)} - \frac{1}{N}\sum_{i=1}^N \frac{(h\cdot g_1)(Y_i,D_i,Z_i)}{\frac{1}{N}\sum_{i=1}^N g_1(Y_i,D_i,Z_i)} \\
=&\frac{1}{N}\sum_{i=1}^N \frac{(h\cdot g_2)(Y_i,D_i,Z_i)}{P(g_2)} + \frac{1}{N}\sum_{i=1}^N \left(\frac{1}{\frac{1}{N}\sum_{i=1}^N g_2(Y_i,D_i,Z_i)} - \frac{1}{P(g_2)}\right)(h\cdot g_2)(Y_i,D_i,Z_i) \\
&-\frac{1}{N}\sum_{i=1}^N \frac{(h\cdot g_1)(Y_i,D_i,Z_i)}{P(g_1)} - \frac{1}{N}\sum_{i=1}^N \left(\frac{1}{\frac{1}{N}\sum_{i=1}^N g_1(Y_i,D_i,Z_i)} - \frac{1}{P(g_1)}\right)(h\cdot g_1)(Y_i,D_i,Z_i)
\end{aligned}
$$

and

$$
\begin{aligned}
&\hat{\phi}_K(h,g) - \phi_K(h,g) \\
=&\frac{1}{N}\sum_{i=1}^N \left\{\frac{(h\cdot g_2)(Y_i,D_i,Z_i)}{P(g_2)} - \frac{P(h\cdot g_2)}{P(g_2)}\right\} - \frac{1}{N}\sum_{i=1}^N \left\{\frac{(h\cdot g_1)(Y_i,D_i,Z_i)}{P(g_1)} - \frac{P(h\cdot g_1)}{P(g_1)}\right\} \\
&-\left(\frac{1}{N}\sum_{i=1}^N [g_2(Y_i,D_i,Z_i) - P(g_2)]\right)\left(\frac{\frac{1}{N}\sum_{i=1}^N (h\cdot g_2)(Y_i,D_i,Z_i)}{\frac{1}{N}\sum_{i=1}^N g_2(Y_i,D_i,Z_i)P(g_2)}\right) \\
&+\left(\frac{1}{N}\sum_{i=1}^N [g_1(Y_i,D_i,Z_i) - P(g_1)]\right)\left(\frac{\frac{1}{N}\sum_{i=1}^N (h\cdot g_1)(Y_i,D_i,Z_i)}{\frac{1}{N}\sum_{i=1}^N g_1(Y_i,D_i,Z_i)P(g_2)}\right). \qquad \text{(A.7)}
\end{aligned}
$$

As defined before, $\mathscr{G}_K = \left\{1_{\mathbb{R}\times\{0,1\}\times\{z_k\}} : k = 1,2,\cdots,K\right\}$, and it has been shown that $\mathscr{G}_K$ is P-Glivenko-Cantelli by Lemma A.1.12.

We now first show the marginal convergence of $\sqrt{N}\left(\hat{\phi}_K(h,g) - \phi_K(h,g)\right)$.

By the multivalue central limit theorem,

$$
S_N = \frac{1}{\sqrt{N}}\sum_{i=1}^N \begin{pmatrix} \frac{(h\cdot g_2)(Y_i,D_i,Z_i)}{P(g_2)} - \frac{P(h\cdot g_2)}{P(g_2)} \\ \frac{(h\cdot g_1)(Y_i,D_i,Z_i)}{P(g_1)} - \frac{P(h\cdot g_1)}{P(g_1)} \\ g_2(Y_i,D_i,Z_i) - P(g_2) \\ g_1(Y_i,D_i,Z_i) - P(g_1) \end{pmatrix} \rightsquigarrow N(0,\Sigma),
$$

where Σ is the covariance matrix of the asymptotic distribution such that

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} & \Sigma_{14} \\ \Sigma_{12} & \Sigma_{22} & \Sigma_{23} & \Sigma_{24} \\ \Sigma_{13} & \Sigma_{23} & \Sigma_{33} & \Sigma_{34} \\ \Sigma_{14} & \Sigma_{24} & \Sigma_{34} & \Sigma_{44} \end{pmatrix},$$

where

$$\begin{aligned}
&\Sigma_{11} = \frac{|P(h\cdot g_2)| - P^2(h\cdot g_2)}{P^2(g_2)}, \Sigma_{12} = \frac{-|P(h\cdot g_2)|}{P(g_2)} \frac{|P(h\cdot g_1)|}{P(g_1)}, \\
&\Sigma_{13} = \frac{P(h\cdot g_2)}{P(g_2)} - P(h\cdot g_2), \Sigma_{14} = \frac{-P(h\cdot g_2)}{P(g_2)} P(g_1), \\
&\Sigma_{22} = \frac{|P(h\cdot g_1)| - P^2(h\cdot g_1)}{P^2(g_1)}, \Sigma_{23} = \frac{-P(h\cdot g_1)}{P(g_1)} P(g_2), \\
&\Sigma_{24} = \frac{P(h\cdot g_1)}{P(g_1)} - P(h\cdot g_1), \Sigma_{33} = P(g_2) - P^2(g_2), \\
&\Sigma_{34} = -P(g_2)P(g_1), \Sigma_{44} = P(g_1) - P^2(g_1).
\end{aligned}$$

We can write

$$\begin{aligned}
&\sqrt{N}\left(\hat{\phi}_K(h,g) - \phi_K(h,g)\right) \\
&= \left(1, -1, -\frac{\frac{1}{N}\sum_{i=1}^N (h\cdot g_2)(Y_i,D_i,Z_i)}{\frac{1}{N}\sum_{i=1}^N g_2(Y_i,D_i,Z_i)P(g_2)}, \frac{\frac{1}{N}\sum_{i=1}^N (h\cdot g_1)(Y_i,D_i,Z_i)}{\frac{1}{N}\sum_{i=1}^N g_1(Y_i,D_i,Z_i)P(g_1)}\right)\cdot S_N.
\end{aligned}$$

Notice that

$$\begin{aligned}
&\left(1, -1, -\frac{\frac{1}{N}\sum_{i=1}^N (h\cdot g_2)(Y_i,D_i,Z_i)}{\frac{1}{N}\sum_{i=1}^N g_2(Y_i,D_i,Z_i)P(g_2)}, \frac{\frac{1}{N}\sum_{i=1}^N (h\cdot g_1)(Y_i,D_i,Z_i)}{\frac{1}{N}\sum_{i=1}^N g_1(Y_i,D_i,Z_i)P(g_1)}\right) \\
&\to \left(1, -1, -\frac{P(h\cdot g_2)}{P^2(g_2)}, \frac{P(h\cdot g_1)}{P^2(g_1)}\right) = A \text{ a.s.}
\end{aligned}$$

which implies

$$\sqrt{N}\left[\hat{\phi}_K(h,g)-\phi_K(h,g)\right] \rightsquigarrow N\left(0,A\Sigma A^T\right).$$

Also,

$$\begin{aligned}
A\Sigma A^T =& \left(\frac{|P(h\cdot g_2)|}{P^2(g_2)}-\frac{P^2(h\cdot g_2)}{P^3(g_2)},-\frac{|P(h\cdot g_1)|}{P^2(g_1)}+\frac{P^2(h\cdot g_1)}{P^3(g_1)},0,0\right)\\
&\cdot\left(1,-1,-\frac{P(h\cdot g_2)}{P^2(g_2)},\frac{P(h\cdot g_1)}{P^2(g_1)}\right)^T\\
=&\frac{|P(h\cdot g_2)|}{P^2(g_2)}-\frac{P^2(h\cdot g_2)}{P^3(g_2)}+\frac{|P(h\cdot g_1)|}{P^2(g_1)}-\frac{P^2(h\cdot g_1)}{P^3(g_1)}\\
=&\frac{|P(h\cdot g_2)|}{P^2(g_2)}\left(1-\frac{|P(h\cdot g_2)|}{P(g_2)}\right)+\frac{|P(h\cdot g_1)|}{P^2(g_1)}\left(1-\frac{|P(h\cdot g_1)|}{P(g_1)}\right).
\end{aligned}$$

This verifies (1.19).

For $\left(\sqrt{N}\left(\hat{\phi}_K(h_1,g_1)-\phi_K(h_1,g_1)\right),\ldots,\sqrt{N}\left(\hat{\phi}_K(h_t,g_t)-\phi_K(h_t,g_t)\right)\right)^T$ with some integer t, similar results as above hold as well, that is, the marginal convergence holds.

Remember we defined a metric on $\mathscr{H}_K\times\mathscr{G}$ by

$$\rho_P\left((h,g),\left(h',g'\right)\right)=\left\|h-h'\right\|_{L^2(P)}+\left\|g_1-g_1'\right\|_{L^2(P)}+\left\|g_2-g_2'\right\|_{L^2(P)}.$$

By Lemma A.1.10, $\mathscr{H}_K\times\mathscr{G}$ is totally bounded under ρ_P. Let

$$X_N(h,g)=\sqrt{N}\left(\hat{\phi}_K(h,g)-\phi_K(h,g)\right).$$

Now we consider asymptotic uniform ρ_P-equicontinuity of X_N in probability. Define another function space by $\mathscr{V}=\left\{v\in L^2(P): v=h\cdot g_K \text{ for some } h\in\mathscr{H}_K \text{ and } g_K\in\mathscr{G}_K\right\}$. Define the empirical process on $\mathscr{V}$ by

$$G_N(v)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{v(Y_i,D_i,Z_i)-E\left[v(Y_i,D_i,Z_i)\right]\right\}.$$

Notice that for each probability measure R on $\mathbb{R}^3$, we have

$$\begin{aligned}\|v_1 - v_2\|_{L^r(R)} &= \|h_1 \cdot g_{K1} - h_2 \cdot g_{K2}\|_{L^r(R)} \\ &\leq \|h_1 \cdot g_{K1} - h_2 \cdot g_{K1}\|_{L^r(R)} + \|h_2 \cdot g_{K1} - h_2 \cdot g_{K2}\|_{L^r(R)} \\ &\leq \|h_1 - h_2\|_{L^r(R)} + \|g_{K1} - g_{K2}\|_{L^r(R)}.\end{aligned}$$

Thus, together with Lemma A.1.5, with $F = 1$ being the envelop function of $\mathscr{V}$, we have for all $\varepsilon > 0$,

$$N\left(\varepsilon \|F\|_{L^r(R)}, \mathscr{V}, L^r(R)\right) \leq \max\left\{N\left(\frac{\varepsilon}{2}, \mathscr{H}_K, L^2(P)\right), N\left(\frac{\varepsilon}{2}, \mathscr{G}_K, L^2(P)\right)\right\}.$$

Similarly to Lemma A.1.6, $\mathscr{V}$ is a R-Donsker. Then for all $\varepsilon, \eta > 0$, by Example 1.5.10 in van der Vaart & Wellner (1996), there is $\delta_G > 0$ such that

$$\limsup_{N\to\infty} \mathbb{P}^* \left(\sup_{\|v - v'\|_{L^2(P)} < \delta_G} \left| G_N(v) - G_N\left(v'\right) \right| > \varepsilon \right) < \eta.$$

Define $H_N(u) = (\sqrt{N})^{-1} \sum_{i=1}^N \{u(Y_i, D_i, Z_i) - P(u)\}$ for all $u \in \mathscr{G}_K$ and it is easy to show that $\mathscr{G}_K$ is a P-Donsker. This implies for all $\varepsilon, \eta > 0$, there is $\delta_H > 0$ such that

$$\limsup_{N\to\infty} \mathbb{P}^* \left(\sup_{\|u - u'\|_{L^2(P)} < \delta_H} \left| H_N(u) - H_N\left(u'\right) \right| > \varepsilon \right) < \eta.$$

As defined before,

$$\left|X_N(h,g) - X_N\left(h',g'\right)\right| = \left| \left[\sqrt{N}\left(\hat{\phi}_K(h,g) - \phi_K(h,g)\right)\right] - \left[\sqrt{N}\left(\hat{\phi}_K\left(h',g'\right) - \phi_K\left(h',g'\right)\right)\right] \right|.$$

Then by (A.7), for all $\varepsilon > 0$,

$$\begin{aligned}
&\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}\left|X_N(h,g)-X_N\left(h',g'\right)\right|>\varepsilon\right)\\
\leq&\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}\left|\frac{G_N(h\cdot g_2)}{P(g_2)}-\frac{G_N(h'\cdot g_2')}{P\left(g_2'\right)}\right|>\frac{\varepsilon}{4}\right)\\
&+\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}\left|\frac{G_N(h\cdot g_1)}{P(g_1)}-\frac{G_N(h'\cdot g_1')}{P\left(g_1'\right)}\right|>\frac{\varepsilon}{4}\right)\\
&+\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}|DH_{N2}|>\frac{\varepsilon}{4}\right)+\mathbb{P}\left(\sup_{\rho_P((h,g),(h',g'))<\delta}|DH_{N1}|>\frac{\varepsilon}{4}\right),
\end{aligned}\tag{A.8}$$

where

$$DH_{N2}=H_N(g_2)\frac{\frac{1}{N}\sum_{i=1}^N(h\cdot g_2)(Y_i,D_i,Z_i)}{\frac{1}{N}\sum_{i=1}^N g_2(Y_i,D_i,Z_i)P(g_2)}-H_N\left(g_2'\right)\frac{\frac{1}{N}\sum_{i=1}^N(h\cdot g_2')(Y_i,D_i,Z_i)}{\frac{1}{N}\sum_{i=1}^N g_2'(Y_i,D_i,Z_i)P\left(g_2'\right)},$$

and

$$DH_{N1}=H_N(g_1)\frac{\frac{1}{N}\sum_{i=1}^N(h\cdot g_1)(Y_i,D_i,Z_i)}{\frac{1}{N}\sum_{i=1}^N g_1(Y_i,D_i,Z_i)P(g_1)}-H_N\left(g_1'\right)\frac{\frac{1}{N}\sum_{i=1}^N(h\cdot g_1')(Y_i,D_i,Z_i)}{\frac{1}{N}\sum_{i=1}^N g_1'(Y_i,D_i,Z_i)P\left(g_1'\right)}.$$

We now consider each term on the right-hand side of the inequality in (A.8). First we have

$$\begin{aligned}
&\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}\left|\frac{G_N(h\cdot g_2)}{P(g_2)}-\frac{G_N(h'\cdot g_2')}{P\left(g_2'\right)}\right|>\frac{\varepsilon}{4}\right)\\
\leq&\mathbb{P}^*\left(\sup_{\|v-v'\|_{L^2(P)}<\delta}\left|G_N(v)-G_N\left(v'\right)\right|>\frac{\varepsilon}{8}\min_{k\leq K}\mathbb{P}(Z=z_k)\right)\\
&+\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}\left|\frac{G_N(h'\cdot g_2')}{P(g_2)}-\frac{G_N(h'\cdot g_2')}{P\left(g_2'\right)}\right|>\frac{\varepsilon}{8}\right)
\end{aligned}$$

and similarly,

$$\begin{aligned}
&\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}\left|G_N(h\cdot g_1)-G_N\left(h'\cdot g_1'\right)\right|>\frac{\varepsilon}{4}\right)\\
\leq&\mathbb{P}^*\left(\sup_{\|v-v'\|_{L^2(P)}<\delta}\left|G_N(v)-G_N\left(v'\right)\right|>\frac{\varepsilon}{8}\min_{k\leq K}\mathbb{P}(Z=z_k)\right)\\
&+\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}\left|\frac{G_N(h'\cdot g_1')}{P(g_1)}-\frac{G_N(h'\cdot g_1')}{P\left(g_1'\right)}\right|>\frac{\varepsilon}{8}\right).
\end{aligned}$$

Also,

$$\begin{aligned}
&\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}|DH_{N2}|>\frac{\varepsilon}{4}\right)\\
\leq&\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}\left|\left[H_N(g_2)-H_N\left(g_2'\right)\right]\cdot\frac{\frac{1}{N}\sum_{i=1}^N(h\cdot g_2)(Y_i,D_i,Z_i)}{\frac{1}{N}\sum_{i=1}^N g_2(Y_i,D_i,Z_i)P(g_2)}\right|>\frac{\varepsilon}{8}\right)\\
&+\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}\left|H_N\left(g_2'\right)\cdot DR_{N2}\right|>\frac{\varepsilon}{8}\right),
\end{aligned}\tag{A.9}$$

where

$$DR_{N2}=\frac{\frac{1}{N}\sum_{i=1}^N(h\cdot g_2)(Y_i,D_i,Z_i)}{\frac{1}{N}\sum_{i=1}^N g_2(Y_i,D_i,Z_i)P(g_2)}-\frac{\frac{1}{N}\sum_{i=1}^N(h\cdot g_2')(Y_i,D_i,Z_i)}{\frac{1}{N}\sum_{i=1}^N g_2'(Y_i,D_i,Z_i)P\left(g_2'\right)}.$$

Now we consider the two terms on the right-hand side of the inequality (A.9).

We have that

$$\begin{aligned}
&\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}\left|\left[H_N(g_2)-H_N\left(g_2'\right)\right]\cdot\frac{\frac{1}{N}\sum_{i=1}^N(h\cdot g_2)(Y_i,D_i,Z_i)}{\frac{1}{N}\sum_{i=1}^N g_2(Y_i,D_i,Z_i)P(g_2)}\right|>\frac{\varepsilon}{8}\right)\\
\leq&\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}\left|\left[H_N(g_2)-H_N\left(g_2'\right)\right]\cdot\frac{P(h\cdot g_2)}{P^2(g_2)}\right|>\frac{\varepsilon}{16}\right)\\
&+\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}\left|\begin{array}{c}[H_N(g_2)-H_N(g_2')]\\ \cdot\left[\frac{\frac{1}{N}\sum_{i=1}^N(h\cdot g_2)(Y_i,D_i,Z_i)}{\frac{1}{N}\sum_{i=1}^N g_2(Y_i,D_i,Z_i)P(g_2)}-\frac{P(h\cdot g_2)}{P^2(g_2)}\right]\end{array}\right|>\frac{\varepsilon}{16}\right)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}^*\left(\sup_{\rho((h,g),(h',g'))<\delta}\left|H_N\left(g_2'\right)\cdot DR_{N2}\right|>\frac{\varepsilon}{8}\right)\\
\leq&\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}\left|H_N\left(g_2'\right)\cdot\left[\frac{P\left(h\cdot g_2\right)}{P^2\left(g_2\right)}-\frac{P\left(h\cdot g_2'\right)}{P^2\left(g_2'\right)}\right]\right|>\frac{\varepsilon}{16}\right)\\
&+\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}\left|H_N\left(g_2'\right)\cdot\left\{DR_{N2}-EDR_{N2}\right\}\right|>\frac{\varepsilon}{16}\right),
\end{aligned}$$

where

$$EDR_{N2}=\frac{P\left(h\cdot g_2\right)}{P^2\left(g_2\right)}-\frac{P\left(h\cdot g_2'\right)}{P^2\left(g_2'\right)}.$$

When δ is small enough,

$$\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}\left|DH_{N2}\right|>\frac{\varepsilon}{4}\right)<\frac{\eta}{4}.$$

Similarly, when δ is small enough,

$$\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}\left|DH_{N1}\right|>\frac{\varepsilon}{4}\right)<\frac{\eta}{4}.$$

Notice that

$$\left\|h\cdot g_1-h'\cdot g_1'\right\|_{L^2(P)}\leq\left\|h\cdot g_1-h'\cdot g_1\right\|_{L^2(P)}+\left\|h'\cdot g_1-h'\cdot g_1'\right\|_{L^2(P)}\leq\rho_P\left(\left(h,g\right),\left(h',g'\right)\right).$$

Because $\mathscr{V}$ is a R-Donsker, for any $\varepsilon,\eta>0$, when δ is small enough,

$$\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}\left|G_N\left(h\cdot g_2\right)-G_N\left(h'\cdot g_2'\right)\right|>\frac{\varepsilon}{8}\min_{k\leq K}\mathbb{P}\left(Z=z_k\right)\right)<\frac{\eta}{8},$$

$$\limsup_{N\to\infty}\mathbb{P}^*\left(\sup_{\rho_P((h,g),(h',g'))<\delta}\left|G_N\left(h\cdot g_1\right)-G_N\left(h'\cdot g_1'\right)\right|>\frac{\varepsilon}{8}\min_{k\leq K}\mathbb{P}\left(Z=z_k\right)\right)<\frac{\eta}{8}.$$

Also, when δ is small, $\rho_P((h,g),(h',g')) < \delta$ implies $\|g_1 - g_1'\|_{L^2(P)} < \delta$ and $\|g_2 - g_2'\|_{L^2(P)} < \delta$, and this implies $|P(g_1 - g_1')| < \delta^2$ and $|P(g_2 - g_2')| < \delta^2$. Thus, small δ implies

$$\limsup_{N\to\infty} \mathbb{P}^* \left(\sup_{\rho_P((h,g),(h',g'))<\delta} \left| \frac{G_N(h' \cdot g_2')}{P(g_2)} - \frac{G_N(h' \cdot g_2')}{P(g_2')} \right| > \frac{\varepsilon}{8} \right) < \frac{\eta}{8}$$

and

$$\limsup_{N\to\infty} \mathbb{P}^* \left(\sup_{\rho_P((h,g),(h',g'))<\delta} \left| \frac{G_N(h' \cdot g_1')}{P(g_1)} - \frac{G_N(h' \cdot g_1')}{P(g_1')} \right| > \frac{\varepsilon}{8} \right) < \frac{\eta}{8}.$$

Putting all above together, we have

$$\limsup_{N\to\infty} \mathbb{P}^* \left(\sup_{\rho_P((h,g),(h',g'))<\delta} \left| X_N(h,g) - X_N(h',g') \right| > \varepsilon \right) < \eta,$$

which indicates X_N is asymptotically uniformly ρ_P-equicontinuous in probability.

With the marginal weak convergence of $\sqrt{N}\{\hat{\phi}_K - \phi_K\}$ and total boundedness of $\mathscr{H}_K \times \mathscr{G}$ by Lemma A.1.10, we can conclude that

$$\sqrt{N}\{\hat{\phi}_K - \phi_K\} \rightsquigarrow \mathbb{G}_K,$$

for some Gaussian process $\mathbb{G}_K$. ■

Proof of Theorem 1.4.1.

By Lemmas A.1.10 and A.1.11, $\mathscr{H}_K \times \mathscr{G}$ is compact under ρ_P.

Let $\varphi_N = \phi_K/(\xi \vee \hat{\sigma}_{KN})$. Given each sequence $r_N \to \infty$, define for all $\omega \in \Omega$,

$$\mathbb{D}_N(\omega) = \left\{ \varphi \in \ell^\infty(\mathscr{H}_K \times \mathscr{G}) : \varphi_N(\omega) + r_N^{-1}\varphi \in \ell^\infty(\mathscr{H}_K \times \mathscr{G}) \right\}.$$

For all $\omega \in \Omega$, define

$$g_N(\omega)(\psi) = r_N\left(\mathscr{S}_K\left(\varphi_N(\omega) + r_N^{-1}\psi\right) - \mathscr{S}_K(\varphi_N(\omega)) \right)$$

for all $\psi \in \mathbb{D}_N(\omega)$. We know $\varphi_N \to \phi_K/(\xi \vee \sigma_K) = \varphi_0$ a.s. Let $\Omega_0 = \{\omega \in \Omega : \varphi_N(\omega) \to \varphi_0\}$ and $\mathbb{P}(\Omega_0) = 1$. Now we want to show that there is some g, for all $\omega \in \Omega_0$, $g_N(\omega)(\psi_N) \to g(\psi)$ for all $\psi_N \in \mathbb{D}_N(\omega)$ with $\psi_N \to \psi$ for some $\psi \in C(\mathscr{H}_K \times \mathscr{G})$. Now we fix an $\omega \in \Omega_0$.

Given $\{Y_i, D_i\}_{i=1}^N$, P_N has finitely many possible values on $\mathscr{H}_K \times \mathscr{G}$. Suppose there are in total J_N possible values of P_N. Since ϕ_K is continuous on $\mathscr{H}_K \times \mathscr{G}$, for all $\psi \in C(\mathscr{H}_K \times \mathscr{G})$, $\sup_{(h,g)\in\mathscr{H}_K\times\mathscr{G}}(\varphi_N + t_N\psi)(h,g) < \infty$.

We define a correspondence $\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}} : \ell^\infty(\mathscr{H}_K \times \mathscr{G}) \to \mathbb{R}$ by

$$\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\psi) = \{(h,g) \in \mathscr{H}_K \times \mathscr{G} : \psi(h,g) = \mathscr{S}_K(\psi)\},$$

for all $\psi \in \ell^\infty(\mathscr{H}_K \times \mathscr{G})$. By the proof similar to that of Theorem 1.3.1, we can show that $\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}$ is upper hemicontinuous at φ_0.

Under H_0, $\sup_{(h,g)\in\mathscr{H}_K\times\mathscr{G}} \varphi_N(h,g) = 0$. And it is easy to show that under H_0,

$$\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\varphi_N) = \bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\varphi_0) = \bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\phi_K).$$

Since $\varphi_N + t_N\psi$ is not continuous on $\mathscr{H}_K \times \mathscr{G}$, $\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\varphi_N + t_N\psi)$ may be empty. As we have shown earlier, P_N at most has J_N possible values. We can construct a modified version of φ_N, denoted by φ'_N such that φ'_N is upper semicontinuous and

(i) $\sup_{(h,g)\in\mathscr{H}_K\times\mathscr{G}} \varphi_N(h,g) = \sup_{(h,g)\in\mathscr{H}_K\times\mathscr{G}} \varphi'_N(h,g)$,

(ii) $\sup_{(h,g)\in\mathscr{H}_K\times\mathscr{G}} (\varphi_N + t_N\psi)(h,g) = \sup_{(h,g)\in\mathscr{H}_K\times\mathscr{G}} (\varphi'_N + t_N\psi)(h,g)$,

(iii) $\varphi'_N + t_N\psi \to \varphi_0$, as $N \to \infty$,

by a similar strategy to that of Theorem 1.3.1. Then $\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\varphi'_N + t_N\psi) \neq \varnothing$.

Let $t_N = r_N^{-1}$. It is easy to show that

$$\left| \sup_{(h,g)\in\mathscr{H}_K\times\mathscr{G}} \{\varphi_N(h,g) + t_N\psi_N(h,g)\} - \sup_{(h,g)\in\mathscr{H}_K\times\mathscr{G}} \{\varphi_N(h,g) + t_N\psi(h,g)\} \right|$$
$$\leq t_N \|\psi_N - \psi\|_\infty = o(t_N).$$

As discussed before, $\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\varphi'_N + t_N\psi)$ is nonempty. Since $\varphi'_N + t_N\psi$ converges to φ_0 and $\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}$ is upper hemicontinuous at φ_0, there is a sequence δ_N such that

$$\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}\left(\varphi'_N + t_N\psi\right) \subset \bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\varphi_0)^{\delta_N},$$

where

$$\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\varphi_0)^{\delta_N} = \left\{ (h,g) \in \mathscr{H}_K\times\mathscr{G} : \inf_{(h',g')\in\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\varphi_0)} \rho_P\left((h,g),(h',g')\right) \leq \delta_N \right\}.$$

Remember that under H_0,

$$\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\varphi_0) = \bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\varphi_N) \text{ and } \sup_{(h,g)\in\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\varphi_0)} \varphi_N(h,g) = 0.$$

Thus, we have that under H_0,

$$\left| \sup_{(h,g)\in\mathscr{H}_K\times\mathscr{G}} \{\varphi_N(h,g) + t_N\psi(h,g)\} - \sup_{(h,g)\in\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\varphi_0)} \{\varphi_N(h,g) + t_N\psi(h,g)\} \right|$$
$$\leq \sup_{(h,g)\in\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\varphi_0)^{\delta_N}} \left\{\varphi'_N(h,g) + t_N\psi(h,g)\right\} - \sup_{(h,g)\in\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\varphi_0)} t_N\psi(h,g)$$
$$\leq \sup_{(h_1,g_1),(h_2,g_2)\in\mathscr{H}_K\times\mathscr{G},\rho_P((h_1,g_1),(h_2,g_2))\leq\delta_N} t_N|\psi(h_1,g_1) - \psi(h_2,g_2)| = o(t_N).$$

Finally,

$$\left|\sup_{(h,g)\in\mathscr{H}_K\times\mathscr{G}}\{\varphi_N(h,g)+t_N\psi_N(h,g)\}-\sup_{(h,g)\in\mathscr{H}_K\times\mathscr{G}}\varphi_N(h,g)-t_N\sup_{(h,g)\in\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\varphi_0)}\psi(h,g)\right|$$
$$\leq\left|\sup_{(h,g)\in\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\varphi_0)}\{\varphi_N(h)+t_N\psi(h)\}-t_N\sup_{(h,g)\in\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\varphi_0)}\psi(h,g)\right|+o(t_N)=o(t_N).$$

This implies that $g_N(\omega)(\psi_N)\to\sup_{(h,g)\in\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\varphi_0)}\psi(h)$.

By Lemmas A.1.2 and 1.4.1,

$$\sqrt{N}\left(\mathscr{S}_K\left(\frac{\hat{\phi}_K}{\xi\vee\hat{\sigma}_{KN}}\right)-\mathscr{S}_K\left(\frac{\phi}{\xi\vee\hat{\sigma}_{KN}}\right)\right)\rightsquigarrow\mathscr{S}_{\Psi_{\mathscr{H}_K\times\mathscr{G}}}\left(\frac{\mathbb{G}_K}{\xi\vee\sigma_K}\right),$$

where by definition, for all $\psi\in\ell^\infty(\mathscr{H}_K\times\mathscr{G})$,

$$\mathscr{S}_{\Psi_{\mathscr{H}_K\times\mathscr{G}}}(\psi)=\sup_{(h,g)\in\Psi_{\mathscr{H}_K\times\mathscr{G}}}\psi(h,g)=\sup_{(h,g)\in\bar{\Psi}_{\mathscr{H}_K\times\mathscr{G}}(\varphi_0)}\psi(h,g),$$

under H_0. ■

For all sets $A_1,A_2\subset\mathscr{H}_K\times\mathscr{G}$, define

$$\overrightarrow{d}_H(A_1,A_2)=\sup_{a\in A_1}\inf_{b\in A_2}\rho_P(a,b),$$

and

$$d_H(A_1,A_2)=\max\left\{\overrightarrow{d}_H(A_1,A_2),\overrightarrow{d}_H(A_2,A_1)\right\}.$$

The following lemma concludes that $\hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}}$ in (1.21) is valid in the sense that $\hat{\mathscr{S}}_{KN}$ satisfies Assumption 3.3 in Fang & Santos (2014).

Lemma A.3.1 *Under Assumption 1.3.1 , if H_0 is true, $d_H\left(\hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}},\Psi_{\mathscr{H}_K\times\mathscr{G}}\right)\to_p 0$.*

Proof of Lemma A.3.1.

The proof strategy is similar to that of Lemma A.2.1.

First, for all $\varepsilon > 0$,

$$\begin{aligned}
\lim_{N\to\infty} \mathbb{P}\left(\overrightarrow{d}_H\left(\Psi_{\mathscr{H}_K\times\mathscr{G}}, \hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}}\right) > \varepsilon\right) &\leq \lim_{N\to\infty} \mathbb{P}\left(\Psi_{\mathscr{H}_K\times\mathscr{G}} \backslash \hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}} \neq \varnothing\right) \\
&\leq \lim_{N\to\infty} \mathbb{P}\left(\sup_{(h,g)\in\Psi_{\mathscr{H}_K\times\mathscr{G}}\backslash\hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}}} \left|\hat{\phi}_K(h,g) - \phi_K(h,g)\right| > \tau_N\right) \\
&\leq \lim_{N\to\infty} \mathbb{P}\left(\sup_{(h,g)\in\mathscr{H}_K\times\mathscr{G}} \sqrt{T_N}\left|\hat{\phi}_K(h,g) - \phi_K(h,g)\right| > \sqrt{T_N}\tau_N\right).
\end{aligned}$$

By Theorem 1.4.1, $\sqrt{N}\left(\hat{\phi}_K - \phi_K\right) \rightsquigarrow \mathbb{G}_K$. Then by Theorem 1.3.6 in van der Vaart & Wellner (1996), $\sup_{(h,g)\in\mathscr{H}_K\times\mathscr{G}} \sqrt{N}\left|\hat{\phi}_K(h,g) - \phi_K(h,g)\right| \rightsquigarrow \sup_{(h,g)\in\mathscr{H}_K\times\mathscr{G}} |\mathbb{G}_K(h,g)|$. If $\sqrt{N}\tau_N \to \infty$, then $\lim_{N\to\infty}\mathbb{P}\left(\overrightarrow{d}_H\left(\Psi_{\mathscr{H}_K\times\mathscr{G}}, \hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}}\right) > \varepsilon\right) = 0$.

Next, consider $\overrightarrow{d}_H\left(\hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}}, \Psi_{\mathscr{H}_K\times\mathscr{G}}\right)$. Define

$$d\left((h,g), \Psi_{\mathscr{H}_K\times\mathscr{G}}\right) = \inf_{(h',g')\in\Psi_{\mathscr{H}_K\times\mathscr{G}}} \rho_P\left((h,g), (h',g')\right)$$

for all $(h,g) \in \mathscr{H}_K \times \mathscr{G}$. For each $\varepsilon > 0$, let $D_\varepsilon = \{(h,g) \in \mathscr{H}_K \times \mathscr{G} : d\left((h,g), \Psi_{\mathscr{H}_K\times\mathscr{G}}\right) \geq \varepsilon\}$. We have shown that $\mathscr{H}_K \times \mathscr{G}$ is compact under ρ_P by Lemmas A.1.10 and A.1.11. Suppose there is $\{(h_n, g_n)\}_n \subset D_\varepsilon$ and $(h_n, g_n) \to (h,g)$ for some $(h,g) \in \mathscr{H}_K \times \mathscr{G}$, then

$$\begin{aligned}
d\left((h,g), \Psi_{\mathscr{H}_K\times\mathscr{G}}\right) &= \inf_{(h',g')\in\Psi_{\mathscr{H}_K\times\mathscr{G}}} \rho_P\left((h,g), (h',g')\right) \\
&\geq \inf_{(h',g')\in\Psi_{\mathscr{H}_K\times\mathscr{G}}} \rho_P\left((h_n,g_n), (h',g')\right) - \rho_P\left((h,g), (h_n,g_n)\right) \geq \varepsilon - \rho_P\left((h,g), (h_n,g_n)\right),
\end{aligned}$$

which is true for all n. Letting $n \to \infty$ gives us $d\left((h,g), \Psi_{\mathscr{H}_K\times\mathscr{G}}\right) \geq \varepsilon$. This implies D_ε is closed in $\mathscr{H}_K \times \mathscr{G}$ which is compact and thus D_ε is compact. If $D_\varepsilon \neq \varnothing$, then $\exists \delta_\varepsilon > 0$ such that

$\inf_{(h,g)\in D_\varepsilon} |\phi_K(h,g)| > \delta_\varepsilon$. Also,

$$
\begin{aligned}
&\lim_{N\to\infty} \mathbb{P}\left(\overrightarrow{d}_H\left(\hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}}, \Psi_{\mathscr{H}_K\times\mathscr{G}}\right) > \varepsilon\right) \\
=&\lim_{N\to\infty} \mathbb{P}\left(\sup_{(h,g)\in\hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}}} \inf_{(h',g')\in\Psi_{\mathscr{H}_K\times\mathscr{G}}} \rho_P\left((h,g),\left(h',g'\right)\right) > \varepsilon\right) \\
\leq&\lim_{N\to\infty} \mathbb{P}\left(\sup_{(h,g)\in\hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}}\setminus\Psi_{\mathscr{H}_K\times\mathscr{G}}} |\phi_K(h,g)| > \delta_\varepsilon, \sup_{(h,g)\in\hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}}\setminus\Psi_{\mathscr{H}_K\times\mathscr{G}}} \left|\hat{\phi}_K(h,g)\right| \leq \tau_N\right).
\end{aligned}
$$

Here we define events

$$
A_N = \left\{ \begin{array}{c} \sup_{(h,g)\in\hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}}\setminus\Psi_{\mathscr{H}_K\times\mathscr{G}}} |\phi_K(h,g)| - \frac{\delta_\varepsilon}{2} \leq \sup_{(h,g)\in\hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}}\setminus\Psi_{\mathscr{H}_K\times\mathscr{G}}} \left|\hat{\phi}_K(h,g)\right| \\ \leq \sup_{(h,g)\in\hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}}\setminus\Psi_{\mathscr{H}_K\times\mathscr{G}}} |\phi_K(h,g)| + \frac{\delta_\varepsilon}{2} \end{array} \right\}.
$$

Now we have

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{(h,g)\in\mathscr{H}_K\times\mathscr{G}} \left|\hat{\phi}_K(h,g) - \phi_K(h,g)\right| \leq \frac{\delta_\varepsilon}{2}\right) \\
\leq&\mathbb{P}\left(\sup_{(h,g)\in\hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}}\setminus\Psi_{\mathscr{H}_K\times\mathscr{G}}} \left|\hat{\phi}_K(h,g) - \phi_K(h,g)\right| \leq \frac{\delta_\varepsilon}{2}\right) \\
\leq&\mathbb{P}\left(\left|\sup_{(h,g)\in\hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}}\setminus\Psi_{\mathscr{H}_K\times\mathscr{G}}} \left|\hat{\phi}_K(h,g)\right| - \sup_{(h,g)\in\hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}}\setminus\Psi_{\mathscr{H}_K\times\mathscr{G}}} |\phi_K(h,g)|\right| \leq \frac{\delta_\varepsilon}{2}\right) \\
=&\mathbb{P}(A_N).
\end{aligned}
$$

By Lemma A.1.8, $\lim_{N\to\infty}\mathbb{P}(A_N)=1$. Thus,

$$
\begin{aligned}
&\lim_{N\to\infty}\mathbb{P}\left(\overrightarrow{d}_H\left(\hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}},\Psi_{\mathscr{H}_K\times\mathscr{G}}\right)>\varepsilon\right)\\
\leq&\lim_{N\to\infty}\mathbb{P}\left(\sup_{(h,g)\in\hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}}\setminus\Psi_{\mathscr{H}_K\times\mathscr{G}}}|\phi_K(h,g)|>\delta_\varepsilon,\sup_{(h,g)\in\hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}}\setminus\Psi_{\mathscr{H}_K\times\mathscr{G}}}\left|\hat{\phi}_K(h,g)\right|\leq\tau_N,A_N\right)\\
\leq&\lim_{N\to\infty}\mathbb{P}\left(\sup_{(h,g)\in\hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}}\setminus\Psi_{\mathscr{H}_K\times\mathscr{G}}}\left|\hat{\phi}_K(h,g)\right|\geq\frac{\delta_\varepsilon}{2},\sup_{(h,g)\in\hat{\Psi}_{\mathscr{H}_K\times\mathscr{G}}\setminus\Psi_{\mathscr{H}_K\times\mathscr{G}}}\left|\hat{\phi}_K(h,g)\right|\leq\tau_N\right)\\
=&0 \text{ as } \tau_N\downarrow 0.
\end{aligned}
$$

■

Appendix B

Proofs for Chapter 2

B.1 Main Results

Proof of Lemma 2.2.1.

Define $\mathscr{T} : L^1[0,1] \to \ell^\infty[0,1]$ by

$$\mathscr{T}(Q)(p) = \frac{\int_0^p Q(t)\mathrm{d}t}{\int_0^1 Q(t)\mathrm{d}t},$$

for any $Q \in L^1[0,1]$ and $p \in [0,1]$. For any $h_n, h \in L^1[0,1]$ such that $h_n \to h$,

$$\begin{aligned}
&\left(\frac{\int_0^p Q(t)+t_n h_n(t)\mathrm{d}t}{\int_0^1 Q(t)+t_n h_n(t)\mathrm{d}t} - \frac{\int_0^p Q(t)\mathrm{d}t}{\int_0^1 Q(t)\mathrm{d}t}\right)\frac{1}{t_n} \\
&- \frac{\int_0^p h(t)\mathrm{d}t \int_0^1 Q(t)\mathrm{d}t - \int_0^p Q(t)\mathrm{d}t \int_0^1 h(t)\mathrm{d}t}{(\int_0^1 Q(t)\mathrm{d}t)^2} \\
&= \left(\frac{\int_0^p Q(t)+t_n h_n(t)\mathrm{d}t \int_0^1 Q(t)\mathrm{d}t - \int_0^p Q(t)\mathrm{d}t \int_0^1 Q(t)+t_n h_n(t)\mathrm{d}t}{\int_0^1 Q(t)+t_n h_n(t)\mathrm{d}t \int_0^1 Q(t)\mathrm{d}t}\right)\frac{1}{t_n} \\
&- \frac{\int_0^p h(t)\mathrm{d}t \int_0^1 Q(t)\mathrm{d}t - \int_0^p Q(t)\mathrm{d}t \int_0^1 h(t)\mathrm{d}t}{(\int_0^1 Q(t)\mathrm{d}t)^2} = \frac{A_n(p)}{B_n(p)},
\end{aligned} \tag{B.1}$$

where

$$
\begin{aligned}
A(p) = & \left(\int_0^p h_n(t)\mathrm{d}t \int_0^1 Q(t)\mathrm{d}t - \int_0^p Q(t)\mathrm{d}t \int_0^1 h_n(t)\mathrm{d}t \right) \left(\int_0^1 Q(t)\mathrm{d}t \right)^2 \\
& - \left(\int_0^p h(t)\mathrm{d}t \int_0^1 Q(t)\mathrm{d}t - \int_0^p Q(t)\mathrm{d}t \int_0^1 h(t)\mathrm{d}t \right) \\
& \cdot \left[\left(\int_0^1 Q(t)\mathrm{d}t \right)^2 + t_n \left(\int_0^1 h_n(t)\mathrm{d}t \right) \left(\int_0^1 Q(t)\mathrm{d}t \right) \right] \\
\leq & 2 \left\| h_n - h \right\|_1 \left(\int_0^1 Q(t)\mathrm{d}t \right)^3 + O(t_n),
\end{aligned} \tag{B.2}
$$

and

$$
\begin{aligned}
B_n(p) = & \left[\left(\int_0^1 Q(t)\mathrm{d}t \right)^2 + t_n \left(\int_0^1 h_n(t)\mathrm{d}t \right) \left(\int_0^1 Q(t)\mathrm{d}t \right) \right] \left(\int_0^1 Q(t)\mathrm{d}t \right)^2 \\
= & \left(\int_0^1 Q(t)\mathrm{d}t \right)^4 + O(t_n).
\end{aligned} \tag{B.3}
$$

Then

$$
\sup_{p\in[0,1]} \left| \frac{A_n(p)}{B_n(p)} \right| \leq \frac{2 \left\| h_n - h \right\|_1 (\int_0^1 Q(t)\mathrm{d}t)^3 + O(t_n)}{(\int_0^1 Q(t)\mathrm{d}t)^4 + O(t_n)} \to 0.
$$

This implies $\mathscr{T}$ is Hadamard differentiable at Q tangentially to $L^1[0,1]$ with derivative

$$
\mathscr{T}'_Q(h)(p) = \frac{\int_0^p h(t)\mathrm{d}t \int_0^1 Q(t)\mathrm{d}t - \int_0^p Q(t)\mathrm{d}t \int_0^1 h(t)\mathrm{d}t}{(\int_0^1 Q(t)\mathrm{d}t)^2}. \tag{B.4}
$$

Notice that $\mathscr{Z}(F_j) = \mathscr{T} \circ \mathscr{V}(F_j)$. By Theorem 3 in Kaji (2017), $\mathscr{V}$ is Hadamard differentiable at F_j tangentially to $C_0[0,\infty) \cap \mathbb{L}$, with Hadamard derivative

$$
\mathscr{V}'_{F_j}(h)(p) = -\frac{h(F_j^{-1}(p))}{f_j(F_j^{-1}(p))},
$$

for all $h \in C_0[0,\infty) \cap \mathbb{L}$. It is easy to show that $\mathscr{V}'_{F_j}(h) \in L^1[0,1]$ when $h \in C_0[0,\infty) \cap \mathbb{L}$. Then by Lemma 3.9.3 (Chain rule) in van der Vaart & Wellner (1996), $\mathscr{T} \circ \mathscr{V}$ is Hadamard differentiable

at F_j tangentially to $C_0[0,\infty)\cap\mathbb{L}$ with derivative $\mathscr{T}'_{\mathscr{V}(F_j)}\circ\mathscr{V}'_{F_j}$. ■

Proof of Lemma 2.2.2.

By Proposition 2 in Kaji (2017), under Assumption 2.2.1, $\sqrt{n_j}(\hat{F}_j - F_j) \rightsquigarrow \mathbb{G}_j$ for some Gaussian process $\mathbb{G}_j$ in $\mathbb{L}$ with mean zero and covariance function $Cov(x,y) = F_j(x)\wedge F_j(y) - F_j(x)F_j(y)$. So we can write $\mathbb{G}_j = \mathscr{W}(F)$, where $\mathscr{W}$ is a standard Brownian bridge. Now we can write $\mathscr{W}(t) = \mathscr{B}(t) - t\mathscr{B}(1)$ for $t\in[0,1]$, where $\mathscr{B}$ is a standard Brownian motion with a.s. continuous paths. This implies $\mathbb{G}_j \in C_0[0,\infty)\cap\mathbb{L}$ a.s..

Then by Lemma 2.2.1 and Theorem 3.9.4 in van der Vaart & Wellner (1996) (Delta method),

$$\sqrt{n_j}(\hat{L}_j - L_j) \rightsquigarrow \mathscr{L}_j, \tag{B.5}$$

for some Gaussian process $\mathscr{L}_j$ with a.s. continuous paths.

Under Assumption 3.3.1, we know the two subsamples are independent, then

$$\begin{aligned}\sqrt{T_n}(\hat{\phi}-\phi) &= \sqrt{T_n}(\hat{L}_2 - L_2) - \sqrt{T_n}(\hat{L}_1 - L_1)\\ &\rightsquigarrow \sqrt{\lambda}\mathscr{L}_2 - \sqrt{1-\lambda}\mathscr{L}_1.\end{aligned} \tag{B.6}$$

■

Proof of Lemma 2.2.3.

Remember $\hat{F}_j^*$ is the bootstrap CDF obtained from sample $\{X_i^{j*}\}_{i=1}^{n_j}$. As we show before, $L_j = \mathscr{Z}(F_j)$.

Let

$$BL_1(\mathbb{L}) = \{h:\mathbb{L}\to\mathbb{R} : \sup_{G\in\mathbb{L}}|h(G)| < 1 \text{ and } |h(G_1)-h(G_2)| \le \|G_1 - G_2\|_{\mathbb{L}}\}.$$

By Lemmas A.16, A.17 and A.18 in Kaji (2017),

$$\sup_{h\in BL_1(\mathbb{L})}\left|E[h(\sqrt{n_j}(\hat{F}_j^*-\hat{F}_j))|\{X_i^1\}_{i=1}^{n_1},\{X_i^2\}_{i=1}^{n_2}]-E[h(\mathbb{G}_j)]\right|\to 0$$

in outer probability and $\sqrt{n_j}(\hat{F}_j^*-\hat{F}_j)$ is asymptotically measurable. Then by Lemma 2.2.1 and Theorem 3.9.11 in van der Vaart & Wellner (1996),

$$\sqrt{n_j}(\hat{L}_j^*-\hat{L}_j)\rightsquigarrow \mathscr{Z}'_{F_j}(\mathbb{G}_j), \tag{B.7}$$

in outer probability measure, where $\mathbb{G}_j$ is the asymptotic distribution of $\sqrt{n_j}(\hat{F}_j-F_j)$. Also Theorem 3.9.11 in van der Vaart & Wellner (1996) implies that $\sqrt{n_j}(\hat{L}_j^*-\hat{L}_j)$ is asymptotically measurable and therefore $\sqrt{T_n}(\hat{\phi}^*-\hat{\phi})$ is asymptotically measurable. The weak convergence result for $\hat{\phi}^*$ follows Assumption 3.3.1.

$\mathscr{Z}$ is a continuous map from $\mathbb{L}_F$ to $\ell^\infty[0,1]$. Thus it is a measurable map. It is not hard to see that $h(\sqrt{T_n}(\hat{\phi}^*-\hat{\phi}))$ is a measurable function of $\left\{W_i^1,W_i^2\right\}$ for any continuous and bounded h from the expression that

$$\begin{aligned}
&\sqrt{T_n}(\hat{\phi}^*-\hat{\phi})\\
=&\sqrt{T_n}([\mathscr{Z}(\hat{F}_2^*)-\mathscr{Z}(\hat{F}_1^*)]-[\mathscr{Z}(\hat{F}_2)-\mathscr{Z}(\hat{F}_1)])\\
=&\sqrt{T_n}([\mathscr{Z}(n_2^{-1}\sum_{i=1}^{n_2}W_i^2 1_{[X_i^2,\infty)})-\mathscr{Z}(n_1^{-1}\sum_{i=1}^{n_1}W_i^1 1_{[X_i^1,\infty)})]\\
&-[\mathscr{Z}(\hat{R}_2)-\mathscr{Z}(\hat{R}_1)]).
\end{aligned} \tag{B.8}$$

■

Proof of Lemma 2.2.4. Recall that the Hausdorff distance between $B(\phi)$ and $\hat{B}_n$ is

$$d_H(B(\phi),\hat{B}_n)=\max\left\{\overrightarrow{d}_H(B(\phi),\hat{B}_n),\overrightarrow{d}_H(\hat{B}_n,B(\phi))\right\}, \tag{B.9}$$

where $\overrightarrow{d}_H(A,B) = \sup_{a\in A}\inf_{b\in B}|a-b|$ for any sets A,B. For $\mathscr{S}$, by Lemma B.3 in Fang & Santos (2014), we only need to verify that

$$d_H(B(\phi),\hat{B}_n,B(\phi)) = o_P(1).$$

First, we consider $\overrightarrow{d}_H(B(\phi),\hat{B}_n)$, where

$$\overrightarrow{d}_H(B(\phi),\hat{B}_n) = \sup_{p_1\in B(\phi)}\inf_{p_2\in\hat{B}_n}|p_1-p_2|. \tag{B.10}$$

Then for any $\varepsilon > 0$,

$$\begin{aligned} &P(\overrightarrow{d}_H(B(\phi),\hat{B}_n) > \varepsilon) \leq \lim_{n\to\infty} P(B(\phi)\backslash\hat{B}_n \neq \varnothing) \\ \leq &P\left(\sup_{p\in B(\phi)\backslash\hat{B}_n}|\hat{\phi}(p)-\phi(p)| > \tau_n\right) \leq P\left(\sup_{p\in[0,1]}\sqrt{T_n}|\hat{\phi}(p)-\phi(p)| > \sqrt{T_n}\tau_n\right). \end{aligned} \tag{B.11}$$

With (2.13), by Theorem 1.3.6 (Continuous Mapping) in van der Vaart & Wellner (1996), it follows that

$$\sqrt{T_n}|\hat{\phi}-\phi| \rightsquigarrow |\sqrt{\lambda}\mathscr{L}_2 - \sqrt{1-\lambda}\mathscr{L}_1|. \tag{B.12}$$

And by Theorem 1.3.6 in van der Vaart & Wellner (1996) again,

$$\sup_{p\in[0,1]}\sqrt{T_n}|\hat{\phi}(p)-\phi(p)| \rightsquigarrow \sup_{p\in[0,1]}|\sqrt{\lambda}\mathscr{L}_2(p) - \sqrt{1-\lambda}\mathscr{L}_1(p)|. \tag{B.13}$$

So if $\sqrt{T_n}\tau_n \to \infty$, the limit probability in (B.11) is 0.

Next, consider $\overrightarrow{d}_H(\hat{B}_n,B(\phi))$. Define $d(p,B(\phi)) = \inf_{p'\in B(\phi)}|p-p'|$. For any $\varepsilon > 0$, $\exists\delta_\varepsilon > 0$ such that

$$\inf_{p\in[0,1],d(p,B(\phi))\geq\varepsilon}|\phi(p)| > \delta_\varepsilon, \tag{B.14}$$

if $\{p \in [0,1] : d(p,B(\phi)) \geq \varepsilon\} \neq \varnothing$. This is because ϕ is continuous and $\{p \in [0,1] : d(p,B(\phi)) \geq \varepsilon\}$ is compact.

To verify that $D_\varepsilon = \{p \in [0,1] : d(p,B(\phi)) \geq \varepsilon\}$ is compact we only prove that it is closed. Suppose there is a sequence $\{p_k\}$ s.t. $p_k \in D_\varepsilon$ and $p_k \to p$ then

$$\begin{aligned} d(p,B(\phi)) &= \inf_{p' \in B(\phi)} \left|p - p'\right| = \inf_{p' \in B(\phi)} \left|p - p_k + p_k - p'\right| \\ &\geq \inf_{p' \in B(\phi)} \left|p_k - p'\right| - |p - p_k| \geq \varepsilon - |p - p_k|, \end{aligned} \tag{B.15}$$

and this is true for any k. So letting $k \to \infty$ gives $d(p,B(\phi)) \geq \varepsilon$.

Then it follows that

$$\begin{aligned} & P\left(\overrightarrow{d}_H(\hat{B}_n, B(\phi)) \geq 2\varepsilon\right) = P\left(\sup_{p_1 \in \hat{B}_n} \inf_{p_2 \in B(\phi)} |p_1 - p_2| \geq 2\varepsilon\right) \\ \leq & P\left(\sup_{p \in \hat{B}_n \setminus B(\phi)} |\phi(p)| > \delta_\varepsilon, \sup_{p \in \hat{B}_n \setminus B(\phi)} |\hat{\phi}(p)| \leq \tau_n\right) \\ = & P\left(\sup_{p \in \hat{B}_n \setminus B(\phi)} |\phi(p)| > \delta_\varepsilon, \sup_{p \in \hat{B}_n \setminus B(\phi)} |\hat{\phi}(p)| \leq \tau_n, A_n\right) \\ & + P\left(\sup_{p \in \hat{B}_n \setminus B(\phi)} |\phi(p)| > \delta_\varepsilon, \sup_{p \in \hat{B}_n \setminus B(\phi)} |\hat{\phi}(p)| \leq \tau_n, A_n^c\right) \\ \leq & P\left(\sup_{p \in \hat{B}_n \setminus B(\phi)} |\hat{\phi}(p)| \geq \frac{\delta_\varepsilon}{2}, \sup_{p \in \hat{B}_n \setminus B(\phi)} |\hat{\phi}(p)| \leq \tau_n\right) + P(A_n^c) \\ \to & 0 \text{ as } \tau_n \downarrow 0, \end{aligned} \tag{B.16}$$

where

$$\begin{aligned} A_n = & \left\{\sup_{p \in \hat{B}_n \setminus B(\phi)} |\phi(p)| - \frac{\delta_\varepsilon}{2} \leq \sup_{p \in \hat{B}_n \setminus B(\phi)} |\hat{\phi}(p)|\right\} \\ & \cap \left\{\sup_{p \in \hat{B}_n \setminus B(\phi)} |\hat{\phi}(p)| \leq \sup_{p \in \hat{B}_n \setminus B(\phi)} |\phi(p)| + \frac{\delta_\varepsilon}{2}\right\}. \end{aligned}$$

In the second equality, we used one result that $P(A_n) \to 1$. This is because $\sup_{p\in\hat{B}_n\setminus B(\phi)} |\hat{\phi}(p) - \phi(p)| \to 0$ a.s. by implication of Theorem 11.1 in Csörgö *et al.* (1986) and

$$\left| \sup_{p\in\hat{B}_n\setminus B(\phi)} |\hat{\phi}(p)| - \sup_{p\in\hat{B}_n\setminus B(\phi)} |\phi(p)| \right| \leq \sup_{p\in\hat{B}_n\setminus B(\phi)} |\hat{\phi}(p) - \phi(p)|, \tag{B.17}$$

which implies $\left|\sup_{p\in\hat{B}_n\setminus B(\phi)} |\hat{\phi}(p)| - \sup_{p\in\hat{B}_n\setminus B(\phi)} |\phi(p)|\right| \to 0$ a.s. As a consequence,

$$\begin{aligned}
&P\left(\left| \sup_{p\in\hat{B}_n\setminus B(\phi)} |\hat{\phi}(p)| - \sup_{p\in\hat{B}_n\setminus B(\phi)} |\phi(p)| \right| \leq \frac{\delta_\varepsilon}{2} \right) \\
=&P\left(\sup_{p\in\hat{B}_n\setminus B(\phi)} |\phi(p)| - \frac{\delta_\varepsilon}{2} \leq \sup_{p\in\hat{B}_n\setminus B(\phi)} |\hat{\phi}(p)| \leq \sup_{p\in\hat{B}_n\setminus B(\phi)} |\phi(p)| + \frac{\delta_\varepsilon}{2} \right) \\
=&P(A_n) \to 1.
\end{aligned} \tag{B.18}$$

We used another fact in (B.16) that for any events A_n s.t. $\lim_{n\to\infty} P(A_n) \to 1$ then for any event C_n, $\lim_{n\to\infty} P(A_n \cap C_n) = \lim_{n\to\infty} P(C_n)$.

■

Lemma B.1.1 *For any $h \in \ell^\infty[0,1]$, $\hat{\mathscr{S}}'_n(h) \leq \mathscr{S}(h)$ almost surely.*

Proof. By definition

$$\hat{\mathscr{S}}'_n(h) = \sup_{p\in\hat{B}_n} h(p), \tag{B.19}$$

where $\hat{B}_n = \left\{p \in [0,1] : |\hat{\phi}(p)| \leq \tau_n\right\}$, and

$$\mathscr{S}(h) = \sup_{p\in[0,1]} h(p). \tag{B.20}$$

Clearly, $\hat{B}_n \subset [0,1]$, which implies $\hat{\mathscr{S}}'_n(h) \leq \mathscr{S}(h)$. ■

Proof of Lemma 2.2.5.

For any τ_n, let $\hat{B}_{0n} = \{p \in [0,1] : |\hat{\phi}(p)| \leq \tau_n\}$ and $\hat{B}_{+n} = \{p \in [0,1] : \hat{\phi}(p) > \tau_n\}$. Then we have

$$B_0(\phi) \triangle \hat{B}_{0n} = (B_0(\phi) \backslash \hat{B}_{0n}) \cup (\hat{B}_{0n} \backslash B_0(\phi)), \tag{B.21}$$

and

$$B_+(\phi) \triangle \hat{B}_{+n} = (B_+(\phi) \backslash \hat{B}_{+n}) \cup (\hat{B}_{+n} \backslash B_+(\phi)). \tag{B.22}$$

Let μ denote a Lebesgue measure. First consider $B_0(\phi) \triangle \hat{B}_{0n}$. For any $\varepsilon > 0$,

$$\begin{aligned} &P(\mu(B_0(\phi) \triangle \hat{B}_{0n}) > \varepsilon) \\ &\leq P(\mu(B_0(\phi) \backslash \hat{B}_{0n}) > \varepsilon/2) + P(\mu(\hat{B}_{0n} \backslash B_0(\phi)) > \varepsilon/2). \end{aligned} \tag{B.23}$$

Remember we have $\sup_{p\in[0,1]} |\hat{\phi}(p) - \phi(p)| \to 0$ a.s. i.e. $P(A_\phi) = 1$,
where $A_\phi = \{\omega : \sup_{p\in[0,1]} |\hat{\phi}_\omega(p) - \phi(p)| \to 0\}$. Now we fix any $\omega \in A_\phi$. For all ε, $\exists N > 0$, s.t. for any $n > N$, $\sup_{p\in[0,1]} |\hat{\phi}_\omega(p) - \phi(p)| < \varepsilon$. Then we can find $\varepsilon_n(\omega)$ s.t. $\sup_{p\in[0,1]} |\hat{\phi}_\omega(p) - \phi(p)| < \varepsilon_n(\omega)$ for each n and $\varepsilon_n(\omega) \to 0$. Thus

$$\hat{B}_{0n}(\omega) \subset \{p \in [0,1] : |\phi(p)| \leq \tau_n + \varepsilon_n(\omega)\}. \tag{B.24}$$

and

$$\hat{B}_{0n}(\omega) \backslash B_0(\phi) \subset \{p \in [0,1] : 0 < |\phi(p)| \leq \tau_n + \varepsilon_n(\omega)\}. \tag{B.25}$$

So we have

$$\mu(\hat{B}_{0n}(\omega) \backslash B_0(\phi)) \to 0, \tag{B.26}$$

for any $\omega \in A_\phi$, which implies $\mu(\hat{B}_{0n} \backslash B_0(\phi)) \to 0$ a.s. Then

$$1\{\mu(\hat{B}_{0n} \backslash B_0(\phi)) > \varepsilon/2\} \to 0 \text{ a.s..} \tag{B.27}$$

By Dominated Convergence Theorem,

$$P(\mu(\hat{B}_{0n} \backslash B_0(\phi)) > \varepsilon/2) = \int 1\{\mu(\hat{B}_{0n} \backslash B_0(\phi)) > \varepsilon/2\} \mathrm{d}P \to 0. \tag{B.28}$$

By arguments similar to (B.11),

$$P(\mu(B_0(\phi) \backslash \hat{B}_{0n}) > \varepsilon/2) \leq P(B_0(\phi) \backslash \hat{B}_{0n} \neq \varnothing) \to 0. \tag{B.29}$$

Hence

$$P(\mu(B_0(\phi) \triangle \hat{B}_{0n}) > \varepsilon) \to 0, \tag{B.30}$$

for any $\varepsilon > 0$, which implies $\mu(B_0(\phi) \triangle \hat{B}_{0n}) = o_p(1)$.

Next consider $B_+(\phi) \triangle \hat{B}_{+n}$. Fix any $\omega \in A_\phi$,

$$B_+(\phi) \backslash \hat{B}_{+n}(\omega) \subset \{p \in [0,1] : 0 < \phi(p) \leq \tau_n + \varepsilon_n(\omega)\}, \tag{B.31}$$

by similar arguments,

$$P(\mu(B_+(\phi) \backslash \hat{B}_{+n}) > \varepsilon/2) \to 0. \tag{B.32}$$

and

$$
\begin{aligned}
P(\mu(\hat{B}_{+n} \backslash B_+(\phi)) > \varepsilon/2) &\leq P(\hat{B}_{+n} \backslash B_+(\phi) \neq \varnothing) \\
&\leq P\left(\sup_{p \in \hat{B}_{+n} \backslash B_+(\phi)} |\hat{\phi}(p) - \phi(p)| > \tau_n \right) \\
&\leq P\left(\sup_{p \in [0,1]} \sqrt{T_n} |\hat{\phi}(p) - \phi(p)| > \sqrt{T_n}\tau_n \right) \to 0, \qquad \text{(B.33)}
\end{aligned}
$$

as shown in (B.11). ■

Lemma B.1.2 *For any $h \in \ell^\infty[0,1]$, $\hat{\mathscr{I}}'_n(h) \leq \mathscr{I}(h)$.*

Proof. By definition,

$$
\mathscr{I}(h) = \int \max(h(p), 0) 1\{0 \leq p \leq 1\} \, \mathrm{d}p. \qquad \text{(B.34)}
$$

and

$$
\begin{aligned}
\hat{\mathscr{I}}'_n(h) &= \int_{\hat{B}_{+n}} h(p) 1\{0 \leq p \leq 1\} \, \mathrm{d}p + \int_{\hat{B}_{0n}} \max(h(p), 0) 1\{0 \leq p \leq 1\} \, \mathrm{d}p \\
&\leq \int \max(h(p), 0) 1\{0 \leq p \leq 1\} \, \mathrm{d}p = \mathscr{I}(h). \qquad \text{(B.35)}
\end{aligned}
$$

■

Proof of Theorem 2.2.1.

We first prove $\hat{c}_{1-\alpha} \to_p c_{1-\alpha}$ by verifying the assumptions of Corollary 3.2 in Fang & Santos (2014).

In the setting of the test, $\mathbb{D} = \ell^\infty[0,1]$, and $\mathbb{E} = \mathbb{R}$. $\ell^\infty[0,1]$ is a Banach space under norm $\|\cdot\|_\infty$ and $\mathbb{R}$ is a Banach space under $|\cdot|$. By Assumption 2.2.4, $\mathscr{F}$ is Hadamard directionally differentiable at ϕ. So Assumption 2.1 in Fang & Santos (2014) holds.

By Assumptions 2.2.1, 3.3.1 and 2.2.4 and (2.12), Assumptions 2.2 and 2.3 in Fang & Santos (2014) hold.

Assumptions 3.1 and 3.2 in Fang & Santos (2014) are satisfied automatically by Lemma 2.2.3.

By Assumption 2.2.5, Assumption 3.3 in Fang & Santos (2014) holds. Then by Corollary 3.2 in Fang & Santos (2014), if the CDF of $\mathscr{F}'_\phi(\sqrt{\lambda}\mathscr{L}_2 - \sqrt{1-\lambda}\mathscr{L}_1)$ is strictly increasing at its $1-\alpha$ quantile $c_{1-\alpha}$, then $\hat{c}_{1-\alpha} \to_p c_{1-\alpha}$.

Under decision rule (2.33), if H_0 is true,

$$\sqrt{T_n}\mathscr{F}(\hat{\phi}) = \sqrt{T_n}(\mathscr{F}(\hat{\phi}) - \mathscr{F}(\phi)) \rightsquigarrow \mathscr{F}'_\phi(\sqrt{\lambda}\mathscr{L}_2 - \sqrt{1-\lambda}\mathscr{L}_1). \tag{B.36}$$

Thus $\sqrt{T_n}\mathscr{F}(\hat{\phi}) - \hat{c}_{1-\alpha} \rightsquigarrow \mathscr{F}'_\phi(\sqrt{\lambda}\mathscr{L}_2 - \sqrt{1-\lambda}\mathscr{L}_1) - c_{1-\alpha}$ by Slutsky's Theorem. Since the CDF of $\mathscr{F}'_\phi(\sqrt{\lambda}\mathscr{L}_2 - \sqrt{1-\lambda}\mathscr{L}_1)$ is continuous at $c_{1-\alpha}$ by assumption, the CDF of $\mathscr{F}'_\phi(\sqrt{\lambda}\mathscr{L}_2 - \sqrt{1-\lambda}\mathscr{L}_1) - c_{1-\alpha}$ is continuous at 0. Then we have

$$\begin{aligned} &P(\sqrt{T_n}\mathscr{F}(\hat{\phi}) - \hat{c}_{1-\alpha} > 0) = 1 - G_n(0) \\ &\to 1 - G(0) = P(\mathscr{F}'_\phi(\sqrt{\lambda}\mathscr{L}_2 - \sqrt{1-\lambda}\mathscr{L}_1) - c_{1-\alpha} > 0), \end{aligned} \tag{B.37}$$

as $n \to \infty$, where G_n and G are the distribution functions for $\sqrt{T_n}\mathscr{F}(\hat{\phi}) - \hat{c}_{1-\alpha}$ and $\mathscr{F}'_\phi(\sqrt{\lambda}\mathscr{L}_2 - \sqrt{1-\lambda}\mathscr{L}_1) - c_{1-\alpha}$ respectively.

If H_0 is false, by Assumption 2.2.2 $\mathscr{F}(\phi) > 0$.

Since $\sqrt{T_n}(\mathscr{F}(\hat{\phi}) - \mathscr{F}(\phi)) \rightsquigarrow \mathscr{F}'_\phi(\sqrt{\lambda}\mathscr{L}_2 - \sqrt{1-\lambda}\mathscr{L}_1)$ still holds, then

$$\begin{aligned} &P(\sqrt{T_n}\mathscr{F}(\hat{\phi}) - \hat{c}_{1-\alpha} > 0) \\ =&P(\sqrt{T_n}(\mathscr{F}(\hat{\phi}) - \mathscr{F}(\phi)) - \hat{c}_{1-\alpha} + \sqrt{T_n}\mathscr{F}(\phi) > 0) \to 1. \end{aligned} \tag{B.38}$$

■

Appendix C

Proofs for Chapter 3

C.1 Oracle Consistency

We first show general results for the oracle properties of penalized regressions and later we will use these results to prove the oracle properties of SFGMM estimator.

Suppose a general form of criterion function is

$$\hat{Q}_S(\alpha_S) = \tilde{Q}_S(\alpha_S) + \sum_{j=1}^{p} P_n\left(|\theta_j|\right) + \lambda_{hn} P_h(h), \tag{C.1}$$

where $\tilde{Q}$ is a loss function, P_n is a penalty function which satisfies Assumption 3.3.4 and P_h is a penalty function for h with tunning parameter λ_{hn}. This criterion function is a more general form than the one used in the main text. We can let $\lambda_{hn} = 0$ and obtain the criterion function used in the main text.

Lemma C.1.1 *Suppose the following conditions hold:*

(i) Almost surely, the loss function $\tilde{Q}_S(\alpha_S)$ has first and second order pathwise derivatives and by functional Taylor expansion in $\mathscr{A}^{(s)}$,

$$\begin{aligned}\tilde{Q}_S(\alpha_S) = \tilde{Q}_S(\Pi_n \alpha_{0S}) &+ \frac{\partial \tilde{Q}_S(\Pi_n \alpha_{0S})}{\partial \alpha_S}\left[\alpha_S - \Pi_n \alpha_{0S}\right] \\ &+ \frac{1}{2!}\frac{\partial^2 \tilde{Q}_S(\Pi_n \alpha_{0S} + \varsigma(\alpha_S - \Pi_n \alpha_{0S}))}{\partial \alpha_S^2}\left[\alpha_S - \Pi_n \alpha_{0S}\right]^2,\end{aligned} \tag{C.2}$$

where $\varsigma \in (0,1)$*;*

(ii) $\max_{m\leq 2}\sup_t \sqrt{\sum_{j=1}^{k}\left(\frac{\partial^m \varphi_j(t)}{\partial t^m}\right)^2} = \delta_{n\varphi} < \infty$ *for each n;*

(iii) $\left|\frac{\partial \tilde{Q}_S(\Pi_n \alpha_{0S})}{\partial \alpha_S}\left[\alpha_S - \Pi_n \alpha_{0S}\right]\right| = O_p(a_n)\left\|\alpha_S - \Pi_n \alpha_{0S}\right\|_s$*, where* $a_n = o(d_n)$*;*

(iv) $\exists C, \gamma_h > 0$ *such that for each* $\gamma_S \in \partial \mathscr{N}_\tau^{(n)}$*,* $h = \sum_{j=1}^{k} b_j \varphi_j$*,*

$$\left|P_h(h) - P_h(\Pi_n h_0)\right| \leq C\delta_{n\varphi}^{\gamma_h}\|b - b_0\|_E, \tag{C.3}$$

with $\lambda_{hn}\delta_{n\varphi}^{\gamma_h} = o(1)$*;*

(v) Let $e_{1n} = a_n + \sqrt{s}P_n'(d_n)$ *and* $e_{2n} = \delta_{n\varphi}a_n + \lambda_{hn}\delta_{n\varphi}^{\gamma_h} = o(1)$*, where* a_n *satisfies (iii). For each* $\varepsilon > 0$*,* $\exists C_2(n) > 0$*, such that with probability* $1-\varepsilon$*, for all large n with* $\|\theta_S - \theta_{0S}\|_E = O(e_{1n}/C_2(n)) = o(1)$*,* $\|b - b_0\|_E = O(e_{2n}/C_2(n)) = o(1)$ *and* $\alpha_S = \left(\theta_S, \sum_j^k b_j\varphi_j\right)$*,*

$$\inf_{\varsigma\in[0,1]} \frac{1}{2}\frac{\partial^2 \tilde{Q}_S\left(\Pi_n\alpha_{0S} + \varsigma\left(\alpha_S - \Pi_n\alpha_{0S}\right)\right)}{\partial \alpha_S^2}\left[\alpha_S - \Pi_n\alpha_{0S}\right]^2 \geq C_2(n)\left\|\gamma_S - \gamma_{0S}\right\|_E^2, \tag{C.4}$$

where $\gamma_S = \left(\theta_S', b'\right)'$ *and* $\gamma_{0S} = \left(\theta_{0S}', b_0'\right)'$*. Also, it holds that* $e_{1n}/\min\{1, C_2(n)\} = o(d_n)$*.*

Then for each $\varepsilon > 0$*, with probability* $1-\varepsilon$*, there is a local minimizer* $\hat{\alpha}_S$ *of* $\hat{Q}_S(\alpha_S)$ *on the sieve space* $\mathscr{A}_k^{(s)}$ *such that* $\left\|\hat{\alpha}_S - \Pi_n\alpha_{0S}\right\|_s = O\left(C_2(n)^{-1}\left(e_{1n} + \delta_{n\varphi}e_{2n}\right)\right)$*.*

In Lemma C.1.1, condition (i) shows that the loss function can be written as a functional Taylor expansion. In this way, we can take advantage of the quadratic form to prove the existence of the local minimizer. Condition (ii) is a restriction for basis functions $\{\varphi_j\}_j$. Condition (iii) basically requires that the first order part of the Taylor expansion is $o(\|\alpha_S - \Pi\alpha_{0S}\|_s)$. Condition (iv) requires the penalty function on the nonparametric component to be continuous with respect to h around Π_n. Condition (v) requires the quadratic term in the Taylor expansion to be large enough, so the difference of the first order and second order terms of the Taylor expansion is positive. Because of the setup of the model, we allow $C_2(n)$ to be a $o(1)$.

Proof of Lemma C.1.1.

The proof closely follows that of Theorem B.1 in Fan & Liao (2014). We extend it to allow for the nonparametric component.

Define

$$\mathscr{N}_{\tau}^{(n)}=\left\{\gamma_S=\left(\theta_S',b'\right)':\left\|\theta_S-\theta_{0S}\right\|_E\leq\tau_1(n)e_{1n},\left\|b-b_0\right\|_E\leq\tau_2(n)e_{2n}\right\}$$

for some $\tau_1(n)$, $\tau_2(n)>0$ such that

$$\tau_1(n)=O\left(\frac{1}{C_2(n)}\right),\tau_2(n)=O\left(\frac{1}{C_2(n)}\right).$$

For all $\gamma_S\in\partial\mathscr{N}_{\tau}^{(n)}$ and $\alpha_S=(\theta_S,h)$ with $h=\sum_{j=1}^{k}b_j\varphi_j$, by condition (i) we have

$$\begin{aligned}\hat{Q}_S(\alpha_S)-\hat{Q}_S(\Pi_n\alpha_{0S})=&\frac{\partial\tilde{Q}_S(\Pi_n\alpha_{0S})}{\partial\alpha_S}[\alpha_S-\Pi_n\alpha_{0S}]\\&+\frac{1}{2!}\frac{\partial^2\tilde{Q}_S(\Pi_n\alpha_{0S}+\varsigma(\alpha_S-\Pi_n\alpha_{0S}))}{\partial\alpha_S^2}[\alpha_S-\Pi_n\alpha_{0S}]^2\\&+\sum_{j=1}^{s}\left[P_n\left(\left|\theta_{Sj}\right|\right)-P_n\left(\left|\theta_{0S,j}\right|\right)\right]+\lambda_{hn}P_h(h)-\lambda_{hn}P_h(\Pi_nh_0).\end{aligned}\tag{C.5}$$

For each $\varepsilon>0$, by condition (iii), $\exists C_1>0$, such that for large n, $P(B_{1n})>1-\varepsilon/2$, where

$$B_{1n}=\left\{\frac{\partial\tilde{Q}_S(\Pi_n\alpha_{0S})}{\partial\alpha_S}[\alpha_S-\Pi_n\alpha_{0S}]>-C_1\left\|\alpha_S-\Pi_n\alpha_{0S}\right\|_s a_n\right\},$$

By condition (v), $\exists C_2(n)>0$, such that for large n, $P(B_{2n})\geq1-\varepsilon/2$, where

$$B_{2n}=\left\{\frac{1}{2!}\frac{\partial^2\tilde{Q}_S(\Pi_n\alpha_{0S}+\varsigma(\alpha_S-\Pi_n\alpha_{0S}))}{\partial\alpha_S^2}[\alpha_S-\Pi_n\alpha_{0S}]^2>C_2(n)\left\|\gamma_S-\gamma_{0S}\right\|_E^2\right\},$$

$\gamma_S=\left(\theta_S',b'\right)'$ and $\varsigma\in(0,1)$.

Since $h - \Pi_n h_0 = \sum_{j=1}^{k} \left(b_j - b_{0j}\right) \varphi_j$, by condition (ii),

$$\begin{aligned}\|h - \Pi_n h_0\|_s &= \max_{m \leq 2} \sup_t \left| \sum_{j=1}^{k} \left(b_j - b_{0j}\right) \frac{\partial^m \varphi_j(t)}{\partial t^m} \right| \\ &\leq \left(\max_{m \leq 2} \sup_t \sqrt{\sum_{j=1}^{k} \left(\frac{\partial^m \varphi_j(t)}{\partial t^m} \right)^2} \right) \|b - b_0\|_E = \delta_{n\varphi} \|b - b_0\|_E . \end{aligned} \tag{C.6}$$

By Lemma B.1 in Fan & Liao (2014), with condition (v),

$$\left| \sum_{j=1}^{s} \left[P_n\left(\left|\theta_{Sj}\right|\right) - P_n\left(\left|\theta_{0S,j}\right|\right) \right] \right| \leq \sqrt{s} P_n'(d_n) \|\theta_S - \theta_{0S}\|_E . \tag{C.7}$$

By conditions (iii), (iv), and (v), for each $\gamma_S \in \partial \mathcal{N}_\tau^{(n)}$ with $h = \sum_{j=1}^{k} b_j \varphi_j$,

$$|P_h(h) - P_h(\Pi_n h_0)| \leq C_h \delta_{n\varphi}^{\gamma_h} \|b - b_0\|_E \tag{C.8}$$

for some $C_h > 0$.

With $e_{1n} = a_n + \sqrt{s} P_n'(d_n) = o(d_n)$ and $e_{2n} = \delta_{n\varphi} a_n + \lambda_{hn} \delta_{n\varphi}^{\gamma_h} = o(1)$, on the event $B_{1n} \cap B_{2n}$,

$$\begin{aligned} &\hat{Q}_S(\alpha_S) - \hat{Q}_S(\Pi_n \alpha_{0S}) \\ \geq& -C_1 \|\alpha_S - \Pi_n \alpha_{0S}\|_s a_n + C_2(n) \|\gamma_S - \gamma_{0S}\|_E^2 \\ &- \sqrt{s} P_n'(d_n) \|\theta_S - \theta_{0S}\|_E - C_h \lambda_{hn} \delta_{n\varphi}^{\gamma_h} \|b - b_0\|_E \\ =& -C_1 \|\theta_S - \theta_{0S}\|_E a_n + C_2(n) \|\theta_S - \theta_{0S}\|_E^2 - \sqrt{s} P_n'(d_n) \|\theta_S - \theta_{0S}\|_E \\ &- C_1 \|h - \Pi_n h_0\|_s a_n + C_2(n) \|b - b_0\|_E^2 - C_h \lambda_{hn} \delta_{n\varphi}^{\gamma_h} \|b - b_0\|_E \\ \geq& \tau_1(n) e_{1n} \left(-C_1 a_n + C_2(n) \tau_1(n) e_{1n} - \sqrt{s} P_n'(d_n) \right) \\ &- C_1 \delta_{n\varphi} \|b - b_0\|_E a_n + C_2(n) \|b - b_0\|_E^2 - C_h \lambda_{hn} \delta_{n\varphi}^{\gamma_h} \|b - b_0\|_E \\ \geq& \tau_1(n) e_{1n}^2 \left(-C_1 + C_2(n) \tau_1(n) - 1 \right) + \tau_2(n) e_{2n}^2 \left(-C_1 + \tau_2(n) C_2(n) - C_h \right) . \end{aligned} \tag{C.9}$$

Then if we take $\tau_1(n) = (C_1 + 1 + \varepsilon_1)/C_2(n)$ and $\tau_2(n) = (C_1 + C_h + \varepsilon_2)/C_2(n)$ for some $\varepsilon_1, \varepsilon_2 > 0$, $\hat{Q}_S(\alpha_S) - \hat{Q}_S(\Pi_n \alpha_{0S}) > 0$. Moreover, because $\hat{Q}_S(\alpha_S)$ is continuous on $\mathscr{N}_\tau^{(n)}$ and $\mathscr{N}_\tau^{(n)}$ is compact, there exists a local minimizer $\hat{\alpha}_S$ such that $\hat{Q}_S(\hat{\alpha}_S) \leq \hat{Q}_S(\alpha_S)$ for all α_S with $\gamma_S \in \mathscr{N}_\tau^{(n)}$. Also, $\|\hat{\alpha}_S - \Pi_n \alpha_{0S}\|_s \leq \left(\tau_1(n) e_{1n} + \tau_2(n) \delta_{n\varphi} e_{2n}\right)$. Let $\delta_{n\alpha} = \tau_1(n) e_{1n} + \tau_2(n) \delta_{n\varphi} e_{2n}$, then $\|\hat{\alpha}_S - \Pi_n \alpha_{0S}\|_s = O(\delta_{n\alpha})$. ■

Lemma C.1.2 *Suppose the assumptions in Lemma C.1.1 hold. Let T be a projection function such that for all $\alpha = \left((\theta_1, ..., \theta_p)', h\right)$,*

$$T\alpha = \left(\left(\bar{\theta}_1, ..., \bar{\theta}_p\right)', h\right), \bar{\theta}_j = \begin{cases} \theta_j & \text{if } j \in S \\ 0 & \text{if } j \notin S \end{cases},$$

where $S = \left\{j \leq p : \theta_{0j} \neq 0\right\}$. We also write

$$T\theta = \left(\left(\bar{\theta}_1, ..., \bar{\theta}_p\right)'\right), \bar{\theta}_j = \begin{cases} \theta_j & \text{if } j \in S \\ 0 & \text{if } j \notin S \end{cases}.$$

Suppose with the local minimizer $\hat{\alpha}_S$ in Lemma C.1.1, it holds that with probability approaching one there exists a neighborhood $B \subseteq \mathscr{A}_k^{(n)}$ of $(\hat{\alpha}_S, 0)$ such that for all $\alpha \in B$ with $\alpha = (\alpha_S, \theta_N)$ but $\theta_N \neq 0$,

$$\hat{Q}(T\alpha) - \hat{Q}(\alpha) < \sum_{j \notin S} P_n\left(\left|\theta_j\right|\right). \tag{C.10}$$

Then with probability approaching one, $\hat{\alpha} = (\hat{\alpha}_S, 0)$ is a local minimizer of

$$\hat{Q}(\alpha) = \tilde{Q}(\alpha) + \sum_{j=1}^{p} P_n\left(\left|\theta_j\right|\right) + \lambda_{hn} P_h\left(\hat{h}\right) \tag{C.11}$$

on $\mathscr{A}_k^{(n)}$, and $\|\hat{\alpha} - \alpha_0\|_s = O(\delta_{n\alpha})$, where $\delta_{n\alpha}$ is obtained in Lemma C.1.1.

The proof of Lemma C.1.2 is similar to that of Theorem B.2 in Fan & Liao (2014).

Lemma C.1.3 *Given $C > 0$, for all α_S such that $\|\alpha_S\|_s \leq C$, under assumptions 3.3.1-3.3.7,*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\partial \rho (Z_i, \alpha_S)}{\partial \alpha_S} [\Delta \alpha_S] V_{Si} \right] \right\|_E = O_p \left(\sqrt{2s^2 + sk} \right) \|\Delta \alpha_S\|_s , \tag{C.12}$$

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\partial^2 \rho_S (Z_i, \alpha_S)}{\partial \alpha_S^2} [\Delta \alpha_S, \Delta \bar{\alpha}_S] V_{Si} \right] \right\|_E = O_p \left(s\sqrt{2s+k} \right) \|\Delta \alpha_S\|_s \|\Delta \bar{\alpha}_S\|_s . \tag{C.13}$$

Proof of Lemma C.1.3.

By Assumption 3.3.5, the first order pathwise derivative of $\rho (Z_i, \alpha)$ exists, that is, for each $\alpha \in \mathscr{A}^{(n)}$,

$$\frac{\partial \rho (Z_i, \alpha)}{\partial \alpha} [\Delta \alpha] = \frac{\partial \rho (Z_i, \alpha + t\Delta \alpha)}{\partial t} |_{t=0}.$$

By simple calculation, for each α such that $\|\alpha\|_s \leq C$, we have

$$\begin{aligned} \frac{\partial \rho (Z_i, \alpha)}{\partial \alpha} [\Delta \alpha] &= \frac{\partial \rho \left(Y, X'(\theta + t\Delta \theta), (h + t\Delta h)(\delta (Y, X'(\theta + t\Delta \theta))) \right)}{\partial t} |_{t=0} \\ &= O(1) X' \Delta \theta + O(1) \Delta h \left(\delta \left(Y, X' \theta \right) \right), \end{aligned} \tag{C.14}$$

and

$$\begin{aligned} \frac{\partial^2 \rho (Z, \alpha)}{\partial \alpha^2} [\Delta \alpha, \Delta \bar{\alpha}] &= O(1) \left(X' \Delta \theta \right) \left(X' \Delta \bar{\theta} \right) + O(1) \left(X' \Delta \theta \right) \Delta \bar{h} \left(\delta \left(Y, X' \theta \right) \right) \\ &+ O(1) \Delta h \left(\delta \left(Y, X' \theta \right) \right) \left(X' \Delta \bar{\theta} \right) + O(1) \Delta h \left(\delta \left(Y, X' \theta \right) \right) \Delta \bar{h} \left(\delta \left(Y, X' \theta \right) \right) \\ &+ O(1) \Delta h' \left(\delta \left(Y, X' \theta \right) \right) X' \Delta \bar{\theta} + O(1) \left(X' \Delta \theta \right) \Delta \bar{h}' \left(\delta \left(Y, X' \theta \right) \right). \end{aligned} \tag{C.15}$$

If $\Delta \alpha = \Delta \bar{\alpha}$,

$$\begin{aligned} \frac{\partial^2 \rho (Z, \alpha)}{\partial \alpha^2} [\Delta \alpha]^2 &= O(1) \left(X' \Delta \theta \right)^2 + O(1) \left(X' \Delta \theta \right) \Delta h \left(\delta \left(Y, X' \theta \right) \right) \\ &+ O(1) \left(\Delta h \left(\delta \left(Y, X' \theta \right) \right) \right)^2 + O(1) \left(X' \Delta \theta \right) \Delta h' \left(\delta \left(Y, X' \theta \right) \right). \end{aligned} \tag{C.16}$$

$\|\alpha\|_s \leq C$ can easily be satisfied if the parameter space is bounded. Then under Assump-

tion 3.3.7, for each $\Delta\alpha_S$ at α_S such that $\|\alpha_S\|_s \leq C$, $\exists M > 0$ such that

$$\begin{aligned}
&\left\|E\left[\left|\left(O(1)X_S'\Delta\theta_S + O(1)\Delta h\left(\delta\left(Y, X_S'(\theta_S)\right)\right)\right)V_S\right|\right]\right\|_E \\
\leq & M\left\{\left\|E\left[\left|V_S X_S'\Delta\theta_S\right|\right]\right\|_E + \left\|E\left[\left|V_S\Delta h\left(\delta\left(Y, X_S'\theta_S\right)\right)\right|\right]\right\|_E\right\} \\
\leq & M\left\{\sqrt{s(2s+k)}\max_{l,m} E\left[\left|V_{Sl}X_{Sm}\right|\right] + \sqrt{2s+k}\sqrt{M_V}\right\}\|\Delta\alpha_S\|_s = O\left(\sqrt{2s^2+sk}\right)\|\Delta\alpha_S\|_s .
\end{aligned} \tag{C.17}$$

Thus,

$$\left\|E\left[\left|\frac{\partial\rho_S(Z,\alpha_S)}{\partial\alpha_S}\left[\Delta\alpha_S\right]V_S\right|\right]\right\|_E = O\left(\sqrt{2s^2+sk}\right)\|\Delta\alpha_S\|_s . \tag{C.18}$$

We know

$$\begin{aligned}
&\left(\frac{\partial\rho_S(Z,\alpha_S)}{\alpha_S}\left[\Delta\alpha_S\right]V_{Sl}\right)^2 \\
=&O(1)\left[\left(X_S'\Delta\theta_S\right)^2 + \Delta h^2\left(\delta\left(Y, X_S'\theta_S\right)\right) + 2X_S'\Delta\theta_S\Delta h\left(\delta\left(Y, X_S'\theta_S\right)\right)\right]V_{Sl}^2.
\end{aligned} \tag{C.19}$$

Then we have

$$E\left[V_{Sl}^2\left(X_S'\Delta\theta_S\right)^2\right] \leq \left\|E\left[V_{Sl}^2 X_S X_S'\right]\right\|_E \|\Delta\theta_S\|_E^2 = O(s)\|\Delta\alpha_S\|_s^2 ,$$

$$E\left[V_{Sl}^2\Delta h^2\left(\delta\left(Y, X_S'\theta_S\right)\right)\right] = O(1)\|\Delta\alpha_S\|_s^2 ,$$

and

$$E\left[\left|V_{Sl}^2 X_S'\Delta\theta_S\Delta h\left(\delta\left(Y, X_S'\theta_S\right)\right)\right|\right] \leq \left\|E\left[\left|V_{Sl}^2 X_S'\right|\right]\right\|_E \|\Delta\alpha_S\|_s^2 = O(s)\|\Delta\alpha_S\|_s^2 .$$

Thus,

$$E\left[\left(\frac{\partial\rho_S(Z,\alpha_S)}{\alpha_S}\left[\Delta\alpha_S\right]V_{Sl}\right)^2\right] = O(s)\|\Delta\alpha_S\|_s^2 .$$

Then for each $\varepsilon > 0$,

$$
\begin{aligned}
&P\left(\frac{\left|\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\rho_S(Z_i,\alpha_S)}{\alpha_S}[\Delta\alpha_S]V_{Si}\right\|_E - \left\|E\left[\frac{\partial\rho_S(Z_i,\alpha_S)}{\alpha_S}[\Delta\alpha_S]V_{Si}\right]\right\|_E\right|}{\sqrt{2s^2+sk}\,\|\Delta\alpha_S\|_s} > \varepsilon\right) \\
\leq & P\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\rho_S(Z_i,\alpha_S)}{\alpha_S}[\Delta\alpha_S]V_{Si} - E\left[\frac{\partial\rho_S(Z_i,\alpha_S)}{\alpha_S}[\Delta\alpha_S]V_{Si}\right]\right\|_E > \varepsilon\sqrt{2s^2+sk}\,\|\Delta\alpha_S\|_s\right) \\
\leq & \sum_{l}^{2s+k} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\rho_S(Z_i,\alpha_S)}{\alpha_S}[\Delta\alpha_S]V_{Sli} - E\left[\frac{\partial\rho_S(Z_i,\alpha_S)}{\alpha_S}[\Delta\alpha_S]V_{Sli}\right]\right|^2 > \frac{\varepsilon^2\left(2s^2+sk\right)\|\Delta\alpha_S\|_s^2}{2s+k}\right) \\
\leq & \sum_{l}^{2s+k}\frac{nE\left[\left(\frac{\partial\rho_S(Z_i,\alpha_S)}{\alpha_S}[\Delta\alpha_S]V_{Sli} - E\left[\frac{\partial\rho_S(Z_i,\alpha_S)}{\alpha_S}[\Delta\alpha_S]V_{Sli}\right]\right)^2\right]}{n^2\varepsilon^2 s\,\|\Delta\alpha_S\|_s^2} = O\left(\frac{2s+k}{n\varepsilon^2}\right)\to 0, \qquad \text{(C.20)}
\end{aligned}
$$

under Assumption 3.3.6(ii). Thus,

$$
\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\rho_S(Z_i,\alpha_S)}{\alpha_S}[\Delta\alpha_S]V_{Si}\right\|_E &= \left\|E\left[\frac{\partial\rho_S(Z_i,\alpha_S)}{\alpha_S}[\Delta\alpha_S]V_{Si}\right]\right\|_E + o_p\left(\sqrt{2s^2+sk}\,\|\Delta\alpha_S\|_s\right) \\
&= O_p\left(\sqrt{2s^2+sk}\,\|\Delta\alpha_S\|_s\right). \qquad \text{(C.21)}
\end{aligned}
$$

This equation shows the relationship between the norm of the first order pathwise derivative $(\partial\rho_S(Z_i,\alpha_S)/\partial\alpha_S)[\Delta\alpha_S]V_{Si}$ and the norm of the increment on the parameter which is $\|\Delta\alpha_S\|_s$.

Consider

$$
\left\|E\left[\left|O(1)V_S\left(X_S'\Delta\theta_S\right)^2\right|\right]\right\|_E \leq M\left\|E\left[\left|V_S X_S'\Delta\theta_S\left(X_S'\Delta\theta_S\right)\right|\right]\right\|_E \qquad \text{(C.22)}
$$

for some $M > 0$. By assumption 3.3.7(i),

$$
E\left[\left|V_{Sl}X_S'\Delta\theta_S\left(X_S'\Delta\theta_S\right)\right|\right] = \left|\Delta\theta_S'\right| E\left[\left|V_{Sl}X_S X_S'\right|\right]|\Delta\theta_S| \leq sM_{VXX}\,\|\Delta\theta_S\|_E^2 \qquad \text{(C.23)}
$$

and then

$$\left\|E\left[\left|O(1)V_S\left(X_S'\Delta\theta_S\right)^2\right|\right]\right\|_E \leq M\cdot s\sqrt{2s+k}M_{VXX}\left\|\Delta\theta_S\right\|_E^2 \leq M\cdot s\sqrt{2s+k}M_{VXX}\left\|\Delta\alpha_S\right\|_s^2$$
$$= O\left(s\sqrt{2s+k}\right)\left\|\Delta\alpha_S\right\|_s^2. \tag{C.24}$$

Similarly,

$$\left\|E\left[\left|O(1)\left(X_S'\Delta\theta_S\right)\left(X_S'\Delta\bar{\theta}_S\right)V_S\right|\right]\right\|_E = O\left(s\sqrt{2s+k}\right)\left\|\Delta\alpha_S\right\|_s\left\|\Delta\bar{\alpha}_S\right\|_s. \tag{C.25}$$

By calculation similar to (C.17), $\exists M>0$ such that

$$\left\|E\left[\left|O(1)V_{Si}X_{Si}'\Delta\theta_S\Delta h\left(\delta\left(Y,X_{Si}'\Delta\theta_S\right)\right)\right|\right]\right\|_E \leq M\left\|E\left[\left|V_{Si}\left(X_{Si}'\Delta\theta_S\right)\right|\right]\right\|_E\left\|\Delta h\right\|_s$$
$$= O\left(\sqrt{s(2s+k)}\right)\left\|\Delta\alpha_S\right\|_s^2. \tag{C.26}$$

By calculation similar to (C.21),

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\rho_S(Z_i,\alpha_S)}{\partial\alpha_S^2}\left[\Delta\alpha_S\right]^2V_{Si}\right\|_E = O_p\left(s\sqrt{2s+k}\right)\left\|\Delta\alpha_S\right\|_s^2 \tag{C.27}$$

and

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\rho_S(Z_i,\alpha_S)}{\partial\alpha_S^2}\left[\Delta\alpha_S,\Delta\bar{\alpha}_S\right]V_{Si}\right\|_E = O_p\left(s\sqrt{2s+k}\right)\left\|\Delta\alpha_S\right\|_s\left\|\Delta\bar{\alpha}_S\right\|_s \tag{C.28}$$

for each α_S such that $\left\|\alpha_S\right\|_s \leq C$. These two equations show the relationship between the norm of the second order pathwise derivative $\left(\partial^2\rho_S(Z_i,\alpha_S)/\partial\alpha_S^2\right)\left[\Delta\alpha_S,\Delta\bar{\alpha}_S\right]V_{Si}$ and the norm of the increment on the parameters, namely $\left\|\Delta\alpha_S\right\|_s$ and $\left\|\Delta\bar{\alpha}_S\right\|_s$. ■

Proof of Theorem 3.3.1.

We verify conditions for Lemma C.1.1 and Lemma C.1.2 to prove the consistency of the SFGMM estimator in Theorem 3.3.1.

For each $\alpha_S \in \mathscr{A}_k^{(s)}$, by Assumption 3.3.5,

$$\tilde{Q}_S(\alpha_S) - \tilde{Q}_S(\Pi_n \alpha_{0S}) = \frac{\partial \tilde{Q}_S(\Pi_n \alpha_{0S})}{\partial \alpha_S} [\alpha_S - \Pi_n \alpha_{0S}] + \frac{1}{2} \frac{\partial^2 \tilde{Q}_S(\bar{\alpha}_S)}{\partial \alpha_S^2} [\alpha_S - \Pi_n \alpha_{0S}]^2, \quad \text{(C.29)}$$

where $\bar{\alpha}_S = \Pi_n \alpha_{0S} + \varsigma(\alpha_S - \Pi_n \alpha_{0S})$ with $\varsigma \in (0,1)$. We write the difference between $\tilde{Q}_S(\alpha_S)$ and $\tilde{Q}_S(\Pi_n \alpha_{0S})$ in the Taylor expansion form. Next, we will show the first order and the second order of the Taylor expansion satisfy the conditions of Lemma C.1.1. For the first term in (C.29),

$$\begin{aligned}
&\left| \frac{\partial \tilde{Q}_S(\Pi_n \alpha_{0S})}{\partial \alpha_S} [\alpha_S - \Pi_n \alpha_{0S}] \right| \\
=&\left| \left[\frac{1}{n} \sum_{i=1}^n \rho_S(Z_i, \Pi_n \alpha_{0S}) V_{Si}' \right] J_S \left[\frac{1}{n} \sum_{i=1}^n \frac{\partial \rho_S(Z_i, \Pi_n \alpha_{0S})}{\partial \alpha_S} [\alpha_S - \Pi_n \alpha_{0S}] V_{Si} \right] \right| \\
\leq&\left| \left[\frac{1}{n} \sum_{i=1}^n \rho_S(Z_i, \alpha_{0S}) V_{Si}' \right] J_S \left[\frac{1}{n} \sum_{i=1}^n \frac{\partial \rho_S(Z_i, \Pi_n \alpha_{0S})}{\partial \alpha_S} [\alpha_S - \Pi_n \alpha_{0S}] V_{Si} \right] \right| \\
&+ \left| \left[\frac{1}{n} \sum_{i=1}^n \frac{\partial \rho_S(Z_i, \tilde{\alpha}_S)}{\partial \alpha_S} [\Pi_n \alpha_{0S} - \alpha_{0S}] V_{Si}' \right] J_S \left[\frac{1}{n} \sum_{i=1}^n \frac{\partial \rho_S(Z_i, \Pi_n \alpha_{0S})}{\partial \alpha_S} [\alpha_S - \Pi_n \alpha_{0S}] V_{Si} \right] \right|.
\end{aligned} \quad \text{(C.30)}$$

By Lemma C.1.3 and Theorem 4.1 in Fan & Liao (2014),

$$\begin{aligned}
&\left| \left[\frac{1}{n} \sum_{i=1}^n \rho_S(Z_i, \alpha_{0S}) V_{Si}' \right] J_S \left[\frac{1}{n} \sum_{i=1}^n \frac{\partial \rho_S(Z_i, \Pi_n \alpha_{0S})}{\partial \alpha_S} [\alpha_S - \Pi_n \alpha_{0S}] V_{Si} \right] \right| \\
\leq&\left\| \left[\frac{1}{n} \sum_{i=1}^n \rho_S(Z_i, \alpha_{0S}) V_{Si}' \right] \right\|_E \|J_S\|_E \left\| \left[\frac{1}{n} \sum_{i=1}^n \frac{\partial \rho_S(Z_i, \Pi_n \alpha_{0S})}{\partial \alpha_S} [\alpha_S - \Pi_n \alpha_{0S}] V_{Si} \right] \right\|_E \\
=&O_p\left(\sqrt{(2s^2 + sk)(2s+k) \log p / n} \right) \|\alpha_S - \Pi_n \alpha_{0S}\|_s
\end{aligned} \quad \text{(C.31)}$$

and

$$\begin{aligned}\left|\frac{\partial \tilde{Q}_S\left(\Pi_n \alpha_{0S}\right)}{\partial \alpha_S}\left[\alpha_S-\Pi_n \alpha_{0S}\right]\right| =& O_p\left(\sqrt{\left(2s^2+sk\right)(2s+k)\log p/n}\right)\left\|\alpha_S-\Pi_n \alpha_{0S}\right\|_s \\ &+O_p\left(2s^2+sk\right)\left\|\alpha_{0S}-\Pi_n \alpha_{0S}\right\|_s\left\|\alpha_S-\Pi_n \alpha_{0S}\right\|_s \\ =& O_p\left(a_n\right)\left\|\alpha_S-\Pi_n \alpha_{0S}\right\|_s, \end{aligned} \tag{C.32}$$

where $a_n = \max\left\{\sqrt{\left(2s^2+sk\right)(2s+k)\log p/n}, \left(2s^2+sk\right)\left\|\alpha_{0S}-\Pi\alpha_{0S}\right\|_s\right\}$. Under Assumption 3.3.7(iii), $a_n = O\left(\sqrt{\left(2s^2+sk\right)(2s+k)\log p/n}\right)$, we let

$$e_{1n} = \sqrt{\left(2s^2+sk\right)(2s+k)\log p/n} + \sqrt{s}P_n'\left(d_n\right) \tag{C.33}$$

and

$$e_{2n} = \delta_{n\varphi}\sqrt{\left(2s^2+sk\right)(2s+k)\log p/n} + \lambda_{hn}\delta_{n\varphi}^{\gamma_h}, \tag{C.34}$$

where e_{1n}, e_{2n} are the notations used in Lemma C.1.1.

Consider

$$\begin{aligned}&\frac{\partial^2 \tilde{Q}_S\left(\bar{\alpha}_S\right)}{\partial \alpha_S^2}\left[\alpha_S-\Pi_n \alpha_{0S}\right]^2 \\ =&\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \rho_S\left(Z_i,\bar{\alpha}_S\right)}{\partial \alpha_S}\left[\alpha_S-\Pi_n \alpha_{0S}\right]V_{Si}'\right] J_S\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \rho_S\left(Z_i,\bar{\alpha}_S\right)}{\partial \alpha_S}\left[\alpha_S-\Pi_n \alpha_{0S}\right]V_{Si}\right] \\ &+\left[\frac{1}{n}\sum_{i=1}^n \rho_S\left(Z_i,\bar{\alpha}_S\right)V_{Si}'\right] J_S\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \rho_S\left(Z_i,\bar{\alpha}_S\right)}{\partial \alpha_S^2}\left[\alpha_S-\Pi_n \alpha_{0S}\right]^2 V_{Si}\right]. \end{aligned} \tag{C.35}$$

First we have

$$
\begin{aligned}
&\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\rho_S(Z_i,\bar{\alpha}_S)}{\partial\alpha_S}[\alpha_S-\Pi_n\alpha_{0S}]V_{Si}'\right]J_S\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\rho_S(Z_i,\bar{\alpha}_S)}{\partial\alpha_S}[\alpha_S-\Pi_n\alpha_{0S}]V_{Si}\right]\\
=&\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\rho_S(Z_i,\Pi_n\alpha_{0S})}{\partial\alpha_S}[\alpha_S-\Pi_n\alpha_{0S}]V_{Si}'\right]J_S\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\rho_S(Z_i,\bar{\alpha}_S)}{\partial\alpha_S}[\alpha_S-\Pi_n\alpha_{0S}]V_{Si}\right]\\
&+\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\rho_S(Z_i,\tilde{\alpha}_S)}{\partial\alpha_S^2}[\alpha_S-\Pi_n\alpha_{0S}][\bar{\alpha}_S-\Pi_n\alpha_{0S}]V_{Si}'\right]\\
&\cdot J_S\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\rho_S(Z_i,\bar{\alpha}_S)}{\partial\alpha_S}[\alpha_S-\Pi_n\alpha_{0S}]V_{Si}\right]\\
=&\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\rho_S(Z_i,\Pi_n\alpha_{0S})}{\partial\alpha_S}[\alpha_S-\Pi_n\alpha_{0S}]V_{Si}'\right]J_S\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\rho_S(Z_i,\Pi_n\alpha_{0S})}{\partial\alpha_S}[\alpha_S-\Pi_n\alpha_{0S}]V_{Si}\right]\\
&+\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\rho_S(Z_i,\tilde{\alpha}_{S1})}{\partial\alpha_S^2}[\alpha_S-\Pi_n\alpha_{0S}][\bar{\alpha}_S-\Pi_n\alpha_{0S}]V_{Si}'\right]\\
&\cdot J_S\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\rho_S(Z_i,\Pi_n\alpha_{0S})}{\partial\alpha_S}[\alpha_S-\Pi_n\alpha_{0S}]V_{Si}\right]\\
&+\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\rho_S(Z_i,\tilde{\alpha}_S)}{\partial\alpha_S^2}[\alpha_S-\Pi_n\alpha_{0S}][\bar{\alpha}_S-\Pi_n\alpha_{0S}]V_{Si}'\right]\\
&\cdot J_S\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\rho_S(Z_i,\bar{\alpha}_S)}{\partial\alpha_S}[\alpha_S-\Pi_n\alpha_{0S}]V_{Si}\right].
\end{aligned}
$$

Given that $\alpha_S\in\mathscr{A}_k^{(s)}$ with $\alpha_S=(\theta_S,h)$ and $h=\sum_{j=1}^{k}b_j\varphi_j$, we write

$$
\begin{aligned}
\rho_S(Z,\alpha_S)&=\rho_S\left(Y,X_S'\theta_S,h\left(\delta\left(Y,X_S'\theta_S\right)\right)\right)\\
&=\rho_S\left(Y,X_S'\theta_S,\sum_{j=1}^{k}b_j\varphi_j\left(\delta\left(Y,X_S'\theta_S\right)\right)\right)=\bar{\rho}_S(Z,\gamma_S).
\end{aligned}
\tag{C.36}
$$

Here we use the map $\rho_S=\rho_S(t_1,t_2,t_3)$. And later we will write $\rho_{Sj}'=\rho_{St_j}'(t_1,t_2,t_3)$ for $j=1,2,3$. Let $L_S(\gamma_S)=\tilde{Q}_S(\alpha_S)$, where $\gamma_S=\left(\theta_S',b'\right)'$.

Now consider

$$\begin{aligned}&\frac{\partial \rho_S(Z_i,\Pi_n\alpha_{0S})}{\partial \alpha_S}[\alpha_S-\Pi_n\alpha_{0S}]=\frac{\partial \rho_S(Z_i,\Pi_n\alpha_{0S}+t(\alpha_S-\Pi_n\alpha_{0S}))}{\partial t}|_{t=0}\\ =&\frac{\partial \bar{\rho}_S(Z_i,\gamma_{0S}+t(\gamma_S-\gamma_{0S}))}{\partial t}|_{t=0}=(\gamma_S-\gamma_{0S})'\frac{\partial \bar{\rho}_S(Z_i,\gamma_{0S})}{\partial \gamma_S}.\end{aligned} \tag{C.37}$$

By simple algebra,

$$\frac{\partial \bar{\rho}_S(Z_i,\gamma_{0S})}{\partial \gamma_S}=\frac{\partial \rho_S\left(Y_i,X_{Si}'\theta_S,\sum_{j=1}^{k(n)}b_j\varphi_j\left(\delta\left(Y_i,X_{Si}'\theta_S\right)\right)\right)}{\partial \gamma_S}|_{\gamma_{0S}}=\begin{pmatrix}r_{1i}\\ r_{2i}\end{pmatrix}, \tag{C.38}$$

where

$$r_{1i}=\rho_{S2}'(\Pi_n\alpha_{0S})X_{Si}+\rho_{S3}'(\Pi_n\alpha_{0S})\sum_{j=1}^{k}b_{0j}\varphi_j'\left(\delta\left(Y_i,X_i'\theta_0\right)\right)\delta_2'\left(Y_i,X_i'\theta_0\right)X_{Si} \tag{C.39}$$

and

$$r_{2i}=\rho_{S3}'(\Pi_n\alpha_{0S})\left(\varphi_1\left(\delta\left(Y_i,X_i'\theta_0\right)\right),...,\varphi_k\left(\delta\left(Y_i,X_i'\theta_0\right)\right)\right)'. \tag{C.40}$$

Also,

$$\begin{aligned}&\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \rho_S(Z_i,\Pi_n\alpha_{0S})}{\partial \alpha_S}[\alpha_S-\Pi_n\alpha_{0S}]V_{Si}'\right]J_S\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \rho_S(Z_i,\Pi_n\alpha_{0S})}{\partial \alpha_S}[\alpha_S-\Pi_n\alpha_{0S}]V_{Si}\right]\\ =&\left[\frac{1}{n}\sum_{i=1}^{n}(\gamma_S-\gamma_{0S})'\frac{\partial \bar{\rho}_S(Z_i,\gamma_{0S})}{\partial \gamma_S}V_{Si}'\right]J_S\left[\frac{1}{n}\sum_{i=1}^{n}(\gamma_S-\gamma_{0S})'\frac{\partial \bar{\rho}_S(Z_i,\gamma_{0S})}{\partial \gamma_S}V_{Si}\right],\end{aligned} \tag{C.41}$$

where

$$\frac{\partial \rho_S(Z_i,\Pi_n\alpha_{0S})}{\partial \gamma_S}V_{Si}'=\begin{pmatrix}r_{1i}V_{Si}'\\ r_{2i}V_{Si}'\end{pmatrix}. \tag{C.42}$$

If

$$\lambda_{\max}\left\{E\left[\frac{\partial\rho_S(Z_i,\Pi_n\alpha_{0S})}{\partial\gamma_S}V_{Si}'\right]E\left[\frac{\partial\rho_S(Z_i,\Pi_n\alpha_{0S})}{\partial\gamma_S}V_{Si}'\right]'\right\}>C_\lambda(n)$$

for some $C_\lambda(n)>0$, then under Assumption 3.3.6(ii), for each $\varepsilon>0$, with probability $1-\varepsilon$, for all large n,

$$\begin{aligned}&\left\|\left[\frac{1}{n}\sum_{i=1}^n(\gamma_S-\gamma_{0S})'\frac{\partial\bar\rho_S(Z_i,\gamma_{0S})}{\partial\gamma_S}V_{Si}'\right]J_S\left[\frac{1}{n}\sum_{i=1}^n(\gamma_S-\gamma_{0S})'\frac{\partial\bar\rho_S(Z_i,\gamma_{0S})}{\partial\gamma_S}V_{Si}'\right]'\right\|_E\\ \geq&\bar C_\lambda(n)\|\gamma_S-\gamma_{0S}\|_E^2\end{aligned}\tag{C.43}$$

for some $\bar C_\lambda(n)>0$. By algebra, we can show that $\bar C_\lambda(n)\geq C_{JL}C_\lambda(n)$. This is a condition similar to the Assumption 4.5 in Fan & Liao (2014). As $n\to\infty$, the basis functions will change, so we allow the lower bound $C_\lambda(n)$ to change as n increases.

With e_{1n}, e_{2n} in (C.33) and (C.34), consider all α_S in the sieve space such that

$$\|\theta_S-\theta_{0S}\|_E=O\left(e_{1n}/\bar C_\lambda(n)\right)$$

and

$$\|b-b_0\|_E=O\left(e_{2n}/\bar C_\lambda(n)\right).$$

By Lemma C.1.3 and Assumption 3.3.7, for $\bar\alpha_S=\Pi_n\alpha_{0S}+\varsigma(\alpha_S-\Pi_n\alpha_{0S})$ and $\Delta\alpha_S=$

$\alpha_S - \Pi_n \alpha_{0S}$,

$$
\begin{aligned}
&\sup_{\varsigma\in[0,1]} \left| \left(\frac{1}{n}\sum_{i=1}^{n} \rho_S\left(Z_i, \bar{\alpha}_S\right) V_{Si} \right)' J_S \left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \rho_S\left(Z_i, \bar{\alpha}_S\right)}{\partial \alpha_S^2} \left[\Delta\alpha_S\right]^2 V_{Si} \right) \right| \\
\leq & \left\| \left(\frac{1}{n}\sum_{i=1}^{n} \rho_S\left(Z_i, \alpha_{0S}\right) V_{Si} \right)' \right\|_E \left\| J_S \right\|_E \left\| \left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \rho_S\left(Z_i, \bar{\alpha}_S\right)}{\partial \alpha_S^2} \left[\Delta\alpha_S\right]^2 V_{Si} \right) \right\|_E \\
&+ \sup_{\varsigma\in[0,1]} \left\| \left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \rho_S\left(Z_i, \tilde{\alpha}_S\right)}{\partial \alpha_S} \left[\bar{\alpha}_S - \alpha_{0S}\right] V_{Si} \right)' \right\|_E \left\| J_S \right\|_E \\
&\cdot \left\| \left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \rho_S\left(Z_i, \bar{\alpha}_S\right)}{\partial \alpha_S^2} \left[\Delta\alpha_S\right]^2 V_{Si} \right) \right\|_E \\
= & O_p\left(s\left(2s+k\right)\sqrt{\log p/n} \right) \left\| \Delta\alpha_S \right\|_s^2 \\
&+ O_p\left(\sqrt{2s^2+sk} \right) \sup_{\varsigma\in[0,1]} \left\| \Pi_n \alpha_{0S} + \varsigma\Delta\alpha_S - \alpha_{0S} \right\|_s O_p\left(1\right) O_p\left(s\sqrt{2s+k} \right) \left\| \Delta\alpha_S \right\|_s^2 \\
= & o_p\left(\bar{C}_\lambda\left(n\right) \right) \left\| \gamma_S - \gamma_{0S} \right\|_E^2 . \qquad \text{(C.44)}
\end{aligned}
$$

Next consider for each bounded α_{S1}, α_{S2},

$$
\begin{aligned}
&\sup_{\varsigma\in[0,1]} \left| \begin{array}{c} \left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \rho_S\left(Z_i, \alpha_{S1}\right)}{\partial \alpha_S^2} \left[\alpha_S - \Pi_n\alpha_{0S}\right] \left[\bar{\alpha}_S - \Pi_n\alpha_{0S}\right] V_{Si}' \right] \\ J_S \left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \rho_S\left(Z_i, \alpha_{S2}\right)}{\partial \alpha_S} \left[\alpha_S - \Pi_n\alpha_{0S}\right] V_{Si} \right] \end{array} \right| \\
\leq & \sup_{\varsigma\in[0,1]} \left\| \left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \rho_S\left(Z_i, \alpha_{S1}\right)}{\partial \alpha_S^2} \left[\alpha_S - \Pi_n\alpha_{0S}\right] \left[\bar{\alpha}_S - \Pi_n\alpha_{0S}\right] V_{Si}' \right] \right\|_E \\
&\cdot \left\| J_S \right\|_E \left\| \left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \rho_S\left(Z_i, \alpha_{S2}\right)}{\partial \alpha_S} \left[\alpha_S - \Pi_n\alpha_{0S}\right] V_{Si} \right] \right\|_E \\
\leq & O_p\left(s\sqrt{2s+k} \right) \left\| \alpha_S - \Pi_n\alpha_{0S} \right\|_s \left\| \left(\alpha_S - \Pi_n\alpha_{0S}\right) \right\|_s O_p\left(\sqrt{2s^2+sk} \right) \left\| \alpha_S - \Pi_n\alpha_{0S} \right\|_s \\
= & O_p\left(s\sqrt{\left(2s+k\right)\left(2s^2+sk\right)}\delta_{n\alpha} \right) \left\| \alpha_S - \Pi_n\alpha_{0S} \right\|_s^2 = o_p\left(\bar{C}_\lambda\left(n\right) \right) \left\| \gamma_S - \gamma_{0S} \right\|_E^2 . \qquad \text{(C.45)}
\end{aligned}
$$

So (C.43)–(C.45) above imply that

$$\inf_{\varsigma\in[0,1]} \frac{1}{2}\frac{\partial^2 \tilde{Q}_S\left(\Pi_n\alpha_{0S}+\varsigma\left(\alpha_S-\Pi_n\alpha_{0S}\right)\right)}{\partial\alpha_S^2}\left[\alpha_S-\Pi_n\alpha_{0S}\right]^2 \geq \left(\bar{C}_\lambda(n)-o_p\left(\bar{C}_\lambda(n)\right)\right)\left\|\gamma_S-\gamma_{0S}\right\|_E^2. \tag{C.46}$$

We let $C_2(n)=\bar{C}_\lambda(n)/2$, then assumptions in Lemma C.1.1 hold.

By Lemma C.1.1, $\left\|\alpha_S-\alpha_{0S}\right\|_s = O\left(\delta_{n\alpha}\right)$, with

$$\delta_{n\alpha}=\delta_{n\varphi}^2\sqrt{\left(2s^2+sk\right)(2s+k)\log p/n}/C_\lambda(n).$$

Next, we check the other assumptions for Lemma C.1.2. We know

$$\begin{aligned}\tilde{Q}(T\alpha)-\tilde{Q}(\alpha) \leq &-\left(\sum_{l\notin S,\theta_l\neq 0}\theta_l\left[\frac{1}{n}\sum_i \frac{\partial\rho\left(Z_i,T\alpha+\varsigma\left(\alpha-T\alpha\right)\right)}{\partial\theta_l}V_i'(T\theta)\right]\right)\\ &\cdot J(T\theta)\cdot\left(\frac{1}{n}\sum_{i=1}^n\rho\left(Z_i,T\alpha+\varsigma\left(\alpha-T\alpha\right)\right)V_i(T\theta)\right).\end{aligned} \tag{C.47}$$

By calculation similar to that of Lemma C.1.3,

$$\left\|E\left[\left|O(1)X_{li}V_{Si}\right|\right]\right\|_E = O\left(\sqrt{2s+k}\right). \tag{C.48}$$

So for each bounded α,

$$\left\|E\left[\frac{\partial\rho\left(Z_i,T\alpha+\varsigma\left(\alpha-T\alpha\right)\right)}{\partial\theta_l}V_i'(T\theta)\right]\right\|_E = O\left(\sqrt{2s+k}\right). \tag{C.49}$$

For all $\varepsilon > 0$,

$$
\begin{aligned}
&P\left(\frac{\left|\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\rho_S(Z_i,T\alpha+\varsigma(\alpha-\alpha))}{\partial\theta_l}V_i(T\theta)\right\|_E-\left\|E\left[\frac{\partial\rho_S(Z_i,T\alpha+\varsigma(\alpha-\alpha))}{\partial\theta_l}V_i(T\theta)\right]\right\|_E\right|}{\sqrt{2s+k}}>\varepsilon\right)\\
\leq& P\left(\begin{array}{c}\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\rho_S(Z_i,T\alpha+\varsigma(\alpha-\alpha))}{\partial\theta_l}V_i(T\theta)-E\left[\frac{\partial\rho_S(Z_i,\alpha_S)}{\alpha_S}[\Delta\alpha_S]V_{Si}\right]\right\|_E\\ >\varepsilon\sqrt{2s+k}\end{array}\right)\\
\leq&\sum_{l}^{2s+k}P\left(\begin{array}{c}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\rho_S(Z_i,T\alpha+\varsigma(\alpha-\alpha))}{\partial\theta_l}V_l(T\theta)-E\left[\frac{\partial\rho_S(Z_i,T\alpha+\varsigma(\alpha-\alpha))}{\partial\theta_l}V_l(T\theta)\right]\right|^2\\ >\varepsilon^2\end{array}\right)\\
\leq&\sum_{l}^{2s+k}\frac{nE\left[\left(\frac{\partial\rho_S(Z_i,T\alpha+\varsigma(\alpha-\alpha))}{\partial\theta_l}V_l(T\theta)\right)^2\right]}{n^2\varepsilon^2}=O\left(\frac{1}{n\varepsilon^2}\right)\to 0
\end{aligned}
$$

under Assumption 3.3.7(i), which implies

$$
\left\|\frac{1}{n}\sum_{i}\frac{\partial\rho\left(Z_i,T\alpha+\varsigma\left(\alpha-T\alpha\right)\right)}{\partial\theta_l}V_i'(T\theta)\right\|_E=O_p\left(\sqrt{2s+k}\right). \tag{C.50}
$$

Notice that

$$
\frac{1}{n}\sum_{i=1}^{n}\rho\left(Z_i,T\alpha+\varsigma\left(\alpha-T\alpha\right)\right)V_i(T\theta)=\frac{1}{n}\sum_{i=1}^{n}\left[\rho\left(Z_i,\hat{\alpha}\right)+\frac{\partial\rho\left(Z_i,\tilde{\alpha}\right)}{\partial\alpha}[\Delta\alpha]\right]V_i(T\theta), \tag{C.51}
$$

where $\Delta\alpha=T\alpha+\varsigma\left(\alpha-T\alpha\right)-\hat{\alpha}$, $\hat{\alpha}=\left(\left(\hat{\theta}_S',0'\right)',\hat{h}\right)$ and remember $\hat{\alpha}_S=\left(\hat{\theta}_S,\hat{h}\right)$. We have

$$
\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^{n}\rho\left(Z_i,\hat{\alpha}\right)V_i(T\theta)\right\|_E\leq\left\|\frac{1}{n}\sum_{i=1}^{n}\rho\left(Z_i,\alpha_0\right)V_i(T\theta)\right\|_E\\
&+\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\rho_S\left(Z_i,\bar{\alpha}_S\right)}{\partial\alpha_S}[\hat{\alpha}_S-\alpha_{0S}]V_i(T\theta)\right\|_E=O_p\left(\left(\sqrt{2s^2+sk}\right)\delta_{n\alpha}\right).
\end{aligned} \tag{C.52}
$$

Also,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \rho \left(Z_i, \tilde{\alpha} \right)}{\partial \alpha} \left[\Delta \alpha \right] V_i \left(T \theta \right) \right\|_E = O_p \left(\sqrt{p \left(2s+k \right)} \right) \left\| \Delta \alpha \right\|_s , \tag{C.53}$$

where $\Delta \alpha = \varsigma \left(\alpha - \hat{\alpha} \right) + \left(1 - \varsigma \right) \left(T \alpha - \hat{\alpha} \right)$ and $\varsigma \in (0,1)$.

We let

$$r_n = o \left(P_n' \left(0^+ \right) / \sqrt{p \left(2s+k \right)^2} \right).$$

We assume $\left\| \alpha - \hat{\alpha} \right\|_s \leq r_n$, then $\left\| T \alpha - \hat{\alpha} \right\|_s \leq r_n$ and $\left\| \Delta \alpha \right\|_s \leq r_n$. With Assumption 3.3.7(iv),

$$\left| \begin{array}{c} \left[\frac{1}{n} \sum_i \frac{\partial \rho \left(Z_i, T \alpha + \varsigma \left(\alpha - T \alpha \right) \right)}{\partial \theta_l} V_i' \left(T \theta \right) \right] J \left(T \theta \right) \\ \cdot \left(\frac{1}{n} \sum_{i=1}^{n} \rho \left(Z_i, T \alpha + \varsigma \left(\alpha - T \alpha \right) \right) V_i \left(T \theta \right) \right) \end{array} \right| = o_p \left(P_n' \left(0^+ \right) \right). \tag{C.54}$$

Let

$$\begin{aligned} D_l \left(\alpha \right) = & \left[\frac{1}{n} \sum_i \frac{\partial \rho \left(Z_i, T \alpha + \varsigma \left(\alpha - T \alpha \right) \right)}{\partial \theta_l} V_i' \left(T \theta \right) \right] J \left(T \theta \right) \\ & \cdot \left(\frac{1}{n} \sum_{i=1}^{n} \rho \left(Z_i, T \alpha + \varsigma \left(\alpha - T \alpha \right) \right) V_i \left(T \theta \right) \right), \end{aligned}$$

then

$$P \left(\left| D_l \right| \leq \frac{1}{2} P_n' \left(0^+ \right) \right) \to 1. \tag{C.55}$$

By mean value theorem and $P_n \left(0 \right) = 0$, we know $\exists \lambda \in (0,1)$ such that

$$\sum_{l \notin S} P_n \left(\left| \theta_l \right| \right) = \sum_{l \notin S, \theta_l \neq 0} \left| \theta_l \right| P_n' \left(\lambda \left| \theta_l \right| \right). \tag{C.56}$$

Also, we know $|\theta_l| \leq r_n$, then because P_n' is non-increasing,

$$P_n'(\lambda |\theta_l|) \geq P_n'(r_n). \tag{C.57}$$

Because P_n' is continuous, we can always find sufficiently small r_n such that $P_n'(r_n) \geq \frac{1}{2} P_n'(0^+)$. Then with probility approaching to 1,

$$\begin{aligned} \hat{Q}(T\alpha) - \hat{Q}(\alpha) &\leq \sum_{l \notin S} (-\theta_l) D_l(\alpha) \leq \sum_{l \notin S} |\theta_l| |D_l(\alpha)| \leq \sum_{l \notin S} |\theta_l| \frac{1}{2} P_n'(0^+) \\ &\leq \sum_{l \notin S} |\theta_l| P_n'(r_n) \leq \sum_{l \notin S, \theta_l \neq 0} |\theta_l| P_n'(\lambda |\theta_l|) \leq \sum_{l \notin S} P_n(|\theta_l|). \end{aligned} \tag{C.58}$$

■

C.2 Asymptotic Normality

Define $\alpha_S^* = \alpha_S \pm \varepsilon_n u_n^*$ for $\varepsilon_n = o\left(n^{-\frac{1}{2}}\right)$, and

$$\mathscr{E}\left(\left[\frac{1}{n}\sum_{i=1}^n A_i'\right] J \left[\frac{1}{n}\sum_{i=1}^n B_i\right]\right) = E\left[A_i'\right] J E\left[B_i\right].$$

Lemma C.2.1 *Local behavior of Criterion and Penalty: Under Assumptions 3.3.1, 3.3.4, 3.3.5 and 3.4.1, it holds that*

(i) $\sup_{\alpha_S \in \mathscr{B}_n^S(\tau)} |\tilde{Q}_S(\alpha_S^*) - \tilde{Q}_S(\alpha_S) - \mathscr{E}\left(\tilde{Q}_S(\alpha_S^*) - \tilde{Q}_S(\alpha_S)\right) - \Delta(Z, \alpha_{0S})[\Delta\alpha_S]| = O_p\left(\varepsilon_n^2\right)$, *where* $\Delta\alpha_S = \alpha_S^* - \alpha_S$;

(ii) $\sum_{j=1}^s \left[P_n\left(\left|\hat{\theta}_{Sj}^*\right|\right) - P_n\left(\left|\hat{\theta}_{Sj}\right|\right)\right] = O\left(\varepsilon_n^2\right)$;

(iii) $\left|\mathscr{E}\left(\hat{Q}_S(\alpha_S^*) - \hat{Q}_S(\alpha_S)\right) - \frac{\|\alpha_S \pm \varepsilon_n u_n^* - \alpha_{0S}\|^2 - \|\alpha - \alpha_{0S}\|^2}{2}\right| = O\left(\varepsilon_n^2\right)$.

Proof of Lemma C.2.1.

First, Lemma C.2.1(ii) holds by Assumptions 3.3.4, 3.3.7(ii), and Lemma B.1 in Fan & Liao (2014). Next, we prove Lemma C.2.1(i). For all $\alpha_S \in \mathscr{V}_n^S(\tau)$,

$$\begin{aligned}
&\tilde{Q}_S(\alpha_S^*) - \tilde{Q}_S(\alpha_S) - \Delta(\alpha_{0S})[\pm\varepsilon_n u_n^*] = \frac{\partial \tilde{Q}_S(\alpha_{S1})}{\partial \alpha_S}[\pm\varepsilon_n u_n^*] - \frac{\partial \tilde{Q}_S(\alpha_{0S})}{\partial \alpha_S}[\pm\varepsilon_n u_n^*] \\
&= \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \rho_S(Z_i,\alpha_{S2})}{\partial \alpha_S}[\alpha_{S1}-\alpha_{0S}]V_{Si}'\right] J_S \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \rho_S(Z_i,\alpha_{S2})}{\partial \alpha_S}[\pm\varepsilon_n u_n^*]V_{Si}\right] \\
&\quad + \left[\frac{1}{n}\sum_{i=1}^n \rho_S(Z_i,\alpha_{S2})V_{Si}'\right] J_S \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \rho_S(Z_i,\alpha_{S2})}{\partial \alpha_S^2}[\pm\varepsilon_n u_n^*, \alpha_{S1}-\alpha_{0S}]V_{Si}\right], \qquad \text{(C.59)}
\end{aligned}$$

where $\alpha_{S2} = \alpha_{0S} + \tau_2(\alpha_{S1} - \alpha_{0S})$ and $\alpha_{S1} = \alpha_S + \tau_1(\pm\varepsilon_n u_n^*)$ with $\tau_1, \tau_2 \in (0,1)$. This is from the pathwise differentiability Assumption 3.3.5. We now apply Lemma C.1.3 to (C.59).

For the first part of (C.59),

$$\begin{aligned}
&\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \rho_S(Z_i,\alpha_{S2})}{\partial \alpha_S}[\alpha_{S1}-\alpha_{0S}]V_{Si}'\right] J_S \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \rho_S(Z_i,\alpha_{S2})}{\partial \alpha_S}[\pm\varepsilon_n u_n^*]V_{Si}\right] \\
&= \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \rho_S(Z_i,\alpha_{0S})}{\partial \alpha_S}[\alpha_{S1}-\alpha_{0S}]V_{Si}'\right] J_S \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \rho_S(Z_i,\alpha_{0S})}{\partial \alpha_S}[\pm\varepsilon_n u_n^*]V_{Si}\right] \\
&\quad + \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \rho_S(Z_i,\alpha_{S3})}{\partial \alpha_S^2}[\alpha_{S1}-\alpha_{0S}, \alpha_{S2}-\alpha_{0S}]V_{Si}'\right] J_S \cdot \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \rho_S(Z_i,\alpha_{0S})}{\partial \alpha_S}[\pm\varepsilon_n u_n^*]V_{Si}\right] \\
&\quad + \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \rho_S(Z_i,\alpha_{0S})}{\partial \alpha_S}[\alpha_{S1}-\alpha_{0S}]V_{Si}'\right] J_S \cdot \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \rho_S(Z_i,\alpha_{S4})}{\partial \alpha_S^2}[\pm\varepsilon_n u_n^*, \alpha_{S2}-\alpha_{0S}]V_{Si}\right] \\
&\quad + \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \rho_S(Z_i,\alpha_{S3})}{\partial \alpha_S^2}[\alpha_{S1}-\alpha_{0S}, \alpha_{S2}-\alpha_{0S}]V_{Si}'\right] \\
&\quad \cdot J_S \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \rho_S(Z_i,\alpha_{S4}')}{\partial \alpha_S^2}[\pm\varepsilon_n u_n^*, \alpha_{S2}-\alpha_{0S}]V_{Si}\right]. \qquad \text{(C.60)}
\end{aligned}$$

By Lemma C.1.3, the construction of $\mathscr{B}_n^S(\tau)$, and Assumption 3.4.1, we know

$$\begin{aligned}
&\left| \begin{array}{c} \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \rho_S(Z_i,\alpha_{S3})}{\partial \alpha_S^2}\left[\alpha_{S1}-\alpha_{0S},\alpha_{S2}-\alpha_{0S}\right]V_{Si}'\right] \\ \cdot J_S\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \rho_S(Z_i,\alpha_{0S})}{\partial \alpha_S}\left[\pm\varepsilon_n u_n^*\right]V_{Si}\right] \end{array} \right| \\
=&O_p\left(s\left(2s+k\right)\sqrt{s}\right)\left\|\alpha_{S1}-\alpha_{0S}\right\|_s\left\|\alpha_{S2}-\alpha_{0S}\right\|_s\varepsilon_n = O_p\left(\varepsilon_n^2\right), \qquad \text{(C.61)}
\end{aligned}$$

$$\begin{aligned}
&\left| \begin{array}{c} \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \rho_S(Z_i,\alpha_{0S})}{\partial \alpha_S}\left[\alpha_{S1}-\alpha_{0S}\right]V_{Si}'\right]J_S \\ \cdot\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \rho_S(Z_i,\alpha_{S4})}{\partial \alpha_S^2}\left[\pm\varepsilon_n u_n^*,\alpha_{S2}-\alpha_{0S}\right]V_{Si}\right] \end{array} \right| \\
=&O_p\left(s\left(2s+k\right)\sqrt{s}\right)\left\|\alpha_{S1}-\alpha_{0S}\right\|_s\left\|\alpha_{S2}-\alpha_{0S}\right\|_s\varepsilon_n = O_p\left(\varepsilon_n^2\right), \qquad \text{(C.62)}
\end{aligned}$$

and

$$\begin{aligned}
&\left| \begin{array}{c} \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \rho_S(Z_i,\alpha_{S3})}{\partial \alpha_S^2}\left[\alpha_{S1}-\alpha_{0S},\alpha_{S2}-\alpha_{0S}\right]V_{Si}'\right]J_S \\ \cdot\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \rho_S(Z_i,\alpha_{S4})}{\partial \alpha_S^2}\left[\pm\varepsilon_n u_n^*,\alpha_{S2}-\alpha_{0S}\right]V_{Si}\right] \end{array} \right| \\
=&O_p\left(s^2\left(2s+k\right)\right)\left\|\alpha_{S1}-\alpha_{0S}\right\|_s\left\|\alpha_{S2}-\alpha_{0S}\right\|_s^2\varepsilon_n = O_p\left(\varepsilon_n^2\right). \qquad \text{(C.63)}
\end{aligned}$$

Then the only problem is the first term in (C.60). We will take care of this term later.

The second part of (C.59) is

$$
\begin{aligned}
&\left[\frac{1}{n}\sum_{i=1}^{n}\rho_S(Z_i,\alpha_{S2})V'_{Si}\right]J_S\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\rho_S(Z_i,\alpha_{S2})}{\partial\alpha_S^2}[\pm\varepsilon_n u_n^*,\alpha_{S1}-\alpha_{0S}]V_{Si}\right]\\
=&\left[\frac{1}{n}\sum_{i=1}^{n}\rho_S(Z_i,\alpha_{0S})V'_{Si}\right]J_S\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\rho_S(Z_i,\alpha_{S2})}{\partial\alpha_S^2}[\pm\varepsilon_n u_n^*,\alpha_{S1}-\alpha_{0S}]V_{Si}\right]\\
&+\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\rho_S(Z_i,\alpha_{S5})}{\partial\alpha_S}[\alpha_{S2}-\alpha_{0S}]V'_{Si}\right]J_S\\
&\cdot\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\rho_S(Z_i,\alpha_{S2})}{\partial\alpha_S^2}[\pm\varepsilon_n u_n^*,\alpha_{S1}-\alpha_{0S}]V_{Si}\right]\\
=&\left[\frac{1}{n}\sum_{i=1}^{n}\rho_S(Z_i,\alpha_{0S})V'_{Si}\right]J_S\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\rho_S(Z_i,\alpha_{S2})}{\partial\alpha_S^2}[\pm\varepsilon_n u_n^*,\alpha_{S1}-\alpha_{0S}]V_{Si}\right]\\
&+\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\rho_S(Z_i,\alpha_{0S})}{\partial\alpha_S}[\alpha_{S2}-\alpha_{0S}]V'_{Si}\right]J_S\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\rho_S(Z_i,\alpha_{S2})}{\partial\alpha_S^2}[\pm\varepsilon_n u_n^*,\alpha_{S1}-\alpha_{0S}]V_{Si}\right]\\
&+\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\rho_S(Z_i,\alpha_{S6})}{\partial\alpha_S^2}[\alpha_{S2}-\alpha_{0S},\alpha_{S5}-\alpha_{0S}]V'_{Si}\right]\\
&\cdot J_S\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\rho_S(Z_i,\alpha_{S2})}{\partial\alpha_S^2}[\pm\varepsilon_n u_n^*,\alpha_{S1}-\alpha_{0S}]V_{Si}\right].
\end{aligned}
\tag{C.64}
$$

By Lemma C.1.3, the construction of $\mathscr{B}_n^S(\tau)$, and Assumption 3.4.1,

$$
\begin{aligned}
&\left|\left[\frac{1}{n}\sum_{i=1}^{n}\rho_S(Z_i,\alpha_{S2})V'_{Si}\right]J_S\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\rho_S(Z_i,\alpha_{S2})}{\partial\alpha_S^2}[\pm\varepsilon_n u_n^*,\alpha_{S1}-\alpha_{0S}]V_{Si}\right]\right|\\
=&O_p\left(\varepsilon_n^2\right).
\end{aligned}
\tag{C.65}
$$

We also have

$$
\begin{aligned}
&\mathscr{E}\left(\tilde{Q}_S\left(\alpha_S^*\right)-\tilde{Q}_S\left(\alpha_S\right)\right)=\mathscr{E}\left(\frac{\partial \tilde{Q}_S\left(\alpha_{S1}\right)}{\partial \alpha_S}\left[\pm\varepsilon_n u_n^*\right]\right)\\
=&E\left[\rho_S\left(Z_i,\alpha_{S1}\right)V_{Si}'\right]J_S\frac{\partial E\left[\rho_S\left(Z_i,\alpha_{S1}\right)V_{Si}\right]}{\partial \alpha_S}\left[\pm\varepsilon_n u_n^*\right]\\
=&E\left[\rho_S\left(Z_i,\alpha_{0S}\right)V_{Si}'\right]J_S E\left[\frac{\partial \rho_S\left(Z_i,\alpha_{0S}\right)V_{Si}}{\partial \alpha_S}\left[\pm\varepsilon_n u_n^*\right]\right]\\
&+E\left[\rho_S\left(Z_i,\alpha_{0S}\right)V_{Si}'\right]J_S E\left[\frac{\partial^2 \rho_S\left(Z_i,\alpha_{S7}\right)V_{Si}}{\partial \alpha_S^2}\left[\pm\varepsilon_n u_n^*,\alpha_{S1}-\alpha_{0S}\right]\right]\\
&+E\left[\frac{\partial \rho_S\left(Z_i,\alpha_{S8}\right)}{\partial \alpha_S}\left[\alpha_{S1}-\alpha_{0S}\right]V_{Si}'\right]J_S E\left[\frac{\partial \rho_S\left(Z_i,\alpha_{0S}\right)}{\partial \alpha_S}\left[\pm\varepsilon_n u_n^*\right]V_{Si}\right]\\
&+E\left[\frac{\partial \rho_S\left(Z_i,\alpha_{S8}'\right)}{\partial \alpha_S}\left[\alpha_{S1}-\alpha_{0S}\right]V_{Si}'\right]J_S E\left[\frac{\partial^2 \rho_S\left(Z_i,\alpha_{S7}\right)}{\partial \alpha_S^2}\left[\pm\varepsilon_n u_n^*,\alpha_{S1}-\alpha_{0S}\right]V_{Si}\right], \qquad \text{(C.66)}
\end{aligned}
$$

with

$$
\left|E\left[\frac{\partial \rho_S\left(Z_i,\alpha_{S8}'\right)}{\partial \alpha_S}\left[\alpha_{S1}-\alpha_{0S}\right]V_{Si}'\right]J_S E\left[\frac{\partial^2 \rho_S\left(Z_i,\alpha_{S7}\right)}{\partial \alpha_S^2}\left[\pm\varepsilon_n u_n^*,\alpha_{S1}-\alpha_{0S}\right]\right]\right|=O\left(\varepsilon_n^2\right), \quad \text{(C.67)}
$$

and

$$
\begin{aligned}
&E\left[\frac{\partial \rho_S\left(Z_i,\alpha_{S8}\right)}{\partial \alpha_S}\left[\alpha_{S1}-\alpha_{0S}\right]V_{Si}'\right]J_S E\left[\frac{\partial \rho_S\left(Z_i,\alpha_{0S}\right)}{\partial \alpha_S}\left[\pm\varepsilon_n u_n^*\right]V_{Si}\right]\\
=&E\left[\frac{\partial \rho_S\left(Z_i,\alpha_{0S}\right)}{\partial \alpha_S}\left[\alpha_{S1}-\alpha_{0S}\right]V_{Si}'\right]J_S E\left[\frac{\partial \rho_S\left(Z_i,\alpha_{0S}\right)}{\partial \alpha_S}\left[\pm\varepsilon_n u_n^*\right]V_{Si}\right]\\
&+E\left[\frac{\partial^2 \rho_S\left(Z_i,\alpha_{S9}\right)}{\partial \alpha_S^2}\left[\alpha_{S1}-\alpha_{0S},\alpha_{S8}-\alpha_{0S}\right]V_{Si}'\right]J_S E\left[\frac{\partial \rho_S\left(Z_i,\alpha_{0S}\right)}{\partial \alpha_S}\left[\pm\varepsilon_n u_n^*\right]V_{Si}\right], \qquad \text{(C.68)}
\end{aligned}
$$

where

$$
\left|E\left[\frac{\partial^2 \rho_S\left(Z_i,\alpha_{S9}\right)}{\partial \alpha_S^2}\left[\alpha_{S1}-\alpha_{0S},\alpha_{S8}-\alpha_{0S}\right]V_{Si}'\right]J_S E\left[\frac{\partial \rho_S\left(Z_i,\alpha_{0S}\right)}{\partial \alpha_S}\left[\pm\varepsilon_n u_n^*\right]V_{Si}\right]\right|=O\left(\varepsilon_n^2\right). \quad \text{(C.69)}
$$

The first term of (C.68) is another problem. Now, we consider the first term of (C.60) and the

first term of (C.68). By Lemma C.1.3, the construction of $\mathscr{B}_n^S(\tau)$ and Assumption 3.4.1,

$$\begin{aligned}
&\sqrt{n}\left|\left\{\begin{array}{c}\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \rho_S(Z_i,\alpha_{0S})}{\partial \alpha_S}\left[\alpha_{S1}-\alpha_{0S}\right]V_{Si}'\right]J_S\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \rho_S(Z_i,\alpha_{0S})}{\partial \alpha_S}\left[\varepsilon_n u_n^*\right]V_{Si}\right] \\ -\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \rho_S(Z_i,\alpha_{0S})}{\partial \alpha_S}\left[\alpha_{S1}-\alpha_{0S}\right]V_{Si}'\right]J_S\left\{E\left[\frac{\partial \rho_S(Z_i,\alpha_{0S})}{\partial \alpha_S}\left[\varepsilon_n u_n^*\right]V_{Si}\right]\right\}\end{array}\right\}\right| \\
=&\left|\begin{array}{c}\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \rho_S(Z_i,\alpha_{0S})}{\partial \alpha_S}\left[\alpha_{S1}-\alpha_{0S}\right]V_{Si}'\right]J_S \\ \cdot\frac{1}{\sqrt{n}}\sum\left(\frac{\partial \rho_S(Z_i,\alpha_{0S})}{\partial \alpha_S}\left[\varepsilon_n u_n^*\right]V_{Si}-E\left[\frac{\partial \rho_S(Z_i,\alpha_{0S})}{\partial \alpha_S}\left[\varepsilon_n u_n^*\right]V_{Si}\right]\right)\end{array}\right| = o_p(\varepsilon_n) \qquad \text{(C.70)}
\end{aligned}$$

and

$$\sqrt{n}\left|\left\{\begin{array}{c}\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \rho_S(Z_i,\alpha_{0S})}{\partial \alpha_S}\left[\alpha_{S1}-\alpha_{0S}\right]V_{Si}'\right]J_S\left\{E\left[\frac{\partial \rho_S(Z_i,\alpha_{0S})}{\partial \alpha_S}\left[\varepsilon_n u_n^*\right]V_{Si}\right]\right\} \\ -E\left[\frac{\partial \rho_S(Z_i,\alpha_{0S})}{\partial \alpha_S}\left[\alpha_{S1}-\alpha_{0S}\right]V_{Si}'\right]J_S E\left[\frac{\partial \rho_S(Z_i,\alpha_{0S})}{\partial \alpha_S}\left[\varepsilon_n u_n^*\right]V_{Si}\right]\end{array}\right\}\right| = o_p(\varepsilon_n), \quad \text{(C.71)}$$

Therefore, Lemma C.2.1(i) holds.

Finally, we consider C.2.1(iii). We have that

$$\begin{aligned}
&\frac{\|\alpha_S \pm \varepsilon_n u_n^* - \alpha_0\|^2 - \|\alpha-\alpha_0\|^2}{2} \\
=&\frac{1}{2}\left(\frac{\partial E\left[\rho_S(Z_i,\alpha_{0S})V_{Si}'\right]}{\partial \alpha_S}\left[\alpha_S \pm \varepsilon_n u_n^* - \alpha_0\right]\right)J_S\left(\frac{\partial E\left[\rho_S(Z_i,\alpha_{0S})V_{si}\right]}{\partial \alpha_S}\left[\alpha_S \pm \varepsilon_n u_n^* - \alpha_0\right]\right) \\
&-\frac{1}{2}\left(\frac{\partial E\left[\rho_S(Z_i,\alpha_{0S})V_{Si}'\right]}{\partial \alpha_S}\left[\alpha_S - \alpha_0\right]\right)J_S\left(\frac{\partial E\left[\rho_S(Z_i,\alpha_{0S})V_{si}\right]}{\partial \alpha_S}\left[\alpha_S - \alpha_0\right]\right) \\
=&\left(\frac{\partial E\left[\rho_S(Z_i,\alpha_{0S})V_{Si}'\right]}{\partial \alpha_S}\left[\pm\varepsilon_n u_n^*\right]\right)J_S\left(\frac{\partial E\left[\rho_S(Z_i,\alpha_{0S})V_{si}\right]}{\partial \alpha_S}\left[\alpha_S - \alpha_0\right]\right) \\
&+\frac{1}{2}\left(\frac{\partial E\left[\rho_S(Z_i,\alpha_{0S})V_{Si}'\right]}{\partial \alpha_S}\left[\varepsilon_n u_n^*\right]\right)J_S\left(\frac{\partial E\left[\rho_S(Z_i,\alpha_{0S})V_{si}\right]}{\partial \alpha_S}\left[\varepsilon_n u_n^*\right]\right) \qquad \text{(C.72)}
\end{aligned}$$

and

$$
\begin{aligned}
&\left|\mathscr{E}\left(\tilde{Q}_S(\alpha_S^*)-\tilde{Q}_S(\alpha_S)\right)-\frac{\|\alpha_S\pm\varepsilon_n u_n^*-\alpha_0\|^2-\|\alpha_S-\alpha_0\|^2}{2}\right| \\
=&\left|\mathscr{E}\left(\frac{\partial\tilde{Q}_S(\alpha_{S1})}{\partial\alpha_S}[\pm\varepsilon_n u_n^*]\right)-\frac{\|\alpha_S\pm\varepsilon_n u_n^*-\alpha_0\|^2-\|\alpha_S-\alpha_0\|^2}{2}\right| \\
&+\frac{1}{2}\left(\frac{\partial E\left[\rho_S(Z_i,\alpha_{0S})V_{si}'\right]}{\partial\alpha_S}[\varepsilon_n u_n^*]\right)J_S\left(\frac{\partial E\left[\rho_S(Z_i,\alpha_{0S})V_{si}\right]}{\partial\alpha_S}[\varepsilon_n u_n^*]\right) \\
\leq&\left|\begin{array}{c} E\left[\frac{\partial\rho_S(Z_i,\alpha_{0S})}{\partial\alpha_S}[\alpha_{S1}-\alpha_{0S}]V_{si}'\right]J_S E\left[\frac{\partial\rho_S(Z_i,\alpha_{0S})}{\partial\alpha_S}[\pm\varepsilon_n u_n^*]V_{si}\right] \\ -\left(\frac{\partial E\left[\rho_S(Z_i,\alpha_{0S})V_{si}'\right]}{\partial\alpha_S}[\pm\varepsilon_n u_n^*]\right)J_S\left(\frac{\partial E\left[\rho_S(Z_i,\alpha_{0S})V_{si}\right]}{\partial\alpha_S}[\alpha_S-\alpha_0]\right)\end{array}\right|+O\left(\varepsilon_n^2\right) \\
=&\left|\left(\frac{\partial E\left[\rho_S(Z_i,\alpha_{0S})V_{si}'\right]}{\partial\alpha_S}[\pm\varepsilon_n u_n^*]\right)J_S E\left[\frac{\partial\rho_S(Z_i,\alpha_{0S})}{\partial\alpha_S}[\alpha_{S1}-\alpha_S]V_{si}\right]\right|+O\left(\varepsilon_n^2\right). \qquad \text{(C.73)}
\end{aligned}
$$

Notice that $\alpha_{S1}=\alpha_S\pm\varsigma\varepsilon_n u_n^*$, then

$$
\left|\mathscr{E}\left(\hat{Q}_S(\alpha_S^*)-\hat{Q}_S(\alpha_S)\right)-\frac{\|\alpha_S\pm\varepsilon_n u_n^*-\alpha_0\|^2-\|\alpha_S-\alpha_0\|^2}{2}\right|=O\left(\varepsilon_n^2\right). \qquad \text{(C.74)}
$$

This implies C.2.1(iii). ■

Proof of Theorem 3.4.1.

The proof of Theorem 3.4.1 closely follows that of theorem 3.4 in Chen *et al.* (2014), while we consider the high-dimensional case. Because $\hat{\alpha}_S$ is a global minimizer in $\mathscr{B}_n^S(\tau)$, by Lemmas C.2.1(i), (ii), and the construction of $\hat{\alpha}_S^*$,

$$
\begin{aligned}
0\leq&\hat{Q}_S(\hat{\alpha}_S^*)-\hat{Q}_S(\hat{\alpha}_S)=\tilde{Q}_S(\hat{\alpha}_S^*)-\tilde{Q}_S(\hat{\alpha}_S)+\sum_{j=1}^{s}\left[P_n\left(\left|\hat{\theta}_{Sj}^*\right|\right)-P_n\left(\left|\hat{\theta}_{Sj}\right|\right)\right] \\
=&\mathscr{E}\left(\tilde{Q}_S(\hat{\alpha}_S^*)-\tilde{Q}_S(\hat{\alpha}_S)\right)+\Delta(\alpha_0)[\hat{\alpha}_S^*-\hat{\alpha}_S]+\tilde{Q}_S(\hat{\alpha}_S^*)-\tilde{Q}_S(\hat{\alpha}_S)-\mathscr{E}\left(\tilde{Q}_S(\hat{\alpha}_S^*)-\tilde{Q}_S(\hat{\alpha}_S)\right) \\
&-\Delta(\alpha_0)[\hat{\alpha}_S^*-\hat{\alpha}_S]+\sum_{j=1}^{s}\left[P_n\left(\left|\hat{\theta}_{Sj}^*\right|\right)-P_n\left(\left|\hat{\theta}_{Sj}\right|\right)\right] \\
\leq&\mathscr{E}\left(\tilde{Q}_S(\hat{\alpha}_S^*)-\tilde{Q}_S(\hat{\alpha}_S)\right)\pm\Delta(\alpha_0)[\varepsilon_n u_n^*]+O_p\left(\varepsilon_n^2\right). \qquad \text{(C.75)}
\end{aligned}
$$

The first inequality holds because we can always make $\tau_1(n)$ and $\tau_2(n)$ to be a little larger so that α_S^* is included in the boundary of $\mathscr{B}_n^S(\tau)$.

By Lemma C.2.1(iii),

$$\mathscr{E}\left(\tilde{Q}_S(\hat{\alpha}_S^*) - \tilde{Q}_S(\hat{\alpha}_S)\right) = \pm\varepsilon_n\langle\hat{\alpha}_S - \alpha_{0S}, u_n^*\rangle + O_p\left(\varepsilon_n^2\right). \tag{C.76}$$

Then

$$\pm\varepsilon_n\langle\hat{\alpha}_S - \alpha_{0S}, u_n^*\rangle \pm \Delta(Z, \alpha_{0S})[\varepsilon_n u_n^*] + O_p\left(\varepsilon_n^2\right) \geq 0, \tag{C.77}$$

which implies

$$\left|\varepsilon_n\langle\hat{\alpha}_S - \alpha_{0S}, u_n^*\rangle + \Delta(Z, \alpha_{0S})[\varepsilon_n u_n^*]\right| = O_p\left(\varepsilon_n^2\right). \tag{C.78}$$

With $\langle\alpha_{n,0} - \alpha_{S0}, u_n^*\rangle = 0$,

$$\sqrt{n}\left|\langle\hat{\alpha}_S - \alpha_{0S,n}, u_n^*\rangle + \Delta(Z, \alpha_{0S})[u_n^*]\right| = o_p(1). \tag{C.79}$$

By assumption 3.4.2(ii) and the Riesz representation theorem,

$$\begin{aligned}
&\frac{f(\hat{\alpha}_S) - f\left(\alpha_{0S,n}\right)}{\|v_n^*\|_{sd}} = \frac{f(\hat{\alpha}_S) - f(\alpha_{0S}) - \frac{\partial f(\alpha_{0S})}{\partial \alpha_S}[\hat{\alpha}_S - \alpha_{0S}]}{\|v_n^*\|_{sd}} \\
&- \frac{f\left(\alpha_{0S,n}\right) - f(\alpha_{0S}) - \frac{\partial f(\alpha_{0S})}{\partial \alpha_S}\left[\alpha_{0S,n} - \alpha_{0S}\right]}{\|v_n^*\|_{sd}} + \frac{\frac{\partial f(\alpha_{0S})}{\partial \alpha_S}[\hat{\alpha}_S - \alpha_{0S}] - \frac{\partial f(\alpha_{0S})}{\partial \alpha_S}\left[\alpha_{0S,n} - \alpha_{0S}\right]}{\|v_n^*\|_{sd}} \\
=&\langle\hat{\alpha}_S - \alpha_{0S,n}, u_n^*\rangle + o_p\left(1/\sqrt{n}\right).
\end{aligned} \tag{C.80}$$

Then we have

$$\left|\sqrt{n}\frac{f(\hat{\alpha}_S) - f\left(\alpha_{0S,n}\right)}{\|v_n^*\|_{sd}} - \sqrt{n}\Delta(Z, \alpha_{0S})[u_n^*]\right| = o_p(1). \tag{C.81}$$

The theorem follows from (C.81) and Assumption 3.4.2(iii). ■

Bibliography

Abadie, Alberto. 2002. Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association*, **97**(457), 284–292.

Abadie, Alberto, Angrist, Joshua, & Imbens, Guido. 2002. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, **70**(1), 91–117.

Ai, Chunrong, & Chen, Xiaohong. 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, **71**(6), 1795–1843.

Aliprantis, Charalambos D, & Border, Kim. 2006. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Science & Business Media.

Andrews, Donald W K, & Shi, Xiaoxia. 2013. Inference based on conditional moment inequalities. *Econometrica*, **81**(2), 609–666.

Angrist, Joshua D. 1990. Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *The American Economic Review*, **80**(3), 313–336.

Angrist, Joshua D, & Imbens, Guido W. 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, **90**(430), 431–442.

Angrist, Joshua D, & Krueger, Alan B. 1991. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, **106**(4), 979–1014.

Angrist, Joshua D, & Krueger, Alan B. 1992. *Estimating the payoff to schooling using the Vietnam-era draft lottery*. Working Paper. National Bureau of Economic Research.

Angrist, Joshua D, & Krueger, Alan B. 1995. Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, **13**(2), 225–235.

Angrist, Joshua D, Imbens, Guido W, & Rubin, Donald B. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, **91**(434),

444–455.

Antoniadis, Anestis. 1996. Smoothing noisy data with tapered coiflets series. *Scandinavian Journal of Statistics*, **23**(3), 313–330.

Armstrong, Timothy B. 2014. Weighted KS statistics for inference on conditional moment inequalities. *Journal of Econometrics*, **181**(2), 92–116.

Armstrong, Timothy B, & Chan, Hock Peng. 2016. Multiscale adaptive inference on conditional moment inequalities. *Journal of Econometrics*, **194**(1), 24–43.

Balke, Alexander, & Pearl, Judea. 1997. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, **92**(439), 1171–1176.

Barrett, Garry F, & Donald, Stephen G. 2003. Consistent tests for stochastic dominance. *Econometrica*, **71**(1), 71–104.

Barrett, Garry F, Donald, Stephen G, & Bhattacharya, Debopam. 2014. Consistent nonparametric tests for Lorenz dominance. *Journal of Business & Economic Statistics*, **32**(1), 1–13.

Beare, Brendan K, & Fang, Zheng. 2017. Weak convergence of the least concave majorant of estimators for a concave distribution function. *Electronic Journal of Statistics*, **11**(2), 3841–3870.

Beare, Brendan K, & Moon, Jong-Myun. 2015. Nonparametric tests of density ratio ordering. *Econometric Theory*, **31**(3), 471–492.

Beare, Brendan K., & Shi, Xiaoxia. 2018. *An improved bootstrap test of density ratio ordering*. Working Paper.

Bhattacharya, Debopam. 2007. Inference on inequality from household survey data. *Journal of Econometrics*, **137**(2), 674–707.

Bishop, John A, Formby, John P, & Smith, W James. 1991a. International comparisons of income inequality: Tests for Lorenz dominance across nine countries. *Economica*, **58**(232), 461–477.

Bishop, John A, Formby, John P, & Smith, W James. 1991b. Lorenz dominance and welfare: Changes in the US distribution of income, 1967-1986. *The Review of Economics and Statistics*, **73**(1), 134–139.

Bondell, Howard D, & Reich, Brian J. 2012. Consistent high-dimensional Bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association*, **107**(500), 1610–1624.

Bradic, Jelena, Fan, Jianqing, & Wang, Weiwei. 2011. Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(3), 325–349.

Carolan, Christopher A, & Tebbs, Joshua M. 2005. Nonparametric tests for and against likelihood ratio ordering in the two-sample problem. *Biometrika*, **92**(1), 159–171.

Chen, Baicheng, Yu, Yao, Zou, Hui, & Liang, Hua. 2012. Profiled adaptive Elastic-Net procedure for partially linear models with high-dimensional covariates. *Journal of Statistical Planning and Inference*, **142**(7), 1733–1745.

Chen, Xiaohong, & Pouzo, Demian. 2009. Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, **152**(1), 46–60.

Chen, Xiaohong, & Pouzo, Demian. 2012. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, **80**(1), 277–321.

Chen, Xiaohong, Liao, Zhipeng, & Sun, Yixiao. 2014. Sieve inference on possibly misspecified semi-nonparametric time series models. *Journal of Econometrics*, **178**, 639–658.

Chernozhukov, Victor, Fernández-Val, Iván, & Galichon, Alfred. 2010. Quantile and probability curves without crossing. *Econometrica*, **78**(3), 1093–1125.

Chernozhukov, Victor, Lee, Sokbae, & Rosen, Adam M. 2013. Intersection bounds: Estimation and inference. *Econometrica*, **81**(2), 667–737.

Chernozhukov, Victor, Kim, Wooyoung, Lee, Sokbae, & Rosen, Adam. 2014. *Implementing Intersection Bounds in Stata*. Working Paper. Centre for Microdata Methods and Practice.

Chetverikov, Denis. 2018. Adaptive tests of conditional moment inequalities. *Econometric Theory*, **34**(1), 186–227.

Cornelissen, Thomas, Dustmann, Christian, Raute, Anna, & Schönberg, Uta. 2016. From LATE to MTE: Alternative methods for the evaluation of policy interventions. *Labour Economics*, **41**, 47–60.

Csörgö, Miklós, Csörgö, Sándor, & Horváth, Lajos. 1986. *An Asymptotic Theory for Empirical Reliability and Concentration Processes*. Vol. 33. Springer.

Dardanoni, Valentino, & Forcina, Antonio. 1999. Inference for Lorenz curve orderings. *The Econometrics Journal*, **2**(1), 49–75.

Daubechies, Ingrid, Defrise, Michel, & De Mol, Christine. 2004. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and*

Applied Mathematics, **57**(11), 1413–1457.

Davidson, Russell, & Duclos, Jean-Yves. 2000. Statistical inference for stochastic dominance and for the measurement of poverty and inequality. *Econometrica*, **68**(6), 1435–1464.

Davydov, Yu. A., Lifshits, M. A., & Smorodina, N. V. 1998. *Local Properties of Distributions of Stochastic Functionals*. Vol. 173. American Mathematical Society.

Delgado, Miguel A, & Escanciano, Juan Carlos. 2012. Distribution-free tests of stochastic monotonicity. *Journal of Econometrics*, **170**(1), 68–75.

Delgado, Miguel A, & Escanciano, Juan Carlos. 2013. Conditional stochastic dominance testing. *Journal of Business & Economic Statistics*, **31**(1), 16–28.

Donald, Stephen G, & Hsu, Yu-Chin. 2016. Improving the power of tests of stochastic dominance. *Econometric Reviews*, **35**(4), 553–585.

Dümbgen, Lutz. 1993. On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*, **95**(1), 125–140.

Fan, Jianqing, & Li, Runze. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**(456), 1348–1360.

Fan, Jianqing, & Liao, Yuan. 2014. Endogeneity in high dimensions. *The Annals of Statistics*, **42**(3), 872–917.

Fan, Jianqing, & Lv, Jinchi. 2011. Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, **57**(8), 5467–5484.

Fang, Zheng, & Santos, Andres. 2014. *Inference on directionally differentiable functions*. Working Paper.

Fu, Wenjiang J. 1998. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, **7**(3), 397–416.

Giacomini, Raffaella, Politis, Dimitris N, & White, Halbert. 2013. A warp-speed method for conducting Monte Carlo experiments involving bootstrap estimators. *Econometric Theory*, **29**(3), 567–589.

Goldie, Charles M. 1977. Convergence theorems for empirical Lorenz curves and their inverses. *Advances in Applied Probability*, **9**(4), 765–791.

Hansen, Bruce E. 2017. Regression kink with an unknown threshold. *Journal of Business & Economic Statistics*, **35**(2), 228–240.

Heckman, James J, & Vytlacil, Edward. 2005. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, **73**(3), 669–738.

Hirano, Keisuke, & Porter, Jack R. 2012. Impossibility results for nondifferentiable functionals. *Econometrica*, **80**(4), 1769–1790.

Hong, Han, & Li, Jessie. 2016. *The Numerical Directional Delta Method*. Working Paper.

Horowitz, Joel L. 1992. A smoothed maximum score estimator for the binary response model. *Econometrica*, **60**(3), 505–531.

Horváth, Lajos, Kokoszka, Piotr, & Zitikis, Ričardas. 2006. Testing for stochastic dominance using the weighted McFadden-type statistic. *Journal of Econometrics*, **133**(1), 191–205.

Huber, Martin, & Mellace, Giovanni. 2015. Testing instrument validity for LATE identification based on inequality moment constraints. *Review of Economics and Statistics*, **97**(2), 398–411.

Imbens, Guido W, & Angrist, Joshua D. 1994. Identification and estimation of local average treatment effects. *Econometrica*, **62**(2), 467–475.

Imbens, Guido W, & Rubin, Donald B. 1997. Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, **64**(4), 555–574.

Kaji, Tetsuya. 2017. *Switching to the new norm: From heuristics to formal tests using integrable empirical processes*. Working Paper.

Kitagawa, Toru. 2015. A test for instrument validity. *Econometrica*, **83**(5), 2043–2063.

Liang, Hua, Liu, Xiang, Li, Runze, & Tsai, Chih-Ling. 2010. Estimation and testing for partially linear single-index models. *Annals of Statistics*, **38**(6), 3811–3836.

Linton, Oliver, Song, Kyungchul, & Whang, Yoon-Jae. 2010. An improved bootstrap test of stochastic dominance. *Journal of Econometrics*, **154**(2), 186–202.

McFadden, Daniel. 1989. Testing for stochastic dominance. *Pages 113–134 of: Studies in the Economics of Uncertainty*. Springer.

Mourifié, Ismael, & Wan, Yuanyuan. 2017. Testing local average treatment effect assumptions. *Review of Economics and Statistics*, **99**(2), 305–313.

Newey, Whitney K, & Powell, James L. 2003. Instrumental variable estimation of nonparametric models. *Econometrica*, **71**(5), 1565–1578.

Ni, Xiao, Zhang, Hao Helen, & Zhang, Daowen. 2009. Automatic model selection for partially

linear models. *Journal of Multivariate Analysis*, **100**(9), 2100–2111.

Peng, Heng, & Huang, Tao. 2011. Penalized least squares for single index models. *Journal of Statistical Planning and Inference*, **141**(4), 1362–1379.

Seo, Juwon. Forthcoming. Tests of stochastic monotonicity with improved power. *Journal of Econometrics*.

Sun, Zhenting, & Beare, Brendan K. 2018. *Improved nonparametric bootstrap tests of Lorenz dominance*. Working Paper.

van der Vaart, Aad W, & Wellner, Jon A. 1996. *Weak Convergence and Empirical Processes*. Springer.

Vytlacil, Edward. 2002. Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, **70**(1), 331–341.

Vytlacil, Edward. 2006. Ordered discrete-choice selection models and local average treatment effect assumptions: Equivalence, nonequivalence, and representation results. *The Review of Economics and Statistics*, **88**(3), 578–581.

Xie, Huiliang, & Huang, Jian. 2009. SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics*, **37**(2), 673–696.

Zhang, Cun-Hui. 2010. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**(2), 894–942.

Zhu, Li-Ping, & Zhu, Li-Xing. 2009. Nonconcave penalized inverse regression in single-index models with high dimensional predictors. *Journal of Multivariate Analysis*, **100**(5), 862–875.