ESSAYS ON SEMIPARAMETRIC MODELS WITH PARTIAL IDENTIFICATION

by

Shengjie Hong

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Economics)

at the

UNIVERSITY OF WISCONSIN-MADISON

2012

Date of final oral examination: 04/18/2012

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

Bruce E. Hansen, Professor, Economics
Jack R. Porter, Professor, Economics
Xiaoxia Shi, Assistant Professor, Economics
Chao Fu, Assistant Professor, Economics
Xiaodong Du, Assistant Professor, Agricultural and Applied Economics

Acknowledgment

Completing this dissertation would not have been possible without the aid and support of many people. I owe my deepest gratitude to my two advisors, Bruce Hansen and Jack Porter. They have devoted countless hours to advising me. Their continuous guidance and encouragement helped me tremendously at all phases of my research over the past five years, and will lift my spirits for years to come. Xiaoxia Shi has been an ideal third member of my committee. She provided insightful advice and always came to my seminars with constructive comments. Steven Durlauf deserves special thanks for providing direction in my pursuit of an academia position. Many other professors contributed to improving the quality of this dissertation through helpful discussion, Andres Aradillas-Lopez, Ivan Canay, Xiaodong Du, Chao Fu, John Kennan, Demian Pouzo, Daniel Quint, William Sandholm, Andres Santos, Alan Sorensen, Christopher Taber, Yu-Chin Xu and Ping Yu.

The econometrics reading group in the department has been an important source of knowledge to me and I am grateful to them: SeoJeong Lee, Ying-ying Lee, Jen-Che Liao, Chu-An Liu, Enrique Pinzon, Nelson Ramirez, Jing Tao, and Jin Yan. Andrew Anderson, Michael Anderson, Yu Fai Choi, and Toshinori Onuma have been perfect office-mates since my first day at graduate school. Chao He, Haixi Li, Hsuan-Chih Lin, and Michael Pistone were always there to cheer me up.

I can never adequately thank my family. My parents have been a constant source of

support throughout my life. Grandpa always encourages me to keep learning. My wife Xiaoyue Wang's dedication, love and persistent confidence in me make me a better and happier person.

Table of Contents

Abstract

This dissertation consists of two self-contained essays on partially identified econometric models, organized in the form of two chapters.

The first chapter develops inference methods for conditional moment models in which the unknown parameter is possibly partially identified and may contain infinite-dimensional components. I consider testing the hypothesis that a given restriction on the parameter is satisfied by at least one element of the identification set. I propose using the sieve minimum of a Kolmogorov-Smirnov type statistic as the test statistic, derive its asymptotic distribution, and provide consistent bootstrap critical values. In this way a broad family of restrictions can be consistently tested, making the proposed procedure applicable to various types of inference. In particular, I show how to: (1) test the semiparametric model specification; (2) construct confidence sets for unknown parametric components; and (3) construct confidence sets for unknown functions at a given point. The specification test is consistent against fixed alternatives. The confidence sets have correct asymptotic coverage probability, excluding any value outside the identification set with asymptotic probability one. My methods are robust to partial identification, and allow for the moment functions to be nonsmooth. A Monte Carlo study demonstrates finite sample performance.

In the second chapter, I consider estimation in dynamic discrete choice panel data models of short time series, in which neither the cross-sectional heterogeneity nor the

initial condition is observed. The major challenges are: (1) point-identification often fails in these models as demonstrated by Honoré and Tamer (2006); and (2) the heterogeneity cannot be differenced out by the standard "within" or first difference transformations due to nonlinearity. I show that the parameter can be equivalently defined by a finite number of conditional moment equalities. And I propose set estimators that are fixed-T consistent with respect to a properly defined Hausdorff distance. Rates of convergence in the Hausdorff distance are derived.

Chapter 1

Inference in Semiparametric Conditional Moment Models with Partial Identification

1.1 Introduction

In this paper, we consider inference in conditional moment models of the following form:

$$E\left[m\left(Y,\,\theta_0,\,h_0\left(X\right)\right)|Z\right]=0 \tag{1.1}$$

where the vector-valued moment function $m\left(\cdot\right)$ is known up to a finite dimensional parameter $\theta_0 \in \Theta$ and an unknown function $h_0 \in \mathcal{H} : \mathcal{X} \to \mathbb{R}^H$. $(X,\,Y,\,Z)$ are the observable variables whose true probability distributions are unknown[1]. Model (1.1) is

[1]We allow for $X,\,Y,\,Z$ to have common components.

a natural semiparametric extension of the parametric conditional moment models

$$E\left[m\left(Y,\, X,\, \theta_0\right)|Z\right] = 0 \tag{1.2}$$

which are well known from the work of Hansen (1982), Chamberlain (1987), Newey (1990), and others. We allow for the unknown parameter $\alpha_0 \equiv (\theta_0,\, h_0)$ to be partially identified in the sense that the identification set

$$\mathcal{A}_I \equiv \{\alpha \in \Theta \times \mathcal{H} :\, E\left[m\left(Y,\, \theta,\, h\left(X\right)\right)|Z\right] = 0 \,\text{a.s. on}\, \mathcal{Z}\,\}$$

may have more than one element.

The conditional moment restrictions in (1.1) have been studied extensively under point-identification assumptions. This line of work led to a rich literature, including important papers such as Newey and Powell (2003) and Ai and Chen (2003). Newey and Powell (2003) provide consistent estimators for α_0. Ai and Chen (2003) establish the $\sqrt{n}$ asymptotic normality and efficiency of their estimators for θ_0. Recently, Chen et al. (2011) derive sufficient conditions for achieving local point-identification in Model (1.1). These regularity conditions provide new insights for understanding both semiparametric identification and its challenges. Canay et al. (2011) study the testability of necessary conditions for point-identification in some nonparametric models with endogeneity.

A growing number of papers have drawn attention to cases where point-

identification fails in special forms of model (1.1). For instance, Santos (2011) shows examples of partial identification in nonparametric instrumental variable (IV) regressions. As another example, Aradillas-Lopez (2010) shows examples of partial identification in the context of incomplete information entry-games. The possible lack of point-identification makes it desirable to allow for partial identification in the conditional moment models. Pioneered by Manski (1990, 2003), partial identification analysis for parametric models has been a significant development. See, for examples, Chernozhukov, Hong, and Tamer (2007), Bugni (2010), Canay (2010), Romano and Shaikh (2010) and Andrews and Shi (2011). The literature on partially identified non/semiparametric models is more recent and very limited. Notably, Santos (2011) considers hypothesis testing under a nonparametric IV specification with partial identification.

We extend the existing literature by developing methods for hypothesis testing under the semiparametric specification of Model (1.1) without assuming point-identification. First, we propose a general procedure for testing the hypothesis that a given restriction on the parameter α is satisfied by at least one element of the identification set $\mathcal{A}_I$. Then we show that a broad family of restrictions can be consistently tested, making the proposed procedure applicable to various types of inference. In particular, we demonstrate how to test the model specification, construct confidence sets for the parametric component θ, and construct confidence sets for the unknown function $h(\cdot)$ at a given point. Our methods are robust to partial identification, and allow for the moment function $m(\cdot)$ to be pointwise nonsmooth in α.

Our confidence sets for θ, denoted by CS_n, have the correct asymptotic coverage probability in the following sense: For a targeted level $1-\rho$,

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta_I} Pr\left(\theta \in \text{CS}_n\left(1-\rho\right)\right) \geq 1-\rho \tag{1.3}$$

where Θ_I is the identification set for θ. The infimum in (1.3) is carried out before the limit is taken, which implies uniformity of correct asymptotic coverage probability over Θ_I.

A potential alternative way of constructing CS's for θ is to project the CS's for $(\theta,\, h)$ onto Θ. However, this alternative has the following drawbacks: (i) As pointed out by Hahn and Ridder (2009), CS's built from projection of some higher dimensional CS's can be very conservative; (ii) To the best of our knowledge, among existing methods, Andrews and Shi (2011)'s is the only one that can be potentially used for constructing CS's for the infinite-dimensional $(\theta,\, h)$ with partial identification. Yet, Andrews and Shi (2011) are primarily concerned with finite-dimensional parameters. The computational cost of using their method for infinite-dimensional parameters can be prohibitive; (iii) In general it is impossible to write out the resulting CS's for θ in a form that is practically useful.

Conditional moment model (1.1) has been recognized as an important branch of semiparametric modeling in applied econometrics. It encompasses nonparametric IV ($E\left[Y-h_0\left(X\right)|Z\right]=0$), partially linear IV ($E\left[Y-X_1'\theta_0-h_0\left(X_2\right)|Z\right]=0$), single index IV ($E\left[Y-h_0\left(X'\theta_0\right)|Z\right]=0$), nonparametric quantile IV ($E\left[1\left\{Y\leq h_0\left(X\right)\right\}|Z\right]=\tau$),

and semiparametric quantile IV as special cases. Model (1.1) also arises in many interaction-based economic environments with optimizing agents. For example, consider the incomplete information entry-game in Aradillas-Lopez (2010): It can be shown that the parameters in the payoff functions are characterized by conditional moment equalities. And the moment functions would contain the unobserved belief under Bayesian-Nash equilibrium as a nuisance parameter.

Our paper is related to the large literature on conditional moment restrictions. To list a few, papers belonging to this literature include Ai and Chen (2003), Newey and Powell (2003), Dominguez and Lobato (2004), Chen and Pouzo (2009), Horowitz (2009), Kim (2009), Andrews and Shi (2011), and Santos (2011). Our paper is also related to the rapidly growing literature on partial identification analysis, such as Manski (2003), Imbens and Manski (2004), Honoré and Tamer (2006), Chernozhukov et al. (2007), Beresteanu and Molinari (2008), Kim (2009), Hahn and Ridder (2009), Andrews and Soares (2010), Bugni (2010), Canay (2010), Romano and Shaikh (2010), Andrews and Shi (2011), Santos (2011), and many more. Our test procedure is most closely related to Santos (2011).

In Section 1.2, we regularize the parameter space $\mathcal{A}$ and formally describe the hypothesis test to be considered. This section serves as a road map for the following sections. In Section 1.3, we introduce the general testing procedure, derive the asymptotic distribution of the test statistic, and show consistency results for the hypothesis test. Bootstrap critical values are provided. We discuss the specification test as a spe-

cial case of the general test at the end of this section. In Section 1.4, we show how to utilize the test procedure developed in Section 1.3 to construct confidence sets. Finite sample performance is studied by Monte Carlo simulations in Section 1.5. Section 1.6 is an empirical illustration of the proposed inference methods applied to study the quantile IV Engel curves for gasoline. Section 1.7 concludes. Mathematical proofs of all the theorems stated in the paper are given in Appendix A.

1.2 Inference with Partial Identification

Model (1.1) defines the true value of the parameter $\alpha = (\theta, h(\cdot))$ by the following conditional moment restriction:

$$E[m(Y, \theta, h(X))|Z] = 0$$

where θ is finite dimensional, and h is infinite dimensional.

We say Model (1.1), or, interchangeably, the parameter $\alpha = (\theta, h)$ in Model (1.1), is partially identified if the identification set

$$\mathcal{A}_I \equiv \{(\theta, h) \in \Theta \times \mathcal{H} : E[m(Y, \theta, h(X))|Z] = 0\}$$

contains more than one element. Santos (2011) provides several examples of partially identified nonparametric IV. In our simulation study (reported in Section 5) we consider

a partially identified semiparametric quantile IV, which is another example of partially identified Model (1.1).

In what follows, we first regularize the parameter space $\mathcal{A} = \Theta \times \mathcal{H}$. Then we formally describe the general form of the hypothesis tests that we consider.

1.2.1 Restrictions on The Parameter Space $\mathcal{A}$

The parameter space $\mathcal{A}$ takes the form of a product space: $\Theta \times \mathcal{H}$. We put the usual compactness assumption on $\Theta \in \mathbb{R}^{d_\theta}$. Smoothness assumptions on the unknown function are often necessary for achieving consistency of inference in non/semiparametric models. For example, Horowitz (2009) shows that it is impossible to consistently test the specification of a nonparametric IV if nonsmooth functions are allowed. It is also reasonable to believe that the unknown function is smooth in many economic applications. For the conditional moment models, we impose the same smoothness assumption on $\mathcal{H}$ as in Santos (2011): we require $\mathcal{H}$ to be a subset of a Sobolev space. Similar assumptions have been made by Ai and Chen (2003), Newey and Powell (2003) and many other authors.

Definition 1.2.1. (i) Let $X \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$. For $\lambda \in \mathbb{N}_+^{d_x}$, $|\lambda| \equiv \sum_{i=1}^{d_x} \lambda_i$ and $D^\lambda h(x) \equiv \partial^{|\lambda|} h(x) / \partial x_1^{\lambda_1} ... \partial x_{d_x}^{\lambda_{d_x}}$. Define[2]:

[2] $\|\cdot\|_E$ is the Euclidean norm, i.e. $\|a\|_E = \sqrt{a'a}$.

$$\|h\|_s^2 = \sum_{|\lambda| \leq d} \int_{\mathcal{X}} \|D^\lambda h\,(x)\,\|_E^2 dx$$

where h is assumed to be d-times differentiable.

(ii) Define the following Sobolev space:

$$W^s\,(\mathcal{X}) \equiv \left\{h:\, \mathcal{X} \to \mathbb{R}^H \text{ s.t. } h \text{ is } d\text{-times differentiable and } \|h\|_s < \infty\right\}.$$

We make the following assumption on the parameter space $\mathcal{A} = \Theta \times \mathcal{H}$:

Assumption 1.2.1. (i) Θ is a compact subset of $\mathbb{R}^{d_\theta}$; (ii) $d \geq d_x + 2$, and $\mathcal{H} = \{h \in W^s\,(\mathcal{X}) :\, \|h\|_s \leq B\}$ for some $B < \infty$.

Assumption 1.2.1 (ii) requires that h is at least $(d_x + 2)$-times differentiable and belongs to a bounded subset of the Sobolev space $W^s\,(\mathcal{X})$. As shown by Gallant and Nychka (1987), Assumption 1.2.1 (ii) implies that $\mathcal{H}$ is compact under the norm $\|\cdot\|_c$ defined as

$$\|h\|_c \equiv \max_{|\lambda| \leq \left[\frac{d}{2}\right]} \sup_{x \in \mathcal{X}} \|D^\lambda h\,(x)\,\|_E,$$

an important property needed in our asymptotic analysis.

1.2.2 Hypothesis Test

For a given restriction on the parameter α, we want to test the hypothesis that at least one element of the identification set $\mathcal{A}_I$ satisfies the restriction. This notion of testing encompasses hypothesis tests under point-identification as a special case where we test whether the true value (the unique element of $\mathcal{A}_I$) satisfies a given restriction.

More specifically, denote by $L(\alpha) = a$ a restriction on α that we want to test for, where L is a mapping on $\mathcal{A}$, and a is a constant that belongs to certain Banach space. Define the restricted set R as the set of parameter values that satisfy the given restriction:

$$R \equiv \{\alpha \in \mathcal{A} : L(\alpha) = a\}. \tag{1.4}$$

Then the hypothesis test takes the following form:

$$\begin{aligned} \mathrm{H}_0 : &\quad \mathcal{A}_I \cap R \neq \varnothing; \\ \mathrm{H}_1 : &\quad \mathcal{A}_I \cap R = \varnothing, \end{aligned}$$

where $\varnothing$ denotes the empty set.

We impose the following assumption on the restriction $L(\alpha) = a$ that we consider testing:

Assumption 1.2.2. We only consider testing restrictions of the form $L(\alpha) = a$ where $L : (\mathcal{A}, \|\cdot\|_c) \to (\mathcal{L}, \|\cdot\|_{\mathcal{L}})$ is a bounded linear operator.

The seemingly restrictive requirements of boundedness and linearity of Assumption 1.2.2 are compensated by flexibility in choosing the Banach space $(\mathcal{L}, \|\cdot\|_{\mathcal{L}})$. And Assumption 1.2.2 is satisfied by a broad family of restrictions. For example, it is satisfied by the restriction $L(\alpha) = 0$ with $L(\cdot) \equiv 0$, which leads to the model specification test: $\mathrm{H}_0 : \mathcal{A}_I \neq \varnothing$ vs $\mathrm{H}_1 : \mathcal{A}_I = \varnothing$. Assumption 1.2.2 is also satisfied by the restrictions (i) $\theta = \bar{\theta}$ for a constant $\bar{\theta} \in \Theta$ and (ii) $h(x_0) = \mu$ for a given $x_0 \in \mathcal{X}$ and a constant μ. As it will be shown, consistent confidence sets for θ and $h(x_0)$ can be built from inverting the tests for restrictions (i) and (ii), respectively. We discuss the specification test in detail at the end of Section 1.3. And we study the asymptotic properties of the confidence sets in Section 1.4. For more examples of restrictions that satisfy Assumption 1.2.2, see Santos (2011).

In the next section, we develop a general procedure for testing a generic restriction which satisfies the assumption specified above.

1.3 The Test Procedure

A popular method of handling conditional moment restrictions is transforming them into a number of (in some case infinitely many) unconditional moment restrictions. This method dates back to Bierens (1982) and is adopted by recent papers such as Dominguez and Lobato (2004), Kim (2009), Santos (2011), and Andrews and Shi (2011). We also adopt this method.

With partial identification, the transformation from conditional moment restrictions into unconditional ones needs to be done carefully in order to prevent any loss of identification power. Therefore, we require a family of instrument functions $\{g(t, \cdot) : \mathcal{Z} \to \mathbb{R}, t \in T\}$ such that for all random variables V with $E(|V|) < \infty$:

$$E(V|Z) = 0 \text{ iff } E[V \cdot g(t, Z)] = 0 \text{ for all } t \in T. \tag{1.5}$$

Consequently, $E[m(Y, \theta, h(X))|Z] = 0$ iff

$$E[m(Y, \theta, h(X)) \cdot g(t, Z)] = 0 \text{ for all } t \in T.$$

And the identification set $\mathcal{A}_I$ can be equivalently written as

$$\mathcal{A}_I = \{(\theta, h) \in \Theta \times \mathcal{H} : E[m(Y, \theta, h(X)) \cdot g(t, Z)] = 0 \, \forall t \in T\}.$$

Instrument functions that satisfy condition(1.5) often exist for example: $g(t, z) = 1(z \leq t)$[3] with $t \in \mathcal{Z}$ as shown by Dominguez and Lobato (2004) and $g(t, z) = e^{t'z}$ with any T of positive Lebesgue measure if Z is bounded almost sure as shown by Bierens (1990). A detailed discussion on valid choices of instrument functions can be found in Andrews and Shi (2011).

In summary, we impose the following assumption on the instrument functions:

[3] $1(z \leq t)$ is the indicator function. It equals 1 if each component in z is less than or equal to the corresponding component in t, and equals 0 otherwise.

Assumption 1.3.1. (i) g satisfies condition (1.5); (ii) $T \in \mathbb{R}^{d_t}$, and g satisfies one of the following conditions: a) $g(t,\, z) = 1\{z \leq t\}$ for $t \in \mathcal{Z}$, or b) T is compact, and $\exists\, M > 0$ s.t. $|g(t_1,\, z) - g(t_2,\, z)| \leq M\|t_1 - t_2\|_E$ for all $t_1,\, t_2 \in T$, and for all $z \in \mathcal{Z}$.

1.3.1 Test Statistic

Define $J_n(\alpha,\, t) \equiv \dfrac{1}{n}\sum_{i=1}^{n} m(y_i,\, \theta,\, h(x_i)) \cdot g(t,\, z_i)$. $J_n(\alpha,\, t)$ is the sample average of the instrument function weighted moments evaluated at $(\alpha,\, t)$. We propose using the following test statistic for hypothesis testing:

$$S_n = \min_{\alpha \in \mathcal{A}_n \cap R} \left\{ \sup_{t \in T_n} n\left[J_n(\alpha,\, t)\right]' W_n(\alpha,\, t)\left[J_n(\alpha,\, t)\right] \right\} \tag{1.6}$$

where $\mathcal{A}_n = \Theta \times \mathcal{H}_n$ with $\mathcal{H}_n \uparrow \mathcal{H}$ as $n \to \infty$, $T_n \uparrow T$ as $n \to \infty$, and W_n is a weighting matrix.

S_n is the minimum of a Kolmogorov-Smirnov type statistic. A Cramér-von Mises (CvM) version of S_n, in which $\sup_{t \in T_n} ...$ is replaced by $\int_{T_n} ...dQ(t)$ with some distribution function $Q(\cdot)$, is an interesting potential alternative. We expect S_n and its CvM alternative to exhibit different power properties, but do not expect one to dominate the other. We leave the CvM alternative for future research. As the expression $W_n(\alpha,\, t)$ suggests, we allow for W_n to be data dependent and/or $(\alpha,\, t)$-dependent. [4]

[4] An example of weighting matrices that depend both on the data and the choice of the parameter value is the continuous-updating type weighting matrices.

We impose the following assumption on the weighting matrix W_n:

Assumption 1.3.2. (the weighting matrix) (i) $W_n(\alpha, t)$ is positive semi-definite; (ii) There is a positive definite $W(\alpha, t)$ such that: a) $\underline{W} \leq W(\alpha, t) \leq \overline{W}$ for all $(\alpha, t) \in \mathcal{A} \times T$ for some positive definite matrices $\underline{W}$ and $\overline{W}$;[5] and (b) $\sup_{\mathcal{A}\times T} |W_n(\alpha, t) - W(\alpha, t)| = O\left(n^{-1/2}\right)$ in a per-element sense.

Assumption 1.3.2 (i) is a natural requirement for weighting matrices. Assumption 1.3.2 (ii) requires that the weighing matrix $W_n(\alpha, t)$ converges at the $1/\sqrt{n}$ rate to a limiting matrix $W(\alpha, t)$ that is bounded from above and bounded away from zero.

1.3.2 Definitions

Our main interest is the asymptotic behavior of the test statistic S_n. In this subsection, we introduce several important definitions that are needed for developing the asymptotic results.

Define [6]

$$\hat{\alpha}_n \in \underset{\alpha\in\mathcal{A}_n\cap R}{\arg\min} \left\{ \sup_{t\in T} n\left[J_n(\alpha, t)\right]' W_n(\alpha, t)\left[J_n(\alpha, t)\right] \right\}. \tag{1.7}$$

as the minimizer of the objective function $\sup_{t\in T} n\left[J_n(\alpha, t)\right]' W_n(\alpha, t)\left[J_n(\alpha, t)\right]$.

[5] For two positive semi-definite matrices A and B, we say $A \geq B$ if $A - B$ is positive semi-definite.

[6] When the "$\arg\min$" is not well defined, $\hat{\alpha}_n$ is defined as $Q_n(\hat{\alpha}_n) \leq \inf_{\alpha\in\mathcal{A}_n} Q_n(\alpha) + o_p(1)$ where $Q_n(\alpha) \equiv \sup_{t\in T} n\left[J_n(\alpha, t)\right]' W_n(\alpha, t)\left[J_n(\alpha, t)\right]$.

Just like in many point-identification analyses, the asymptotic behavior of $\hat{\alpha}_n$ provides important insight on that of S_n within the partial identification framework. In particular, we find the distance between $\hat{\alpha}_n$ and the identification set $\mathcal{A}_I$ with respect to norms $\|\cdot\|_{L_2}$ and $\|\cdot\|_w$ (to be defined immediately) plays a crucial role in our analysis.

$\|\cdot\|_{L_2}$ is the $L2$ norm, defined as

$$\begin{aligned} \|\alpha\|_{L_2} &\equiv \|\theta\|_E + \|h\|_{L_2} \\ &= \|\theta\|_E + \sqrt{E\left[\left(h\left(X\right)\right)^2\right]}. \end{aligned}$$

And the distance between $\hat{\alpha}_n$ and $\mathcal{A}_I$ with respect to $\|\cdot\|_{L_2}$ is defined as

$$d_{L_2}\left(\hat{\alpha}_n, \mathcal{A}_I\right) \equiv \inf_{\alpha \in \mathcal{A}_I} \|\hat{\alpha}_n - \alpha\|_{L_2}.$$

The second norm $\|\cdot\|_w$ is a pseudo norm that is weaker than $\|\cdot\|_{L_2}$. To define $\|\cdot\|_w$, we need to introduce the following notion of functional derivatives:

Definition 1.3.1. (pathwise derivatives) (i) For any functional $f\left(\cdot\right): \mathcal{H} \to \mathbb{R}$, define

$$\frac{df\left(h_0\right)}{dh}\left[\Delta_h\right] = \frac{df\left(h_0 + \tau\Delta_h\right)}{d\tau} \mid_{\tau=0}$$

as the pathwise first derivative of f at h_0 in the direction of Δ_h;

(ii) Define

$$\frac{d^2 f\left(h_0\right)}{dh^2}\left[\Delta, \Delta\right] = \frac{d^2 f\left(h_0 + \tau\Delta\right)}{d\tau^2} \mid_{\tau=0}$$

as the pathwise second derivative of f at h_0 in the direction of Δ;

(iii) For any function $f(\cdot): \mathcal{A} \to \mathbb{R}$, define

$$\begin{aligned} \frac{df(\alpha_0)}{d\alpha}[\Delta] &= \frac{df(\alpha_0 + \tau\Delta)}{d\tau}\mid_{\tau=0} \\ &= \frac{\partial f(\theta_0, h_0)}{\partial \theta'}(\Delta_\theta) + \frac{df(\theta_0, h_0)}{dh}[\Delta_h] \end{aligned}$$

as the pathwise first derivative of f at $\alpha_0 = (\theta_0, h_0)$ in the direction of $\Delta = (\Delta_\theta, \Delta_h)$.

Definition 1.3.2. (pseudo norm $\|\cdot\|_w$) Denote by $\rho(\alpha, t) \equiv E[m(Y, \theta, h(X)) \cdot g(t, Z)]$. Define

$$\|\tilde{\alpha}\|_w = \sup_{\alpha_0 \in \mathcal{A}_I} \sup_{t \in T} \| \frac{d\rho(\alpha_0, t)}{d\alpha}[\tilde{\alpha}] \|_E.$$

And the distance between $\hat{\alpha}_n$ and $\mathcal{A}_I$ with respect to $\|\cdot\|_w$ is defined as

$$d_w(\hat{\alpha}_n, \mathcal{A}_I) \equiv \inf_{\alpha \in \mathcal{A}_I} \|\hat{\alpha}_n - \alpha\|_w.$$

We derive the convergence rates of $d_{L_2}(\hat{\alpha}_n, \mathcal{A}_I)$ and $d_w(\hat{\alpha}_n, \mathcal{A}_I)$, and summarize the results in Lemma 2 and Lemma 3 in Appendix A. These results are used in deriving the asymptotic distribution of our test statistic. Notably, they parallel the convergence rates results under point-identification in Chen and Pouzo (2011).

Remark 1.3.1. (i) According to Definition 1.3.1, in the direction of $\Delta = (\Delta_\theta,\, \Delta_h)$

$$\begin{aligned} \frac{d\,\rho\,(\alpha_0,\, t)}{d\alpha}\,[\Delta] &= \frac{d\,\rho\,(\alpha_0 + \tau\Delta,\, t)}{d\tau}\,|_{\tau=0} \\ &= \frac{\partial\rho\,(\alpha_0,\, t)}{\partial\theta'}\,(\Delta_\theta) + \frac{d\,\rho\,(\alpha_0,\, t)}{dh}\,[\Delta_h]\,. \end{aligned}$$

(ii) To better understand Definition 1.3.1, it is helpful to consider examples.

(a) For a partially linear IV where

$$\rho\,(\theta,\, h,\, t) = E\left[\left(Y - X_1'\theta - h\,(X_2)\right)\cdot g\,(t,\, Z)\right],$$

in the direction of $(\Delta_\theta,\, \Delta_h)$, its pathwise first derivative at α_0 is

$$-E\left[\left(X_1'\Delta_\theta + \Delta_h\,(X_2)\right)\cdot g\,(t,\, Z)\right], \tag{1.8}$$

and its pathwise second derivative is 0;

(b) For a partially linear quantile IV where

$$\rho\,(\theta,\, h,\, t) = E\left[\left[1\left(Y - X_1'\theta - h\,(X_2) \le 0\right) - \tau\right]\cdot g\,(t,\, Z)\right],$$

in the direction of $(\Delta_\theta,\, \Delta_h)$, its pathwise first derivative at α_0 is

$$E\left[f_{Y|X,Z}\left(X_1'\theta_0 + h_0\,(X_2)\right)\cdot\left(X_1'\Delta_\theta + \Delta_h\,(X_2)\right)\cdot g\,(t,\, Z)\right], \tag{1.9}$$

and its pathwise second derivative is

$$E\left[f'_{Y|X,Z}\left(X'_1\theta_0 + h_0\left(X_2\right)\right)\cdot\left(X'_1\Delta_\theta + \Delta_h\left(X_2\right)\right)^2\cdot g\left(t,\, Z\right)\right]. \tag{1.10}$$

(*iii*) (a) In the point-identified case, $\|\alpha\|_w = \max_{t\in T}\|\frac{d\,\rho\left(\alpha_0,\, t\right)}{d\alpha}\left[\alpha\right]\|_E$. Compared with

$$\|\alpha\|_{0w} = \sqrt{E\left\{\|\frac{dE\left(m\left(Y,\, \alpha_0\right)|Z\right)}{d\alpha}\left[\alpha\right]\|_E^2\right\}}$$

which is the standard weak norm considered in the literature (see, for example, Ai and Chen (2003) and Chen and Pouzo (2011)),

$$\begin{aligned}
\|\alpha\|_w &= \sup_{t\in T}\|\frac{dE\left[m(Y,\, \alpha_0)\cdot g\left(t,\, Z\right)\right]}{d\alpha}\left[\alpha\right]\|_E \\
&= \sup_{t\in T}\|E\left\{\left[\frac{dE\left[m(Y,\, \alpha_0)|Z\right]}{d\alpha}\left[\alpha\right]\right]\cdot g\left(t,\, Z\right)\right\}\|_E \\
&\leq C\cdot\|E\left[\frac{dE\left[m(Y,\, \alpha_0)|Z\right]}{d\alpha}\left[\alpha\right]\right]\|_E \\
&\leq C\cdot\|\alpha\|_{0w}\,.
\end{aligned}$$

Therefore, $\|\cdot\|_w \preceq \|\cdot\|_{0w}$;

(b) When $m\left(y,\, h\left(x\right)\right) = y - h\left(x\right)$, the $\|h\|_w = \max_{t\in T}|E\left[h\left(X\right)g\left(Z,\, t\right)\right]|$ which coincides with Santos (2011). ■

We introduce the following measure of ill-posedness to account for the ill-posed problem that may arise in our semiparametric specification:

Definition 1.3.3. (sieve measure of ill-posedness) The sieve measure of ill-posedness is defined as

$$\psi_n = \sup_{\alpha \in \mathcal{A}_n : \inf_{\alpha_0 \in \mathcal{A}_I} \|\alpha - P_n(\alpha_0)\|_w \neq 0} \left(\frac{\inf_{\alpha_0 \in \mathcal{A}_0} \|\alpha - P_n(\alpha_0)\|_{L2}}{\inf_{\alpha_0 \in \mathcal{A}_0} \|\alpha - P_n(\alpha_0)\|_w} \right),$$

where $P_n(\alpha_0)$ is the projection of α_0 onto $\mathcal{A}_n$.[7]

Under point-identification, the sieve measure of ill-posedness in Definition 1.3.3 is similar to the one defined in Chen and Pouzo (2011).

Finally, we introduce two families of functions, $V_{k_n}^{\alpha_0}$ and $V_{\infty}^{\alpha_0}$. As it will be shown, $V_{\infty}^{\alpha_0}$ appears in the asymptotic null distribution of S_n. $\left\{V_{k_n}^{\alpha_0}\right\}$ forms an increasing series of sets with $V_{\infty}^{\alpha_0}$ being the limiting set.

Definition 1.3.4. Let $\mathcal{H}_{nR} = \left\{(r_\theta, r_h) \in \mathbb{R}^{d_\theta + k_n} : L\left((r_\theta, r_h' p^{k_n})\right) = 0\right\}$ in which $p^{k_n} \equiv (p_1, ..., p_{k_n})'$ is the vector of basis functions for the sieve space $\mathcal{H}_n$.

(i) Define the following family of functions:

$$V_{k_n}^{\alpha_0} = \left\{ v : T \to \mathbb{R}^{d_m} \, s.t. \, v(t) = \frac{\partial \rho(\theta_0, h_0, t)}{\partial \theta'} \cdot r_\theta + \frac{d\rho(\theta_0, h_0, t)}{dh} \left[p^{k_n}\right]' r_h, \, (r_\theta, r_h) \in \mathcal{H}_{nR} \right\}$$

[7] More specifically, $P_n(\alpha_0) = P_n((\theta_0, h_0)) = (\theta_0, P_n(h_0))$ where $P_n(h_0)$ is the projection of h_0 onto $\mathcal{H}_n$.

for each $\alpha_0 \in \mathcal{A}_I$, where $\frac{d\,\rho(\theta_0,\, h_0,\, t)}{d\,h}\left[p^{k_n}\right]'$ is a $d_m \times k_n$ matrix defined as

$$\left(\frac{d\,\rho(\theta_0,\, h_0,\, t)}{d\,h}[p_1],\, ...,\, \frac{d\,\rho(\theta_0,\, h_0,\, t)}{d\,h}[p_{k_n}]\right);$$

(ii) Define $V_\infty^{\alpha_0}$ as the closure of $\cup_{n=1}^{\infty} V_{k_n}^{\alpha_0}$ under the supreme norm $||\cdot||_\infty$ for each $\alpha_0 \in \mathcal{A}_0$.

In words, $V_\infty^{\alpha_0}$ consists of all possible first order deviations (in the sense of the pathwise derivatives) of the population moment $E[m(Y,\, \cdot) \cdot g(t,\, Z)]$ (without violating the given restriction $L(\alpha) = a$) from its value at α_0. Note that $\mathcal{H}_{nR}$ forms a linear subspace in $\mathbb{R}^{d_\theta + d_{k_n}}$ because of the linearity of $L(\cdot)$.

1.3.3 Asymptotic Results on S_n

Assumptions 1.2.1, 1.2.2, 1.3.1 and 1.3.2 state regularity conditions on the parameter space, the form of the restriction to be tested, the instrument functions, and the weighting matrix, respectively. In addition, we impose the following assumptions in order to obtain the asymptotic distribution of S_n.

Assumption 1.3.3. (i) $\{X_i,\, Y_i,\, Z_i\}_{i=1}^n$ is i.i.d. with $X \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$, $Y \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}$, $Z \in \mathcal{Z} \subseteq \mathbb{R}^{d_z}$, and $\mathcal{X}$ is bounded; (ii) $\{m(\cdot,\, \alpha) : \alpha = (\theta,\, h) \in \mathcal{A}\}$ is a uniformly bounded *Donsker class;* (iii) In a $\|\cdot\|_{L_2}$-neighborhood of $\mathcal{A}_I$ defined as $B(\mathcal{A}_I,\, \xi) \equiv \{\alpha \in \mathcal{A} :\, d_{L_2}(\alpha,\, \mathcal{A}_I) \leq \xi\}$ for some $\xi > 0$, $\rho(\alpha,\, t)$ is pathwise differentiable with re-

spect to α; (iv) $\exists$ positive constants c_1, c_2 and δ s.t. $c_1\left(\inf_{\alpha_0\in\mathcal{A}_I}\|\alpha-\alpha_0\|_w\wedge\delta\right)\leq \max_{t\in T}\|\rho(\alpha,\,t)\|_E\leq c_2\inf_{\alpha_0\in\mathcal{A}_I}\|\alpha-\alpha_0\|_w$ for all $\alpha\in\mathcal{A}$; (v) $\psi_n=o\left(n^{1/4}\right)$.

Assumption 1.3.4. (i) The dimension of the sieve space $\mathcal{H}_n$, denoted by k_n, satisfies $k_n<\infty$, $k_n\to\infty$ and $\frac{k_n}{n}\to 0$ as $n\to\infty$; (ii) The eigenvalues of $E\left[p^{k_n}(X)\,p^{k'_n}(X)\right]$ are bounded, uniformly on n, i.e. denote by τ_n the largest eigenvalue of $E\left[p^{k_n}(X)\,p^{k'_n}(X)\right]$, then there is $J>0$ s.t. $\tau_n\leq J$ for all n;

Assumption 1.3.5. There exists $\Pi_n h\in\mathcal{H}_n$ for each $h\in\mathcal{H}$ such that: (i) $||\Pi_n h-h||_{L^2}=o(1)$ for all $h\in\mathcal{H}$; (ii) $\delta_{w,n}\equiv\sup_{h_0\in\mathcal{H}_I}\|h_0-\Pi_n h_0\|_w=o\left(n^{-1/2}\right)$ and $\delta_{s,n}\equiv\sup_{h_0\in\mathcal{H}_I}\|h_0-\Pi_n h_0\|_{L2}=o\left(n^{-1/4}\right)$; (iii) For some $\gamma_n=o(1)$ s.t. $\gamma_n\sqrt{n}\to\infty$, $\sup_{h\in\mathcal{H}_I}||\Pi_n h||_s\leq B-\gamma_n$; (iv) There is $\Pi_n t\in T_n$ for each $t\in T$ such that $\sup_{t\in T}\|t-\Pi_n t\|_E=o\left(n^{-1/2}\right)$.

Assumption 1.3.6. (i) In a $\|\cdot\|_{L_2}$-neighborhood of $\mathcal{A}_I$, $\rho(\alpha,\,t)$ is pathwise twice differentiable with respect to α; (ii) Define the pathwise sieve Hessian matrix of $\rho(\alpha,\,t)$ as a $k_n\times k_n$ matrix $\Psi_n(\alpha,\,t)$ whose ijth element is $\frac{d^2\,\rho(\alpha,\,t)}{dh^2}[p_i,\,p_j]$. The eigenvalues of $\Psi_n(\alpha,\,t)$ are uniformly bounded on $\mathcal{A}_I\times T$; (iii) The eigenvalues of the $d_\theta\times d_\theta$ Hessian matrix $\frac{\partial^2\,\rho(\alpha,\,t)}{\partial\theta^2}$ are uniformly bounded on $\mathcal{A}_I\times T$.

Remark 1.3.2. Given the compactness of $\mathcal{A}$, Assumption 1.3.3 (ii) is usually satisfied. For example, it is satisfied by the linear moment functions in the partially linear IV

as shown in Santos (2011). Assumption 1.3.3 (ii) is also satisfied if $m(\cdot)$ is pointwise Lipschitz continuous in α and continuous in y, and $\mathcal{X} \times \mathcal{Y}$ is compact, which is shown in Appendix B. For quantile IV regressions where m is an indicator function, Assumption 1.3.3 (ii) is satisfied if we restrict the variability of the unknown function $h(\cdot)$ such that $\{\{(x, y) : y \leq h(x)\} : h \in \mathcal{H}\}$ forms a *Vapnik–Chervonenkis* class of sets. This can be done, for example, by requiring the number of time for which the second order derivative of $h(\cdot)$ switches signs (between ≥ 0 and < 0) to be bounded by a finite integer. Assumption 1.3.3 (iii) implies that the pseudo-metric $\|\cdot\|_w$ is well defined in a neighborhood of the $\mathcal{A}_I$. This condition trivially holds for partially linear IV as shown by formula (1.8), and also holds for partially linear quantile IV as long as the probability density $f_{Y|X,Z}(\cdot)$ is well defined as shown by formula (1.9). Assumption 1.3.3 (iv) implies that the pseudo weak distance $d_w(\alpha, \mathcal{A}_I) \equiv \inf_{\alpha_0 \in \mathcal{A}_I} \|\alpha - \alpha_0\|_w$ is Lipschitz continuous with respect to the population criterion. Similar assumptions are commonly imposed in the semiparametric literature, see, for example Chen and Pouzo (2011). In essence, Assumption 1.3.3 (iv) is analogous to the full rank and continuity of the Jacobian (near the true value) assumption made for parametric moment restrictions in the point-identified case. Assumption 1.3.3 (v) allows for the measure of ill-posedness ψ_n to grow to ∞ but requires that the growth rate is slower than $n^{1/4}$. Assumption 1.3.4 (i) requires that we use finite dimensional linear sieve space. Assumption 1.3.5 (iii) requires that any element of $\mathcal{H}_I$ can be approximated well by some sieve space element that is in the interior of $\mathcal{H}$. It is worth noticing that Assumption 1.3.5 (iii) does not rule out scenarios in which some elements of $\mathcal{H}_I$ are on the boundary – it only requires that

each element of $\mathcal{H}_I$ can be approximated well by some element of the sieve space that is in the interior of $\mathcal{H}$. Assumption 1.3.4 and Assumption 1.3.5 are satisfied by many commonly used sieves such as polynomials, P-splines and B-splines. Assumption 1.3.6 (i) is a stronger restriction than Assumption 1.3.3 (ii). It holds trivially for partially linear IV because the pathwise second derivative in such a model is always zero as shown in Remark 1.3.1 (ii). Assumption 1.3.6 (i) also holds for partially linear quantile IV as long as $f_{Y|X,Z}(\cdot)$ is differentiable as shown by formula (1.10). Assumption 1.3.6 (ii) and (iii) together implies that $|\frac{d^2 \rho(\alpha, t)}{d\alpha^2}[\Delta, \Delta]| \leq D \cdot \|\triangle\|_{L2}^2$ for some $D > 0$ for all $(h, \Delta, t) \in \mathcal{H} \times \mathcal{H} \times T$. Assumption 1.3.6 (ii) and (iii) are needed to control the second order term in our asymptotic analysis. They hold trivially for partially linear IV, and hold for locally linear quantile IV as long as the first derivative of $f_{Y|X,Z}(\cdot)$ exists and is bounded according to (1.10). We note that for vector-valued $m(\cdot)$, Assumption 1.3.6 (ii) and (iii) should be viewed as restrictions imposed on each element of $\rho(\alpha, t)$. ■

Now we present the main asymptotic results.

Theorem 1.3.1. Let Assumptions 1.2.1-1.2.2, 1.3.1-1.3.6 hold. If $\mathcal{A}_I \cap R \neq \varnothing$, we have

$$S_n \Rightarrow \inf_{\alpha \in \mathcal{A}_I \cap R} \inf_{v \in V_\infty^\alpha} \sup_{t \in T} \left\{ [G(\alpha, t) + v(t)]' W(\alpha, t) [G(\alpha, t) + v(t)] \right\}$$

where $G(\alpha, t)$ is a tight Gaussian process on $\mathcal{A} \times T$.

Theorem 1.3.2. Let Assumptions 1.2.1-1.2.2, 1.3.1-1.3.6 hold. We have

$$n^{-1}S_n \xrightarrow{a.s.} \min_{\alpha \in \mathcal{A} \cap R} \sup_{t \in T} \left\{ E\left[m\left(Y,\, \theta,\, h\left(X\right)\right) w\left(t,\, Z\right) \right]' W\left(\alpha,\, t\right) E\left[m\left(Y,\, \theta,\, h\left(X\right)\right) w\left(t,\, Z\right) \right] \right\}.$$

Theorem 1.3.1 shows that S_n converges in distribution to a tight distribution under the null. It is worth noticing that the limiting distribution is nonpivotal in the sense that it depends on the underlying data generating process through $\mathcal{A}_I$ and V_∞^α. Theorem 1.3.2 shows the asymptotic behavior of S_n under a fixed alternative. Let

$$C \equiv \min_{\alpha \in \mathcal{A} \cap R} \sup_{t \in T} \left\{ E\left[m\left(Y,\, \theta,\, h\left(X\right)\right) w\left(t,\, Z\right) \right]' W\left(\alpha,\, t\right) E\left[m\left(Y,\, \theta,\, h\left(X\right)\right) w\left(t,\, Z\right) \right] \right\}. \quad (1.11)$$

When $\mathcal{A}_I \cap R = \varnothing$, for each $\alpha \in \mathcal{A} \cap R$ there exists some $t \in T$ such that $E\left[m\left(Y,\, \theta,\, h\left(X\right)\right) w\left(t,\, Z\right) \right] \neq 0$. Consequently, $C > 0$ due to the compactness of $\mathcal{A} \cap R$. Therefore, Theorem 1.3.2 implies that S_n goes to ∞ almost surely at the rate $O\left(n\right)$ when H_0 is false. Theorems 1.3.1 and 1.3.2 are the basis for our inference procedure.

1.3.4 Implementation and Critical Values

Denote by D the asymptotic null distribution of S_n. According to Theorem 3.1,

$$D = \inf_{\alpha \in \mathcal{A}_I \cap R} \inf_{v \in V_\infty^\alpha} \sup_{t \in T} \left\{ \left[G\left(\alpha,\, t\right) + v\left(t\right) \right]' W\left(\alpha,\, t\right) \left[G\left(\alpha,\, t\right) + v\left(t\right) \right] \right\}. \quad (1.12)$$

For a targeted nominal size ρ, using D's $(1-\rho)$th quantile as the asymptotic critical value seems to be the most natural way of implementing Theorem 1.3.1 and 1.3.2. However, the $(1-\rho)$th quantile is infeasible because we are unfamiliar with the distribution of D. We propose using critical values based on a bootstrap procedure that is specified below.

The Bootstrap Statistic

Recall the original test statistic is

$$S_n = \min_{\alpha \in \mathcal{A}_n \cap R} \left\{ \sup_{t \in T_n} n \left[J_n(\alpha, t) \right]' W_n(\alpha, t) \left[J_n(\alpha, t) \right] \right\}$$

where $J_n(\alpha, t) \equiv \frac{1}{n} \sum_{i=1}^{n} m(y_i, \theta, h(x_i)) \cdot g(t, z_i)$.

For a bootstrap sample $\{x_i^*, y_i^* z_i^*\}$, let

$$J_n^*(\alpha, t) \equiv \frac{1}{n} \sum_{i=1}^{n} \left[m(y_i^*, \theta, h(x_i^*)) \cdot g(t, z_i^*) - \frac{1}{n} \sum_{i=1}^{n} m(y_i, \theta, h(x_i)) g(t, z_i) \right]. \quad (1.13)$$

And let $W_n^*(\alpha, t)$ be the weighting matrix calculated using the bootstrap sample. We define the bootstrap version of the test statistic as

$$S_n^* = \min_{\alpha \in \mathcal{A}_n \cap R} \left\{ \sup_{t \in T_n} n \left[J_n^*(\alpha, t) \right]' W_n^*(\alpha, t) \left[J_n^*(\alpha, t) \right] + \lambda_n P_n(\alpha, t) \right\} \quad (1.14)$$

where $P_n(\alpha, t) \equiv \left[\frac{1}{n}\sum_{i=1}^{n} m(y_i, \theta, h(x_i))\, g(t, z_i)\right]^2$.

When setting $\lambda_n = 0$, our bootstrap procedure becomes the standard bootstrap procedure. However, we require that $\lambda_n > 0$ because the term J_n in S_n is only properly centered on $\mathcal{A}_I$ while its counterpart J_n^* in S_n^* is properly centered on the whole parameter space $\mathcal{A}$. By penalizing $\alpha \notin \mathcal{A}_I$, $P_n(\alpha, t)$ effectively ensures the minimization in S_n^* to be carried out asymptotically only on $\mathcal{A}_I \cap R$, just like S_n. In essence, the need for $P_n(\alpha, t)$ is due to the nonpivotality of the asymptotic null distribution. $\lambda_n \in \mathbb{R}$ is a tuning parameter that determines how much weight to put on $P_n(\alpha, t)$.

Assumption 1.3.7. $\lambda_n \to \infty$ and $\lambda_n = o(n/\log(\log(n)))$.

The following theorem summarizes the asymptotic properties for S_n^*:

Theorem 1.3.3. Let Assumptions 1.2.1-1.2.2, 1.3.1-1.3.7 hold. If $\mathcal{A}_I \cap R \neq \varnothing$, then there exists a tight distribution D' such that $S_n^* \Rightarrow D'$ and

$$Pr(D > a) \leq Pr\left(D' > a\right)$$

for any given $a \in \mathbb{R}$. In addition, If $\mathcal{A}_I \cap R = \varnothing$

$$\lambda_n^{-1} S_n^* \xrightarrow{a.s.} \min_{\alpha \in \mathcal{A} \cap R} \max_{t \in T} \left\{ E[m(Y, \theta, h(X))\, w(t, Z)]'\, W(\alpha, t)\, E[m(Y, \theta, h(X))\, w(t, Z)] \right\}.$$

Theorem 1.3.3. shows that, under the null, the bootstrap statistic S_n^* converges in

law to a tight distribution that stochastically dominates the asymptotic null distribution D (in a non-strict sense). When the null hypothesis is false, both S_n^* and S_n go to infinity almost sure, but S_n^* does so at a slower rate ($O(\lambda_n)$). Theorem 1.3.3 guarantees the consistency of the hypothesis test based on the bootstrap critical value described below.

Calculating the Critical Value

For a targeted nominal size ρ, the critical value is calculated as follows:

1. Generate B bootstrap samples from the original sample $\{x_i, y_i, z_i\}_{i=1}^n$;
2. From each bootstrap sample, calculate $S_{n,b}^*$ according to formula(1.14);
3. Calculate the critical value $C_n^*(1-\rho)$ as the $(1-\rho)$th percentile of the set of values $\left\{S_{n,b}^*\right\}_{b=1}^B$.

The null hypothesis $\mathrm{H}_0 : \mathcal{A}_I \neq \varnothing$ is rejected if $S_n > C_n^*(1-\rho)$.

We conclude by showing the consistency of the hypothesis test based on the bootstrap critical value $C_n^*(1-\rho)$:

Theorem 1.3.4. (consistency of the hypothesis test) Consider hypothesis test of $\mathrm{H}_0 : \mathcal{A}_I \cap R \neq \varnothing$ against $\mathrm{H}_1 : \mathcal{A}_I \cap R = \varnothing$. Let Assumptions 1.2.1-1.2.2, 1.3.1-1.3.7 hold.

Then the test has the following properties:

$$\begin{aligned} \limsup_{n\to\infty} Pr\left(S_n > C_n^*\left(1-\rho\right) \mid \mathrm{H}_0\right) &\leq \rho; \\ \lim_{n\to\infty} Pr\left(S_n \leq C_n^*\left(1-\rho\right) \mid \mathrm{H}_1\right) &= 0. \end{aligned}$$

Clearly, Theorem 1.3.4 shows that the probability of type I error is asymptotically no larger than the nominal size ρ, and the probability of type II error is asymptotically 0.

Model Specification Test

In the specification test for conditional moment models, we test whether the conditional moment restriction is satisfied by some element of the parameter space. It fits in to our testing framework as a special case in which the restricted set $R = \mathcal{A}$.[8]

Formally, in the specification test, we consider the following H_0 and H_1:

$$\begin{aligned} \mathrm{H}_0: &\quad \mathcal{A}_I \neq \varnothing; \\ \mathrm{H}_1: &\quad \mathcal{A}_I = \varnothing, \end{aligned}$$

Consequently, the test statistic in this case becomes

[8] Recall $R = \{\alpha \in \mathcal{A}: L(\alpha) = a\}$. When $L(\cdot) \equiv 0$ which is clearly a bounded linear operator and $a = 0$, $R = \mathcal{A}$.

$$S_n = \min_{\alpha \in \mathcal{A}_n} \left\{ \sup_{t \in T_n} n \left[J_n(\alpha, t) \right]' W_n(\alpha, t) \left[J_n(\alpha, t) \right] \right\}.$$

And the corresponding bootstrap statistic becomes

$$S_n^* = \min_{\alpha \in \mathcal{A}_n} \left\{ \sup_{t \in T_n} n \left[J_n^*(\alpha, t) \right]' W_n^*(\alpha, t) \left[J_n^*(\alpha, t) \right] + \lambda_n P_n(\alpha, t) \right\}.$$

Following the general procedure developed in this section, we calculate the bootstrap critical value $C_n^*(1-\rho)$ for a targeted size ρ and reject the null hypothesis that the model is correctly specified if $S_n > C_n^*(1-\rho)$.

It follows directly from Theorem 3.4 that the specification test above is consistent. This result is reported as the following corollary:

Corollary 1.3.1. (consistency of the specification test) Let Assumptions 1.2.1-1.2.2, 1.3.1-1.3.7 hold. The specification test is consistent against fixed alternative in the sense that

$$\begin{aligned} \limsup_{n \to \infty} Pr\left(\text{Type I Error}\right) &\leq \rho; \\ \lim_{n \to \infty} Pr\left(\text{Type II Error}\right) &= 0. \end{aligned}$$

1.4 Confidence Sets

In the previous section, we developed a consistent procedure for hypothesis testing in conditional moment models of the form

$$E\left[m\left(Y,\,\theta,\,h\left(X\right)\right)|Z\right]=0$$

where the parameter $\alpha=(\theta,\,h)$ is possibly partially identified. As previously mentioned, confidence sets for θ can be built by inverting the test for the restriction $\theta=\bar{\theta}$. And confidence sets for $h\left(\cdot\right)$ at a given point x_0 can be built by inverting the test for $h\left(x_0\right)=\mu$. In this section, we show exactly how. And we study the asymptotic properties of the resulting confidence sets.

Although we only consider in this paper the two types of confidence sets above, confidence sets for other features of the parameter can be constructed by similarly inverting the corresponding test. For example, we can build confidence sets for the whole h by inverting the test for $h=\bar{h}$ where $\bar{h}$ is any given element of $\mathcal{H}$.

1.4.1 Confidence Sets for θ

To construct confidence sets for θ, we start with considering testing the restriction $\theta=\bar{\theta}$ for a given constant $\bar{\theta}\in\Theta$. This test fits into our framework as a special case where $L\left(\cdot\right)$ is a projection operator such that $L\left((\theta,\,h)\right)=\theta$.

Testing $\theta = \bar{\theta}$

When testing the restriction $\theta = \bar{\theta}$, $\mathcal{A} \cap R = \left\{\left(\bar{\theta},\, h\right) :\, h \in \mathcal{H}\right\}$. In this case, the test statistic becomes

$$S_n\left(\bar{\theta}\right) = \min_{h \in \mathcal{H}_n} \left\{ \sup_{t \in T_n} n \left[J_n\left(\left(\bar{\theta},\, h\right),\, t\right)\right]' W_n\left(\left(\bar{\theta},\, h\right),\, t\right) \left[J_n\left(\left(\bar{\theta},\, h\right),\, t\right)\right] \right\}. \tag{1.15}$$

And the corresponding bootstrap statistic and critical value are

$$S_n^*\left(\bar{\theta}\right) = \min_{h \in \mathcal{H}_n} \left\{ \sup_{t \in T_n} n \left[J_n^*\left(\left(\bar{\theta},\, h\right),\, t\right)\right]' W_n^*\left(\left(\bar{\theta},\, h\right),\, t\right) \left[J_n^*\left(\left(\bar{\theta},\, h\right),\, t\right)\right] + \lambda_n P_n\left(\left(\bar{\theta},\, h\right),\, t\right) \right\} \tag{1.16}$$

and $C_n^*\left(1-\rho,\, \bar{\theta}\right)$, respectively.

Notice in (1.15) and (1.16) how the minimization are carried out: different values of h are searched on while θ is fixed at $\bar{\theta}$. Therefore both $S_n\left(\bar{\theta}\right)$ and $S_n^*\left(\bar{\theta}\right)$ can be viewed as a profiled minimum. We write S_n, S_n^* and $C_n^*\left(1-\rho\right)$ as functions of $\bar{\theta}$ to emphasize their dependence on the conjectured value $\bar{\theta}$.

The Confidence Sets

We construct confidence sets for θ by inverting the above test. For a targeted nominal level $1-\rho$, the confidence set is constructed as

$$\mathrm{CS}_n\left(1-\rho\right) = \left\{\bar{\theta} \in \Theta :\, S_n\left(\bar{\theta}\right) \leq C_n^*\left(1-\rho,\, \bar{\theta}\right)\right\}. \tag{1.17}$$

In words, the confidence set consists of all values of $\bar{\theta}$ at which the corresponding test fails to reject the null.

Theorem 1.3.4 guarantees the resulting confidence sets are consistent in the following sense:

$$\inf_{\theta \in \Theta_I} \liminf_{n \to \infty} Pr\left(\theta \in \mathbf{CS}_n\left(1-\rho\right)\right) \geq 1-\rho. \tag{1.18}$$

We are able show that another notion of consistency, which is stronger than (1.18), holds for the confidence sets without additional assumptions:

Theorem 1.4.1. Let Assumptions 1.2.1-1.2.2, 1.3.1-1.3.7 hold. Then:

$$\liminf_{n \to \infty} \inf_{\theta \in \Theta_I} Pr\left(\theta \in \mathbf{CS}_n\left(1-\rho\right)\right) \geq 1-\rho \tag{1.19}$$

and $\lim_{n \to \infty} Pr\left(\theta \in \mathbf{CS}_n\left(1-\rho\right)\right) = 0$ for any $\theta \notin \Theta_I$.

Compared with (1.18), the infimum in (1.19) is carried out before the limit is taken. Therefore, Theorem 1.4.1 shows that the confidence sets have correct asymptotic coverage probability uniformly over the identification set Θ_I for a fixed data generating process (DGP). We note that our confidence sets do not necessarily achieve the uniformity over drifting sequences of data generating processes as in Andrews and Soares (2010). Theorem 1.4.1 also shows that the confidence sets are consistent against fixed alternatives.

1.4.2 Confidence Sets for $h(x_0)$

We construct confidence sets for $h(\cdot)$ at a given point, say x_0, by inverting the test for the restriction $h(x_0) = \mu$ where μ is constant in the range of $h(\cdot)$.

Testing $h(x_0) = \mu$

When testing the restriction $h(x_0) = \mu$, the test statistic becomes

$$S_n(\mu) = \min_{\alpha \in \mathcal{A}_n : h(x_0) = \mu} \left\{ \sup_{t \in T_n} n \left[J_n(\alpha, t) \right]' W_n(\alpha, t) \left[J_n(\alpha, t) \right] \right\}. \tag{1.20}$$

And the corresponding bootstrap statistic and critical value are

$$S_n^*(\mu) = \min_{\alpha \in \mathcal{A}_n : h(x_0) = \mu} \left\{ \sup_{t \in T_n} n \left[J_n^*(\alpha, t) \right]' W_n^*(\alpha, t) \left[J_n^*(\alpha, t) \right] + \lambda_n P_n(\alpha, t) \right\} \tag{1.21}$$

and $C_n^*(1-\rho, \mu)$, respectively. We write S_n, S_n^* and $C_n^*(1-\rho)$ as functions of μ to emphasize their dependence on the conjectured value μ.

The Confidence Sets

For a targeted nominal level $1-\rho$, the confidence set from inverting the above test is constructed as

$$\mathrm{CS}_n(1-\rho) = \{ l : \; S_n(l) \leq C_n^*(1-\rho, \mu) \}.$$

Define $\mathcal{H}_{I\,x_0} \equiv \{\mu : h(x_0) = \mu \text{ for some } h \in \mathcal{H}_I\}$ as the identification set for $h(x_0)$. The following theorem shows the consistency of the confidence sets:

Theorem 1.4.2. Let Assumptions 1.2.1-1.2.2, 1.3.1-1.3.7 hold. Then the confidence sets are consistent in the sense that

$$\liminf_{n\to\infty} \inf_{\mu\in\mathcal{H}_{I\,x_0}} Pr\left(\mu \in \mathbf{CS}_n\left(1-\rho\right)\right) \geq 1-\rho \tag{1.22}$$

and $\lim_{n\to\infty} Pr\left(\mu \in \mathbf{CS}_n\left(1-\rho\right)\right) = 0$ for any $\mu \notin \mathcal{H}_{I\,x_0}$.

Similar to the result in Theorem 1.4.1, this notion of CS consistency implies that the correctness of asymptotic coverage probability holds uniformly over the identification set $\mathcal{H}_{I\,x_0}$.

1.5 Monte Carlo Simulations

In this section we study the finite sample performance of the proposed inference procedures using Monte Carlo simulations. We consider the following partially linear quantile IV:

$$E\left[1\left\{Y \leq X_1\theta + h\left(X_2\right)\right\}|Z\right] = 0.75. \tag{1.23}$$

We simulate the data from the following DGP:

$$\begin{aligned} Y = & \quad \sin(\pi X_2) + U, \\ U = & \quad \frac{1}{10}\left[E\left(X_2^2|Z\right) - X_2^2\right] - 0.75 + \varepsilon. \end{aligned} \tag{1.24}$$

where

$$\begin{aligned} Z &\sim \textbf{Uniform}\left[0.5,\, 1\right], \\ X_1 &\sim \textbf{Uniform}\left[0,\, 1\right] + Z, \\ X_2 &\sim Z \cdot \textbf{Uniform}\left[-1,\, 1\right], \\ \varepsilon &\sim \textbf{Uniform}\left[0,\, 1\right], \end{aligned} \tag{1.25}$$

Identification

For any $\lambda \in \mathbb{R}$, $(0,\, \sin(\pi x_2) + \lambda x_2)$ satisfies the conditional moment equality (1.23) because

$$\begin{aligned} & E\left[1\left\{Y \leq \sin(\pi X_2) + \lambda X_2\right\}|Z\right] \\ =& Pr\left(U - \lambda X_2 \leq 0 | Z\right) \\ =& Pr\left(\varepsilon \leq 0.75 + \lambda X_2 - \frac{1}{10}\left[E\left(X_2^2|Z\right) - X_2^2\right]|Z\right) \\ =& E\left[Pr\left(\varepsilon \leq 0.75 + \lambda X_2 - \frac{1}{10}\left[E\left(X_2^2|Z\right) - X_2^2\right]|X_2,\, Z\right)|Z\right] \\ =& E\left[0.75 + \lambda X_2 - \frac{1}{10}\left[E\left(X_2^2|Z\right) - X_2^2\right]|Z\right] \\ =& 0.75 + \lambda E\left(X_2|Z\right) \\ =& 0.75. \end{aligned} \tag{1.26}$$

Therefore, under the DGP specified in ((A.23)) and (1.25), the identification set $\mathcal{A}_I$ contains $(0,\, \sin(\pi x_2))$ and $(0,\, \sin(\pi x_2) + \lambda x_2)$ for $\lambda > 0$, and the parameter $(\theta_0,\, h_0)$ in model (1.23) is partially identified.

Equation (1.26) also shows clearly that zero is an element of θ's identification set Θ_I. A complete characterization of Θ_I has not been established analytically. But through computation, we find that specific none-zero values, such as 0.3, 0.5, and 1, do not belong to Θ_I.

$1-\rho$ Confidence Sets for θ

For the instrument functions, we select the family of indicator functions

$$g(t,\, z) = 1\{t \leq z\},\, t \in T = \mathcal{Z}\,(= [0.5,\, 1]),$$

and set T_n to be the grid $\{0.5,\, 0.55,\, 0.6,\, 0.65,\, 0.7,\, 0.75,\, 0.8,\, 0.85,\, 0.9,\, 0.95,\, 1\}$. The sieve space we use is a B-Splines of order 3 with knot $\{-1,\, -1,\, -1,\, 0,\, 1, 1, 1\}$, which implies $k_n = 4$. This sieve is able to provide a good approximation of $\sin(\pi x) + \lambda x$. The critical values are based on 100 bootstrap evaluations. In our study, we performed 100 Monte Carlo repetitions of 500 observations $(n = 500)$.

Nominal Level = 0.9				
	θ = 0	θ = 0.3	θ = 0.5	θ = 1
$\lambda_n = 0$	0.52	0.06	0.00	0.00
$\lambda_n = 22.36$	0.71	0.29	0.15	0.00
$\lambda_n = 63.00$	0.87	0.52	0.21	0.02
Nominal Level = 0.95				
	θ = 0	θ = 0.3	θ = 0.5	θ = 1
$\lambda_n = 0$	0.56	0.06	0.01	0.00
$\lambda_n = 22.36$	0.80	0.36	0.18	0.00
$\lambda_n = 63.00$	0.93	0.59	0.27	0.02

TABLE 1.1 COVERAGE PROBABILITY FOR θ, UNDER DIFFERENT CHOICES OF λ_n

For different values of θ (0, 0.3, 0.5, and 1), Table 1.1 reports the simulated probability of θ being covered by the CS's as a function of the targeted nominal level $1-\rho$ and the tuning parameter λ_n. We have the following observations: (i) When moving the value of θ away from the identification set Θ_I, from zero (a value in Θ_I) to values such as 0.3, 0.5, or 1 (values outside Θ_I), we see a significant decrease in the coverage probability for all three choices of λ_n. This is an indicator that the hypothesis test, by inverting which our CS's are built, holds good power against fixed alternatives in finite samples; (ii) The coverage probability is somewhat sensitive to the choice of λ_n. Overall, choosing $\lambda_n = 63.0$, which equals approximately $n^{2/3}$, gives good size control; (iii) The choice $\lambda_n = 0$, which corresponds to the standard bootstrap procedure, is not warranted by our theory. And, as shown in Table 1.1, setting $\lambda_n = 0$ leads to severe size distortion.

1.6 Empirical Illustration

An Engel curve for a consumer describes how their share of expenditures on a given good or service is related to their total expenditure level. An Engel curve for a population, or for the representative consumer of this population, then describes the consumption behavior of the population. Empirical study of Engel curves has been focusing on mean Engel curves

$$Y_i = h\left(X_i\right) + e_i, \;\; E\left(e_i \left| Z_i\right.\right) = 0 \tag{1.27}$$

which describe how the mean share of expenditure Y_i on a given good or service is related to the total expenditures X_i. When estimating or conducting inference on mean Engel curves, the total expenditures are often viewed as possibly endogenous since they may be correlated with variables that are likely captured in the error term e_i, such as expenditures on other goods or service, and preferences between current and future consumption (which in turn affects the saving rate). Total income is commonly employed as an instrumental variable such as in Blundell, Chen, and Kristensen (2007) and Santos (2011). As pointed out by Blundell et al. (2007), total income will be exogenous from consumption expenditures, and therefore is a valid instrumental variable, under the assumption that heterogeneity in earnings is uncorrelated with (or, more strictly, independent of) preferences over consumption expenditures.

Quantile Engel curves

$$Y_i = h(X_i) + e_i, \;\; Pr(e_i \leq 0 \,|Z_i) = \tau, \;\; 0 < \tau < 1 \tag{1.28}$$

describe how a given quantile level of Y_i is related to X_i. By using quantile Engel curves we are able to obtain distributional information that we are unable to learn from mean Engel curves. However, empirical study of quantile Engel curves is very limited. Chen and Pouzo (2009) estimate quantile Engel curves from the UK Family Expenditure Survey data under the point-identification assumption, while still using total income as the instrumental variable.

In this section, I apply the proposed inference methods to study the quantile IV Engel curves for gasoline in Brazil using the Pesquisa de Orcamentos Familiares 2002-2003 (POF) data. The same data set was used to study the mean IV Engel curves for gasoline by Santos (2011). The quantile IV Engel curves in (1.28) can be equivalently written as:

$$E[1\{Y_i \leq h(X_i)\} \,|Z_i] = \tau, \;\; 0 < \tau < 1 \tag{1.29}$$

where Y_i is the share of total nondurable expenditures of household i spent on gasoline, X_i is the log of total nondurable expenditures of household i, and Z_i is the log of total household income. The data set contains $n = 4994$ observations. For the instrument function $g(t, z)$, I pick $g(t, z) = \phi\left(\frac{t_1 - z}{t_2}\right)$ with grid $\{-0.8, -0.4, 0, 0.4, 0.8\} \times \{0.05, 0.2\}$ for (t_1, t_2) as in Santos (2011). I set $\lambda_n = 292.168$ which corresponds to $n^{2/3}$. A limited analysis of the robustness of the results to these choices is considered below

I consider the 50th quantile, and consequently set $\tau = 0.5$. A log linear specification (i.e., linear with respect to the log of total nondurable expenditure X) is commonly used to parametrize Engel curves in the literature, including papers such as Leser (1963), Prais and Houthakker (1955), and Working (1943). Quantile Engel curves under the log linear specification take the form $E\left[1\left\{Y_i \leq \alpha + X_i\beta\right\}|Z_i\right] = \tau$. The log linear specification can be tested through the hypothesis test:

$$\begin{aligned} \mathrm{H}_0: &\quad \mathcal{H}_I \cap R_L \neq \varnothing; \\ \mathrm{H}_1: &\quad \mathcal{H}_I \cap R_L = \varnothing, \end{aligned}$$

where $\mathcal{H}_I$ denotes the identification set of h in Model (1.29), and R_L is defined as

$$R_L = \left\{h \in \mathcal{H}: \, h\left(x\right) = \alpha + x\beta \text{ for some } (\alpha, \, \beta) \in \mathbb{R}^2\right\}.$$

The calculated test statistic yields a p-value of 0.357. Therefore, the log linear specification is not rejected at a significance level of either 0.1 or 0.05. Interestingly, a linear specification (i.e., linear with respect to the total nondurable expenditure, e^X) is rejected at any significance level (p-value = 0).

After failing to reject the log linear specification, I calculate the 95% confidence region for the intercept α and the slope β by inverting test of $(\alpha, \, \beta) = \left(\tilde{\alpha}, \, \tilde{\beta}\right)$ as discussed in Section 1.4. The resulting confidence region for $(\alpha, \, \beta)$[9] is depicted in Figure 1.1

[9]I parametrize the linear functions on the range of X, $[7.468, \, 12.441]$, by a B-Spline of order 2 with knots $\{7.468, \, 7.468, \, 12.441, \, 12.441\}$. I obtain the the joint confidence region for the two coefficients of the B-Spline through a grid search with grid $[0.012 : \, 0.005 : \, 0.212] \times [-0.024 : \, 0.005 : \, 0.176]$, after confirming that the confidence region is in the interior of $[0.012, \, 0.212] \times [-0.024, \, 0.176]$ through an initial grid search with grid $[-3 : \, 0.1 : \, 3] \times [-3 : \, 0.1 : \, 3]$. Then I obtain the confidence regions for $(\alpha, \, \beta)$ as depicted in Figure

(a). Projecting this confidence region onto the α axis and the β axis yields a confidence region of $[0.0660,\ 0.2461]$ for α and a confidence region of $[-0.0153,\ 0.0028]$ for β, respectively. Figure 1.2 (a) shows the pointwise 95% confidence bounds for the 50th quantile Engel curve under the log linear specification, obtained by plotting the maximum and minimum of $\alpha + x\beta$ over the confidence region for $(\alpha,\ \beta)$ for each $x \in [7,\ 13]$.

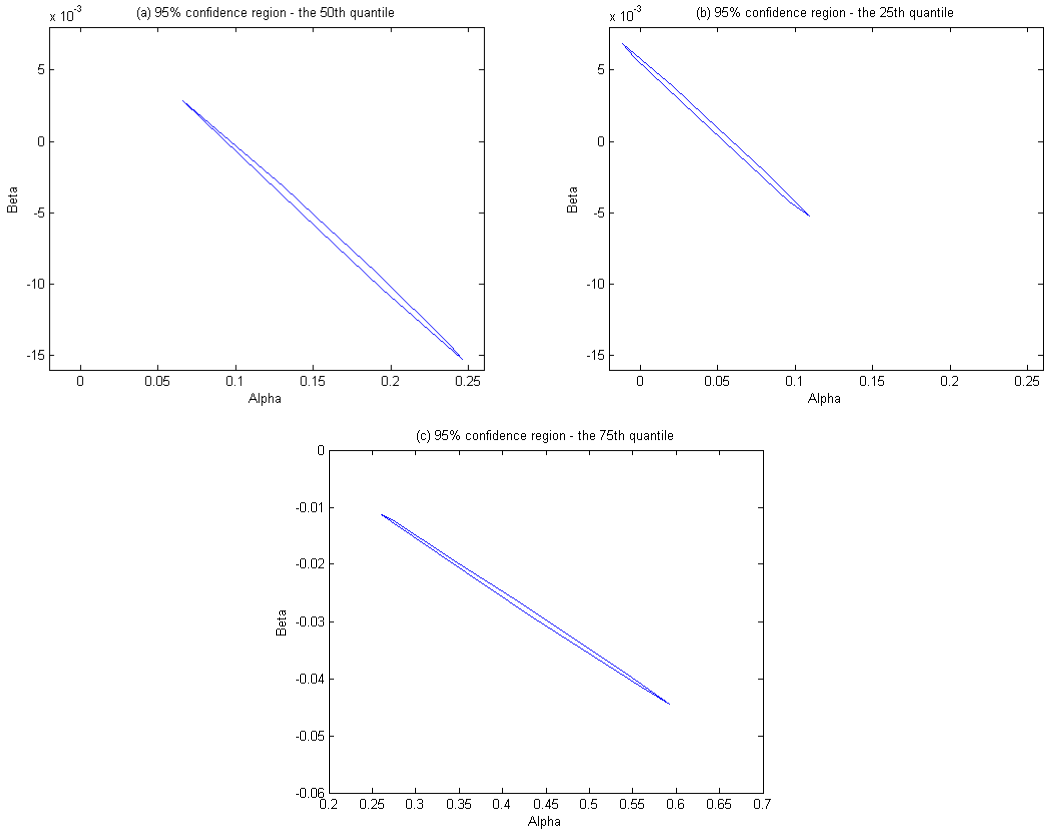


FIGURE 1.1 95% CONFIDENCE REGION FOR $(\alpha,\ \beta)$

1.1 (a) through linear transformation of the confidence region for the B-Spline coefficients.

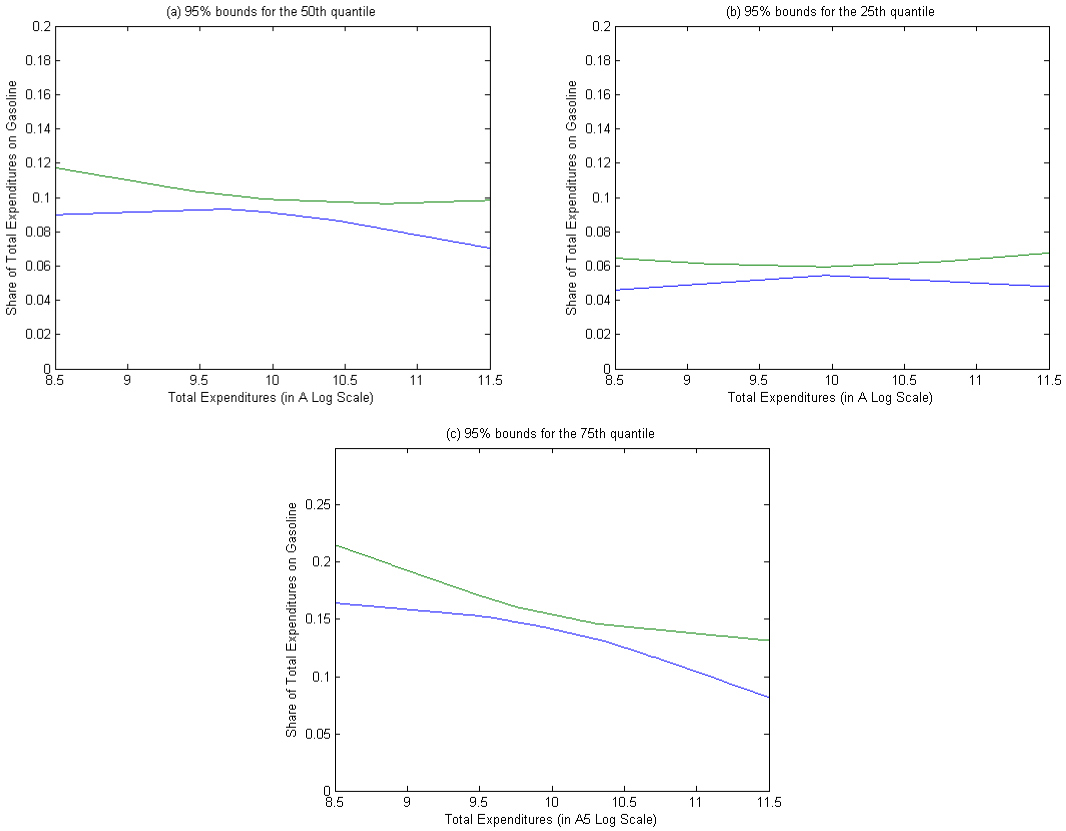


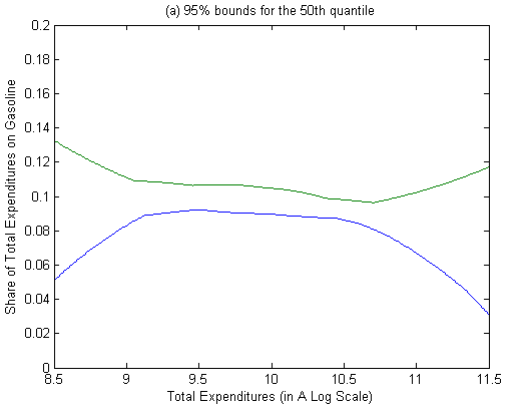
FIGURE 1.2 POINTWISE 95% CONFIDENCE BOUNDS UNDER LOG LINEAR SPECIFICATION

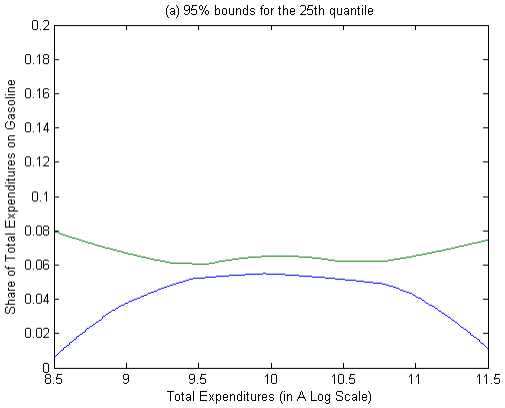
I also examine the 25th quantile ($\tau = 0.25$) and the 75th quantile ($\tau = 0.75$). For $\tau = 0.25$, a log linear specification for the corresponding Engel curve yields a p-value of 0.2467, and therefore is not rejected by the proposed test. Figure 1.1 (b) shows the 95% confidence region for (α, β). The projected confidence regions for α and β are $[-0.0111, 0.1090]$ and $[-0.0052, 0.0068]$, respectively. Yet, a linear specification is rejected at any significance level that is larger than 0.001(p-value = 0.001). Figure 1.2 (b) shows the pointwise 95% confidence bounds for the 25th quantile Engel curve under the log linear specification.

For $\tau = 0.75$, a log linear specification for the corresponding Engel curve yields a

p-value of 0.26. Figure 1.1 (c) shows the 95% confidence region for (α, β). The projected confidence regions for α and β are $[0.2601, 0.5928]$ and $[-0.0444, -0.0113]$, respectively. Yet, again, a linear specification is rejected at any significance level (p-value = 0). Figure 1.2 (c) shows the pointwise 95% confidence bounds for the 75th quantile Engel curve under the log linear specification.

In addition, I examine the quantile Engel curves using a log quadratic specification $\left[1\left\{Y_i \leq c_0 + c_1 X_i + c_2 X_i^2\right\} | Z_i\right] = \tau$, which yields p-values of 0.640, 0.760, and 0.990 for $\tau = 0.25, 0.5$ and 0.75, respectively. Figure 1.3 shows the pointwise 95% confidence bounds for the 25th, 50th and 75th quantile Engel curves under the log quadratic specification. Compared with their counterparts under the log linear specification depicted in Figure 1.2, the pointwise confidence bands in Figure 1.3 are slightly wider around the average level of total expenditures, while they are much wider at the high and low levels of total expenditures.





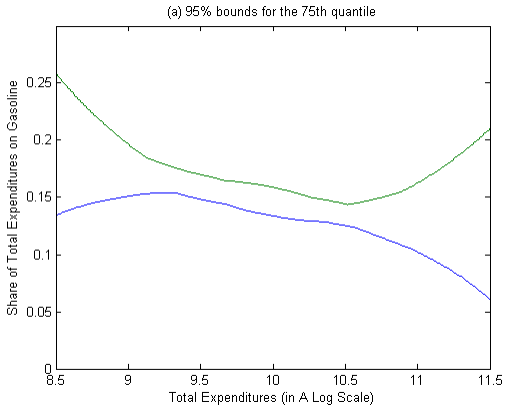


FIGURE 1.3 POINTWISE 95% CONFIDENCE BOUNDS UNDER LOG QUADRATIC SPECIFICATION

To check the robustness of the empirical results, I conduct inference on the median Engel curve ($\tau = 0.5$) with either the tuning parameter λ_n or the instrument function, or both, being altered to $\lambda_n = 70.668$, which corresponds to $n^{1/2}$, and $g_{\text{alt}}(t, z) = 1\{t \leq z\}$ with grid $\{3.8, 4.3, 3.8, ..., 12.8\}$. A summary of results from different combination of choices of λ_n and instrument function is reported as Table 1.2 below.

	Log Linear					*Linear*
	p-val.	*CI for α*	*CI for β*	*CI for h($\bar{X}$)*	*Santos(2011)*	*p-val.*
λ_n=$n^{2/3}$, g	0.357	[0.0660, 0.2461]	[-0.0153, 0.0028]	[0.0917, 0.0994]	[0.110, 0.116]	0
λ_n=$n^{1/2}$, g	0.293	[0.0660, 0.2461]	[-0.0153, 0.0028]	[0.0917, 0.0994]	[0.110, 0.116]	0
λ_n=$n^{2/3}$, g_{alt}	0.593	[0.1260, 0.2862]	[-0.0193, -0.0032]	[0.0941, 0.0971]	n/a	0
λ_n=$n^{1/2}$, g_{alt}	0.250	[0.1260, 0.2862]	[-0.0193, -0.0032]	[0.0941, 0.0971]	n/a	0

TABLE 1.2 P-VALUE AND 95% CONFIDENCE REGIONS FROM DIFFERENT λ_n AND INSTRUMENT FUNCTION

The second column of Table 1.2 lists the p-value for the log linear specification, associated with different choices of λ_n and instrument function. Although the p-values differ, the log linear specification is not rejected regardless of the choices of λ_n and instrument function. To the contrary, a linear specification yields a p-value of zero, and

therefore is rejected, under all four different choices of λ_n and instrument function, as shown in the last column.

The third and fourth columns of Table 1.2 show the corresponding confidence regions for the intercept and slope, respectively. To get a better sense of scale and magnitude of these confidence regions, I report in the fifth column the confidences region for the median share of gasoline expenditures at the sample average of the log total expenditures $\bar{X} = 9.9151$. As shown in the fifth column, all four different choices of λ_n and instrument function yield similar confidence regions for the median share at $\bar{X}$.

For comparison purposes, I also quote in the sixth column the confidence regions for the mean share of gasoline expenditures at $\bar{X}$ from Table 2 of Santos (2011). Comparing the sixth column with the fifth column shows a noticeable shift to the left when moving from the confidence regions for the mean to the confidence regions for the median. This observation suggests that the distribution of share of gasoline expenditures at the average total expenditure level in Brazil is likely to have a mean larger than the median, and consequently is likely to be skewed to the right.

The log quadratic specification yields slightly wider confidence regions for the median share at $\bar{X}$ as shown in Table 1.3.

	Log Linear	*Log Quadratic*
λ_n=$n^{2/3}$, g	[0.0917, 0.0994]	[0.0900, 0.1054]
λ_n=$n^{1/2}$, g	[0.0917, 0.0994]	[0.0900, 0.1054]
λ_n=$n^{2/3}$, g_{alt}	[0.0941, 0.0971]	[0.0804, 0.1154]
λ_n=$n^{1/2}$, g_{alt}	[0.0941, 0.0971]	[0.0804, 0.1154]

TABLE 1.3 CONFIDENCE REGIONS FOR $h\left(\bar{X}\right)$ UNDER TWO DIFFERENT SPECIFICATIONS

1.7 Conclusions

In this paper, I combine ideas and techniques from semiparametric modeling and partial identification analysis to develop general methods for inference in the conditional moment model $E\left[m\left(Y,\,\theta_0,\,h_0\left(X\right)\right)|Z\right]\,=\,0$. Without assuming point-identification, I propose a consistent procedure for testing the hypothesis that a given restriction on the parameter $\alpha\,=\,(\theta,\,h)$ is satisfied by at least one element of the identification set. Based on the proposed testing procedure, I show how to consistently test the conditional moment restriction specification, construct confidence sets for θ and for the unknown function $h\left(\cdot\right)$ at a given point. My methods are robust to partial identification. They extend the inference methods developed in Santos (2011) for nonparametric IV to more general conditional moment models with nonlinear moment functions.

Chapter 2

Estimation in Dynamic Discrete Choice Panel Data Models with Partial Identification

2.1 Introduction

Dynamic discrete choice panel data models arise in many econometric applications. Identification of the parameters in these models is delicate when lags of dependent variable and unobservable individual-specific effects are included in the explanatory variable. Under a random effects specification, the conditional distributions of the individual effects on the explanatory variables are assumed to belong to certain parametric family of functions. Although such an assumption helps to identify the parameter, point-identification can still fail, in particular when the process started prior to the initial period of observation which causes the initial conditions problem. The fixed effects approach is more robust in the sense that the assumptions it imposes on the distribution of the individual effects are less restrictive. However, as pointed out in Honoré

and Tamer (2006), the fixed effects approach is unsatisfactory for at least two reasons. First, it is not known how to do fixed effects estimation for many dynamic panel data models, and, even when it is known, the maintained assumptions are often very strong.

A number of recent papers have drawn attention to the partial identification issue in dynamic discrete choice panel data models. Notably, for the following autoregressive panel data model

$$y_{it} = \mathbf{1}\left\{y_{i\,t-1}\gamma + \alpha_i \geq \varepsilon_{it}\right\} \quad (t = 1, 2, ..., T), \tag{2.1}$$

Honoré and Tamer (2006) show that γ in model (2.1) is not point-identified when $T \leq 3$ and when $f(\alpha_i)$ and $f_0(y_{i0}|\alpha_i)$ are not fully specified, where $f(\cdot)$ the is the probability distribution function of α_i and $f_0(\cdot|\alpha_i)$ is the conditional distribution function of y_{i0} on α_i. They also show, through computational method, that the size of the identification set is small under their specifications, which suggests that partial identification is still informative. Therefore, it is desirable to develop estimation methods for the dynamic discrete choice panel data models that are robust to the possible lack of point-identification.

In this paper, I consider estimation in panel data models of the following form:

$$y_{it} = \mathbf{1}\left\{y_{i\,t-1}\gamma + x_{it}'\beta + \alpha_i \geq \varepsilon_{it}\right\} \quad (t = 1, 2, ..., T) \tag{2.2}$$

where (β, γ) is the parameter of interest, and both α_i and $y_{i\,0}$ are unobserved. Note that

model (2.1) is a special case of model (2.2) where $\beta_0 = 0$.

The major challenges facing estimation of model (2.2) are: 1). Point-identification often fails in model (2.2) as we just discussed; and 2). The heterogeneity cannot be differenced out by the standard "within" or first difference transformations due to non-linearity. I first show that the parameter can be equivalently defined by a finite number of conditional moment equalities. Then I propose using the Chernozhukov et al. (CHT) (2007) type set-estimators for estimating the identification set of (β, γ). The set-estimators are consistent in the sense that their Hausdorff distances to the identification set converge to zero in probability as sample size goes to infinity. Rates of convergence in the Hausdorff distance are derived. As it will be shown, the rates of convergence can be made arbitrarily close to $O_p\left(1/\sqrt{n}\right)$.

This paper is closely related to a growing literature on partial identification analysis, including papers such as Manski and Tamer (2002), Manski (2003), Imbens and Manski (2004), Chernozhukov et al. (2007), and many others. The paper is also closely related to Honoré and Kyriazidou (2000), Honoré and Tamer (2006), and many other papers in the large literature on panel data models.

The rest of the paper is organized as follows. Section 2.2 studies the identification in model (2.1). Section 2.3 constructs set estimators for the identification set, and derives all the consistency results. Section 2.4 extends the results developed in Sections 2.2 and 2.3 to the more general case of model (2.2). Section 2.5 studies finite sample performance by conducting Monte Carlo simulations. Section 2.6 concludes the paper.

The proofs of the theorems and lemmas stated in the paper are given in the Appendix C.

2.2 Identification

In Model (2.2),

$$y_{it} = \mathbf{1}\{y_{i\,t-1}\gamma + \alpha_i \geq \varepsilon_{it}\} \quad (t = 1, 2, ..., T),$$

each y_{it} takes value on $\{0, 1\}$. Consequently, the observable variable $y_i = (y_{i1}, y_{i2}, ..., y_{iT})$ takes value on a finite set, which is denoted by Ω. Hence, Ω is the sample space. Ω has $J = 2^T$ different elements, and can be described as $\Omega = \{A_j : j = 1, 2, ..., J\}$. Each A_j is a $T-$dimensional vector of zeros and ones. y_{i0} is not observable for any individual i. Assume that ε_{it} are jointly independent and that they are identically distributed over individuals with a known cdf F_t in the model. This assumption specifies that the conditional pmf of y_{it}, which depends only on $y_{i,t-1}$, α_i, and γ, has the form

$$f_t(y_{it}|y_{i\,t-1}, \alpha_i; \gamma) = y_{it}F_t\left(y_{i\,t-1}\gamma + \alpha_i\right) + (1 - y_{it})\left[1 - F_t\left(y_{i\,t-1}\gamma + \alpha_i\right)\right]. \tag{2.3}$$

Assume throughout this paper that it is known that α_i takes values on $\{a_m : m = 1,2,..,M\}$, but the associated probability masses $P_m := Pr\left(\alpha_i = a_m\right)$ are unknown.

For instance, when $T = 2$, the probability of $y_{i2} = 1$ given y_{i1}, α_i, andγ equals

$Pr\left(y_{i\,1}\gamma+\alpha_i\geq\varepsilon_{i2}\right)=Pr\left(\varepsilon_{i2}\leq y_{i\,1}\gamma+\alpha_i\right)=F_2\left(y_{i\,1}\gamma+\alpha_i\right)$. For $A_j=(0,1)$ $i.e.\,y_{i1}=0$ and $y_{i2}=1$, the model suggests that the probability of event $\{A_j\}$ happening should be

$$\begin{aligned}
\pi\left(A_j\right) &= Pr\left(y_{i1}=0,\,y_{i2}=1\right)\\
&= \sum_{m=1}^{M}Pr\left(y_{i0}=0,\,y_{i0}\gamma+\alpha_i<\varepsilon_{i1},\,y_{i1}\gamma+\alpha_i\geq\varepsilon_{i2}\,|\alpha_i=\alpha_m\right)P_m\\
&\quad+\sum_{m=1}^{M}Pr\left(y_{i0}=1,\,y_{i0}\gamma+\alpha_i<\varepsilon_{i1},\,y_{i1}\gamma+\alpha_i\geq\varepsilon_{i2}\,|\alpha_i=\alpha_m\right)P_m\\
&= \sum_{m=1}^{M}Pr\left(y_{i0}=0|\alpha_i=\alpha_m\right)Pr\left(\varepsilon_{i1}>\alpha_m\right)Pr\left(\varepsilon_{i2}\leq\alpha_m\right)P_m\\
&\quad+\sum_{m=1}^{M}Pr\left(y_{i0}=1|\alpha_i=\alpha_m\right)Pr\left(\varepsilon_{i1}>\gamma+\alpha_m\right)Pr\left(\varepsilon_{i2}\leq\alpha_m\right)P_m\\
&= \sum_{m=1}^{M}Pr\left(y_{i0}=0|\alpha_i=\alpha_m\right)\left[1-F_1\left(\alpha_m\right)\right]F_2\left(\alpha_m\right)P_m\\
&\quad+\sum_{m=1}^{M}Pr\left(y_{i0}=1|\alpha_i=\alpha_m\right)\left[1-F_1\left(\gamma+\alpha_m\right)\right]F_2\left(\alpha_m\right)P_m.
\end{aligned}$$

where the third equation follows from the independence assumption of ε_{it}.

In general, for $T>0$, and any given $A_j=(y_{i1},\,y_{i2},\,...,\,y_{iT})\in\Omega$, the induced probability of event $\{A_j\}$ happening is

$$\pi\left(A_j\right)=\sum_{l\in\{0,1\}}\left\{\sum_{m=1}^{M}\left[Pr(y_{i0}=l|\alpha_i=a_m)\prod_{t=1}^{T}f_t(y_{it}|y_{i\,t-1},\,\alpha_i=a_m;\,\gamma)\right]P_m\right\},\qquad(2.4)$$

where $Pr(y_{i0}=l|\alpha_i=a_m)$ is the probability of $y_{i0}=l$ conditional on $\alpha_i=a_m$, $l=0,\,1$. Here, $Pr(y_{i0}=l|\alpha_i=a_m)$ is not specified, as well as P_m. Set $\theta_m=Pr(y_{i0}=0|\alpha_i=$

$a_m)P_m$, and $\theta_{M+m} = f_0(y_{i0} = 1|\alpha_i = a_m)P_m$ for $m = 1, 2, ..., M$. By definition, $\theta_m \geq 0$ for $m = 1, 2, ..., 2M$, and $\sum_{m=1}^{2M} \theta_m = 1$. Then

$$\begin{aligned} \pi(A_j) &= \sum_{m=1}^{M} \left[\prod_{t=1}^{T} f_t(y_{it}|y_{i\,t-1},\, \alpha_i = a_m;\, \gamma)\theta_m \right] \\ &\quad + \sum_{m=M+1}^{2M} \left[\prod_{t=1}^{T} f_t(y_{it}|y_{i\,t-1},\, \alpha_i = a_{m-M};\, \gamma)\theta_m \right]. \end{aligned} \tag{2.5}$$

$\pi(A_j)$ depends on the value of $\theta = (\gamma, \theta_1, ..., \theta_{2M})$. I emphasize such dependence on parameter value by writing $\pi(A_j)$ as $\pi(A_j; \theta)$. $\pi(A_j; \theta)$ is the probability of the event $\{A_j\}$ predicted by the model, given the parameter value θ.

The idea underlying identification is to find the parameter value(s) that generates the same distribution for the observable endogenous variable as the true data generating process (DGP) does. Let $P(A_j)$ be the actual probability of event $\{A_j\}$, i.e. $P(A_j)$ is generated by the true DGP. When implemented with the true parameter value, the induced probability $\pi(A_j; \theta)$ should equal the true probability$P(A_j)$ on each single event $\{A_j\}$. The identification set of the unknown vector-valued parameter θ, therefore, is the set of parameter values that make $\pi(A_j, \theta)$ agree with $P(A_j)$ on each single event $\{A_j\}$. Specifically, the identification set

$$\Theta_I = \{\theta \in \Theta : \pi(A_j; \theta) = P(A_j) \text{ for } j = 1, ..., J\} \tag{2.6}$$

where Θ is the parameter space, and $J = 2^T$, which is the number of possible single events.

Since $P(A_j) = E_p[\mathbf{1}(y_i = A_j)]$, Θ_I is the set of parameter values satisfying the following moment conditions:

$$E_p[m_j(\theta)] := E_P[\pi(A_j; \theta) - \mathbf{1}(y_i = A_j)] = 0 \text{ for } j = 1, ..., J, \tag{2.7}$$

where the subscript P refers to the fact that the expectation is calculated under the true distribution $P(\cdot)$. Define $m(\theta) = (m_1(\theta), m_2(\theta), ..., m_J(\theta))'$. Then Θ_I is the set of parameter values satisfying the moment condition vector $E_p[m(\theta)] = 0$.

The above moment equality motivates us to use the following population criterion function

$$Q(\theta) = ||E_p[m(\theta)]||^2, \tag{2.8}$$

where $||.||$ is the euclidean distance on $\mathbb{R}^J$. It is easy to verify that Θ_I is the set of parameter values that minimize $Q(\theta)$.

2.3 Estimation

The identification set Θ_I is fully characterized by moment condition (6). Define:

$$\begin{aligned} E_n[m_j(\theta)] &= \frac{1}{n}\sum_{i=1}^{n}[\pi(A_j; \theta) - \mathbf{1}(y_i = A_j)] \\ &= \pi(A_j; \theta) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(y_i = A_j), \end{aligned} \tag{2.9}$$

which is the sample analog of (6). The sample criterion function is naturally defined as

$$Q_n(\theta) = ||E_n[m(\theta)]||^2. \tag{2.10}$$

The set estimator constructed in this paper takes the following form:

$$C_n(\hat{c}) := \{\theta \in \Theta : nQ_n(\theta) \leq \hat{c}\}, \tag{2.11}$$

which is the contour set of level $\hat{c}$ of the function $nQ_n(\theta)$. $\hat{c}$ is possibly data dependent. This is a special case of the sets constructed in CHT (2007)[1]. To analyze the asymptotic properties of the set estimators, a proper definition of distance between sets is required. I use the Hausdorff distance between sets, defined as[2]

$$d_H(A, B) := [sup_{a \in A} d(a, B)] \vee [sup_{b \in B} d(b, A)],$$

where $d(a, B) := inf_{b \in B}||b - a||$ and $d(A, B) := +\infty$ if either A or B is empty.

Throughout this paper, the parameter space Θ is a nonempty compact subset of R^d where d is the dimension of θ. (In this section $\theta = (\gamma, \theta_1, ..., \theta_{2M})$. Therefore, $d = 2M + 1$.) And both $Q(\theta)$ and $Q_n(\theta)$ are well defined on a neighborhood of Θ. Define $\mathcal{C}_n = sup_{\theta \in \Theta_I} nQ_n(\theta)$. The construction of consistent estimators of form (11) includes finding proper $\hat{c}$, s.t. $\hat{c} \geq C_n$ with probability approaching 1 but $\hat{c}/n \to_P 0$. The following

[1] In CHT (2007), $C_n(c) := \{\theta \in \Theta : a_n Q_n(\theta) \leq c\}$,where a_n is some sequence s.t. $a_n \to +\infty$.

[2] $a \vee b := max\{a, b\}$, and $a \wedge b := min\{a, b\}$.

several lemmas and theorems present the consistency results.

Lemma 2.3.1. (1) $(E_n[m(\theta)] - E_p[m(\theta)])$ is independent of θ, and

$$\sqrt{n}\left(E_n[m(\theta)] - E_p[m(\theta)]\right) \to_d \Delta \tag{2.12}$$

where $\Delta \sim N(\mathbf{0}, \Sigma)$, $\Sigma_{ii} = P(A_i)[1 - P(A_i)]$ and $\Sigma_{ij} = -P(A_i)P(A_j)$ when $i \neq j$, and (2) for any sequence of real numbers $b_n \to \infty$, $Pr(\mathcal{C}_n \leq b_n) \to 1$.

Lemma 2.3.2. The following regularity conditions (RC) hold: (a) both Q and Q_n are continuous on Θ. (b) $sup_{\Theta}|Q - Q_n| = O_p(1/\sqrt{n})$. (c) $sup_{\Theta_I} Q_n = O_p(1/n)$.

Theorem 2.3.1. Let $\widehat{\Theta}_I = C_n(b_n)$, where $b_n \to \infty$ but $b_n/n \to 0$. Suppose that $\Theta_I \neq \Theta$. Then $\Theta_I \subseteq \widehat{\Theta}_I$ with probability approaching 1, and $d_H\left(\widehat{\Theta}_I, \Theta_I\right) = o_p(1)$.

Theorem 2.3.1 states that for any sequence of real numbers b_n approaching infinity but at a slower rate than n, $C_n(b_n) = \{\theta \in \Theta : nQ_n(\theta) \leq b_n\}$ is a Hausdorff-consistent estimator of the identification set.

I assume the following partial identification condition holds:

Condition C2.3.1. There exist positive constants K and δ such that for all $\theta \in \Theta$,

$$||E_p[m(\theta)]|| \geq K \cdot (d(\theta, \Theta_I) \wedge \delta). \tag{2.13}$$

This condition is the partial-identification analogy to the condition that the Jacobian $\nabla_\theta E_P[[m(\theta)]|]$ is full-ranked and continuous near Θ_I in the point-identified case. The same condition is discussed and required by CHT (2007) for their asymptotic results to hold.

Lemma 2.3.3. under C2.3.1, there exist positive constants (δ, κ) such that for any $\varepsilon \in (0, 1)$ there are $(\kappa_\varepsilon, n_\varepsilon)$ such that for all $n \geq n_\varepsilon$,

$$Q_n(\theta) \geq \kappa \cdot [d(\theta, \Theta_I) \wedge \delta]^2$$

uniformly on $\left\{\theta \in \Theta : d(\theta, \Theta_I) \geq (\kappa_\varepsilon/n)^{1/2}\right\}$, with probability at least $1 - \varepsilon$.

The following theorem establishes the rate of convergence of the set estimators of form (2.11).

Theorem 2.3.2. Under C2.3.1, let $\widehat{\Theta}_I = C_n(b_n)$,where $b_n \to \infty$ and $b_n/n \to 0$,

$d\left(\widehat{\Theta}_I,\, \Theta_I\right) = O_p\left((b_n/n)^{1/2}\right)$.

Theorem 2.3.2 states that for any sequence of real numbers b_n approaching infinity but at a slower rate than n, $C_n\left(b_n\right) = \{\theta \in \Theta :\, nQ_n\left(\theta\right) \leq b_n\}$ converge to Θ_I at rate $(b_n/n)^{1/2}$. Theorem 2.3.2 implies that the rate of convergence can be made arbitrarily close to $O_p\left(1/\sqrt{n}\right)$ by letting b_n approach infinity at an arbitrarily slow rate. For example, by letting $b_n \propto log\, n$, or $b_n \propto log(log\, n)$, the rate of convergence is $O_p\left(\sqrt{\frac{log\, n}{n}}\right)$, or $O_p\left(\sqrt{\frac{log(log\, n)}{n}}\right)$, either of which is greater than $O_p\left(n^{-(1/2+\rho)}\right)$ for any positive ρ.

2.4 Models with Continuous Exogenous x_{it}

In many economic applications, the underlying panel data models include exogenous variable in the explanatory variable set. And, often, the exogenous variable is continuous. For instance, the exogenous variable can be the price of a certain product, which takes value on an interval of positive real numbers. In this section I consider the model

$$y_{it} = \mathbf{1}\left\{y_{i\,t-1}\gamma + x_{it}'\beta + \alpha_i \geq \varepsilon_{it}\right\} \quad (t = 1, 2, ..., T),$$

where x_{it} are continuous random variables with bounded support. Again, the observable variable $y_i = (y_{i1},\, y_{i2},\, ...,\, y_{iT})$ takes value on the finite sample space Ω. Ω has

$J = 2^T$ different elements, and can be described as $\Omega = \{A_j : j = 1, 2, ..., J\}$. Each A_j is a $T - dimension$ vector of zeros and ones.

I still maintain the assumptions made in section 2.2 on the distributions of ε_{it} and α_i. ε_{it} are jointly independent and that they are identically distributed over individuals with a known cdf F_t in the model. And α_i takes values on $\{a_m : m = 1, 2, .., M\}$ with unknown associated probability masses P_m. In addition, assume that ε_{it} has a bounded pdf $f_{\varepsilon t}$.

Now the conditional pmf of y_{it} depends on $y_{i,t-1}$, x_{it}, α_i, γ, and β. And it takes the following form

$$\begin{aligned} f_t(y_{it}|y_{i\,t-1},\, x_{it},\, \alpha_i;\, \gamma,\, \beta) &= y_{it}F_t\left(y_{i\,t-1}\gamma + x_{it}'\beta + \alpha_i\right) \\ &\quad +(1 - y_{it})\left[1 - F_t\left(y_{i\,t-1}\gamma + x_{it}'\beta + \alpha_i\right)\right]. \end{aligned} \tag{2.14}$$

For any $A_j = (y_{i1},\, y_{i2},\, ...,\, y_{iT}) \in \Omega$, the induced probability of event $\{A_j\}$ conditioned on $x_i = (x_{i1},\, x_{i2},\, \ldots,\, x_{iT})$ is

$$\pi\left(A_j;\, x_i\right) = \sum_{l\in\{0,1\}} \left\{ \sum_{m=1}^{M} \left[Pr(y_{i0} = l|\alpha_i = a_m) \prod_{t=1}^{T} f_t(y_{it}|y_{i\,t-1},\, x_{it},\, \alpha_i = a_m;\, \gamma,\, \beta) \right] P_m \right\}. \tag{2.15}$$

Again, set $\theta_m = Pr(y_{i0} = 0|\alpha_i = a_m)P_m$, and $\theta_{M+m} = f_0(y_{i0} = 1|\alpha_i = a_m)P_m$ for $m =$

$1,\ 2,\ ...,\ M$. By definition, $\theta_m \geq 0$ for $m = 1,\ 2,\ ...,\ 2M$, and $\sum_{m=1}^{2M} \theta_m = 1$. Then

$$\begin{aligned} \pi\left(A_j;\ x_i\right) &= \sum_{m=1}^{M} \left[\prod_{t=1}^{T} f_t(y_{it}|y_{i\,t-1},\ x_{it},\ \alpha_i = a_m;\ \gamma,\ \beta)\theta_m \right] \\ &\quad + \sum_{m=M+1}^{2M} \left[\prod_{t=1}^{T} f_t(y_{it}|y_{i\,t-1},\ x_{it},\ \alpha_i = a_m;\ \gamma,\ \beta)\theta_m \right]. \end{aligned} \tag{2.16}$$

$\pi\left(A_j;\ x_i\right)$ depends on the value of $\theta = (\gamma,\ ,\beta,\ \theta_1,\ ...,\ \theta_{2M})$. I emphasize this dependence on parameter value by writing $\pi\left(A_j;\ x_i\right)$ as $\pi\left(A_j;\ x_i,\ \theta\right)$. $\pi\left(A_j;\ x_i,\ \theta\right)$ is the probability of the event $\{A_j\}$ conditioned on x_i predicted by the model, given the parameter value θ.

Let $P\left(A_j|x_i\right)$ be the actual probability of event $\{A_j\}$ conditioned on x_i, i.e. $P\left(A_j|x_i\right)$ is generated by the true DGP. When implemented with the true parameter value, the induced probability $\pi\left(A_j;\ x_i,\ \theta\right)$ equals the true probability$P\left(A_j|x_i\right)$ on each single event $\{A_j\}$ almost surely over the support of x. So the identification set of θ is

$$\Theta_I = \left\{\theta \in \Theta :\ Pr\left(\pi\left(A_j;\ x_i,\ \theta\right) = P\left(A_j|x_i\right)\right) = 1\ for\ j = 1,\ ...,\ J\right\}. \tag{2.17}$$

Define $m_j\left(\theta\right) = \pi\left(A_j;\ x_i,\ \theta\right) - \mathbf{1}\left(y_i = A_j\right)$; $m\left(\theta\right) = \left(m_1\left(\theta\right),\ m_2\left(\theta\right),\ ...,\ m_J\left(\theta\right)\right)'$. Since $P\left(A_j|x_i\right) = E_p\left[\mathbf{1}\left(y_i = A_j\right)|x_i\right]$, and $E_p\left[\pi\left(A_j;\ x_i,\ \theta\right)|x\right] = \pi\left(A_j;\ x_i,\ \theta\right)$, equivalently,

$$\Theta_I = \left\{\theta \in \Theta :\ E_p\left[m\left(\theta\right)|x_i\right] = 0\ a.s.\right\}. \tag{2.18}$$

Therefore, the identification set is defined by a number of conditional moment equali-

ties.

Denote by $\mathcal{G}$ a family of vector-valued functions of x_i. Define

$$\Theta_{\mathcal{G}} = \{\theta \in \Theta : E_p [g_j (x_i) m_j (\theta)] = 0, j = 1, ..., J, \text{for all } g = (g_1, ..., g_J) \in \mathcal{G}\}. \quad (2.19)$$

$E_p [m (\theta) | x_i] = 0$ a.s. implies that $E_p [g_j (x_i) m_j (\theta)] = 0$,$j = 1, ..., J,$ for all $g \in \mathcal{G}$. Therefore $\Theta_I \subseteq \Theta_{\mathcal{G}}$.

As shown by Andrews and Shi (2011), by properly choosing instrument functions, one can transform conditional moment equalities into unconditional ones without losing identification power. In other words, by properly constructing $\mathcal{G}$, one can make $\Theta_{\mathcal{G}} = \Theta_I$. In this paper, I consider estimating $\Theta_{\mathcal{G}}$, where $\mathcal{G}$ is a finite subset of $L^d (S)$, where d is the dimension of the parameter θ, and S is the support for x. Namely, $\mathcal{G} = \{g^1, ..., g^H\}$, and

$$\Theta_{\mathcal{G}} = \left\{\theta \in \Theta : E_P \left[g_j^h (x_i) m_j (\theta)\right] = 0, for j = 1, ..., J, and h = 1, ..., H\right\}. \quad (2.20)$$

Define $M (\theta) = \left(g_1^1 m_1, ..., g_J^1 m_J, ..., g_1^H m_1, ..., g_J^H m_J\right)'$. Then $\Theta_{\mathcal{G}} = \{\theta \in \Theta : E_p [M (\theta)] = 0\}$. The estimation of $\Theta_{\mathcal{G}}$ follows the same algorithm adopted in Section 2: Define the population criterion by $Q (\theta) = ||E_P [M (\theta)]||^2$; The sample criterion $Q_n (\theta) = ||E_n [M (\theta)]||^2$, where $E_n \left[g_j^h m_j (\theta)\right] = \frac{1}{n} \sum_{i=1}^n g_j^h (x_i) m_j (\theta; x_i, y_i)$, and $E_n [M (\theta)] = (E_n \left[g_1^1 m_1 (\theta)\right], ..., E_n \left[g_J^H m_J (\theta)\right])'$. Also, still define $\mathcal{C}_n = sup_{\theta \in \Theta_{\mathcal{G}}} nQ_n (\theta)$.

And the estimator takes the following form:

$$C_n(\hat{c}) := \{\theta \in \Theta : nQ_n(\theta) \leq \hat{c}\}, \tag{2.21}$$

where $\hat{c}$ is possibly data dependent.

Lemma 2.4.1. (1) In the metric space $L^\infty(\Theta')$

$$\sqrt{n}(E_n[M(\theta)] - E_p[M(\theta)]) \Rightarrow \Delta(\theta) \tag{2.22}$$

where $\Delta(\theta)$ is a mean zero Gaussian process with almost sure continuous paths, and Θ' is a neighborhood of Θ. (2) For any sequence of real numbers $b_n \to \infty$, $Pr(\mathcal{C}_n \leq b_n) \to 1$.

Lemma 2.4.2. The following regularity conditions (RC) hold: (a) both Q and Q_n are continuous on Θ. (b) $sup_\Theta|Q - Q_n| = O_p(1/\sqrt{n})$. (c) $sup_{\Theta_\mathcal{G}} Q_n = O_p(1/n)$.

Lemma 2.4.1 and Lemma 2.4.2 are generalized versions of Lemma 2.3.1 and Lemma 2.3.2, respectively.

Theorem 2.4.1. Let $\widehat{\Theta}_\mathcal{G} = C_n(b_n)$, where $b_n \to \infty$ but $b_n/n \to 0$. Suppose that $\Theta_\mathcal{G} \neq \Theta$. Then $\Theta_\mathcal{G} \subseteq \widehat{\Theta}_\mathcal{G}$ with probability approaching 1, and $d_H\left(\widehat{\Theta}_\mathcal{G}, \Theta_\mathcal{G}\right) = o_p(1)$.

Theorem 2.4.1 states that for any sequence of real numbers b_n approaching infinity but at a slower rate than n, $C_n(b_n) = \{\theta \in \Theta : nQ_n(\theta) \leq b_n\}$ is a Hausdorff-consistent estimator of $\Theta_{\mathcal{G}}$.

I assume the same partial identification condition holds for $M(\theta)$:

Condition C2.4.1. There exist positive constants K and δ such that for all $\theta \in \Theta$,

$$||E_P[M(\theta)]|| \geq K \cdot (d(\theta, \Theta_{\mathcal{G}}) \wedge \delta). \tag{2.23}$$

Lemma 2.4.3. under C2.4.1, there exist positive constants (δ, κ) such that for any $\varepsilon \in (0, 1)$ there are $(\kappa_\varepsilon, n_\varepsilon)$ such that for all $n \geq n_\varepsilon$,

$$Q_n(\theta) \geq \kappa \cdot [d(\theta, \Theta_{\mathcal{G}}) \wedge \delta]^2$$

uniformly on $\left\{\theta \in \Theta : d(\theta, \Theta_{\mathcal{G}}) \geq (\kappa_\varepsilon/n)^{1/2}\right\}$, with probability at least $1 - \varepsilon$.

The following theorem establishes the rate of convergence of the set estimator.

Theorem 2.4.2. Under C2.4.1, let $\widehat{\Theta}_{\mathcal{G}} = C_n(b_n)$,where $b_n \to \infty$ and $b_n/n \to 0$, $d\left(\widehat{\Theta}_{\mathcal{G}}, \Theta_{\mathcal{G}}\right) = O_p\left((b_n/n)^{1/2}\right)$.

Theorem 2.4.2 states that for any sequence of real numbers b_n approaching infinity but at a slower rate than n, $C_n(b_n) = \{\theta \in \Theta : nQ_n(\theta) \leq b_n\}$ converge to $\Theta_{\mathcal{G}}$ at rate $(b_n/n)^{1/2}$. Similar to Theorem 2.3.2, Theorem 2.4.2 implies that the rate of convergence can be made arbitrarily close to $O_p(1/\sqrt{n})$ by letting b_n approach infinity at an arbitrarily slow rate.

2.5 Monte Carlo Simulations

In this section I present results from a number of simulations designed to show how the proposed estimators perform in finite samples.

The Basic Design:

I consider dynamic discrete choice model

$$y_{it} = \mathbf{1}\{y_{i\,t-1}\gamma + \alpha_i \geq \varepsilon_{it}\} \quad (t = 1, 2, 3)$$

where the true parameter values are as followings: (1) ε_{it} are i.i.d standard normal, i.e. $\varepsilon_{it} \sim N(0, 1)$; (2) $f_0(y_{i0} = 1|\alpha_i) = 0.5$ for any possible value of α_i; (3) α_i has support $\{-3.0, -2.8, ..., 2.8, 3.0\}$, with

$$P(\alpha_i = a_j) = \Phi\left((a_j + a_{j+1})/2\right) \qquad \text{for } a_j = -3.0,$$

$$P(\alpha_i = a_j) = \Phi\left((a_j + a_{j+1})/2\right) - \Phi\left((a_j + a_{j-1})/2\right) \text{ for } a_j = -2.8, ..., 2.8,$$

$$P(\alpha_i = a_j) = 1 - \Phi\left((a_j + a_{j-1})/2\right) \qquad \text{for } a_j = 3.0.$$

This distribution can be viewed as a discrete approximation of the standard normal distribution. All the results presented in this section are based on 10 replications[3] of the above model.

I focus on estimating the identification set of γ when $\gamma = -0.3,\ 0,\ and\ 0.3$, respectively. In calculating the identification set, it is known that α_i is a discrete random variable with support $\{-3.0, -2.8, ..., 2.8, 3.0\}$, but the associated probability masses are unknown. In the estimations, I use the estimator $\widehat{\Theta}_I = C_n\left(0.2 log\, n\right)$[4].

True Value	*Identification Set*	*Average Hausdorff Distance*		
		$n = 1000$	$n = 10000$	$n = 50000$
$\gamma = -0.3$	$[-0.346,\ -0,228]$	1.0693	0.2505	0.0250
$\gamma = 0$	$[-0.059,\ 0.056]$	0.8087	0.1347	0.0253
$\gamma = 0.3$	$[0.239,\ 0.350]$	0.6651	0.1725	0.0344

TABLE 2.1

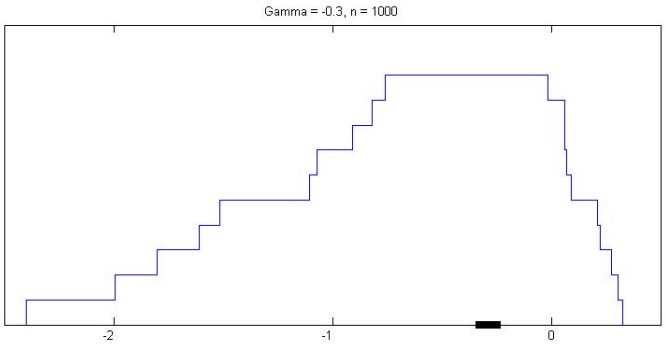
The Hausdorff distance between the identification set and the estimated set is an measure of the bias and precision of the set estimators. In Table 2.1 above, I present the identification sets and their average Hausdorff distances from the estimated sets for $\gamma = -0.3,\ 0,\ and\ 0.3$, and sample sizes $n = 1000,\ 10000,\ and\ 50000$. According to Table 2.1, the distance between the identification set and the estimated set decreases

[3]The number of replications are relatively small due to lack of an automatic way of refining and recording the result of each single experiment.

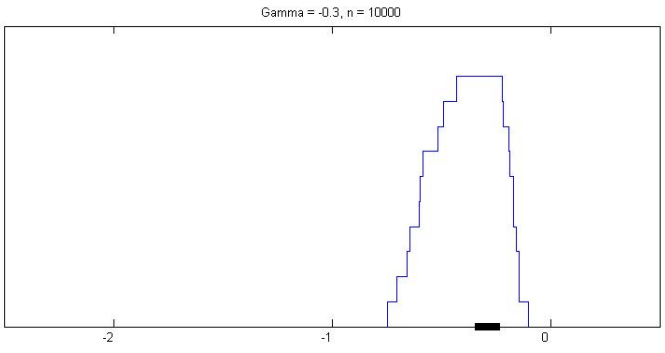
[4]I find that when b_n is set equal to either $0.2 log\, n$ or $log(log\, n)$ the set estimators behave well for samples of size 1000 and 10000. The estimators do not behave as well when b_n is set equal to $log\, n$

rapidly as the sample size increases for each $\gamma = -0.3,\ 0$, and 0.3 case. This observation supports Theorem 3.1 which implies that $\widehat{\Theta}_I = C_n\left(0.2 log\, n\right)$ is a consistent estimator of the Θ_I. However, the results from finite sample simulations have nothing to say on the rate of convergence.

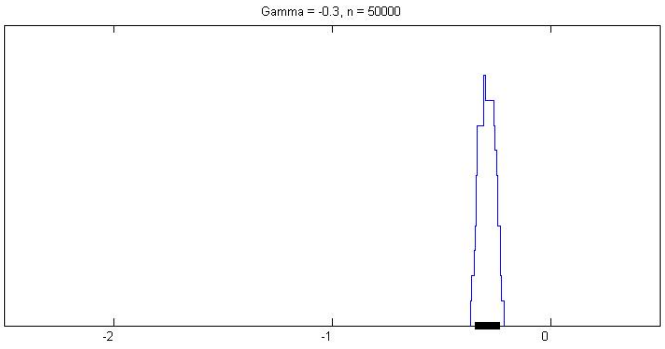
Figure 2.1 below shows, graphically, that the estimated sets converge to the identification set for each different value of γ as the sample size n increases. Each graph plots a step function $I(x) := \sum_{j=1}^{10} \mathbf{1}\left(x \in \widehat{\Theta}_j\right)$, where $\widehat{\Theta}_j$ is the set estimated in the jth replication for a given true value of γ and sample size n. The histogram-like graphs do not imply any kind of probability distribution or probability of coverage, though.



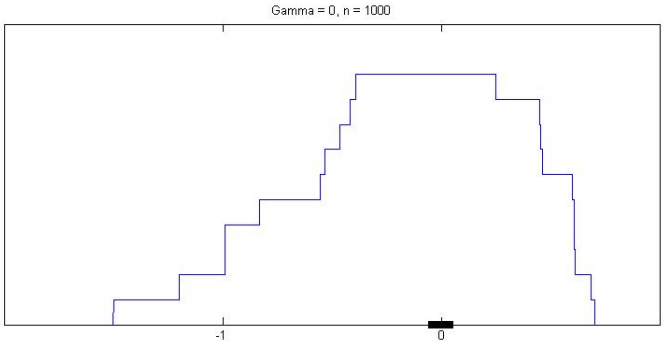
$\gamma = -0.3,\ n = 1000$



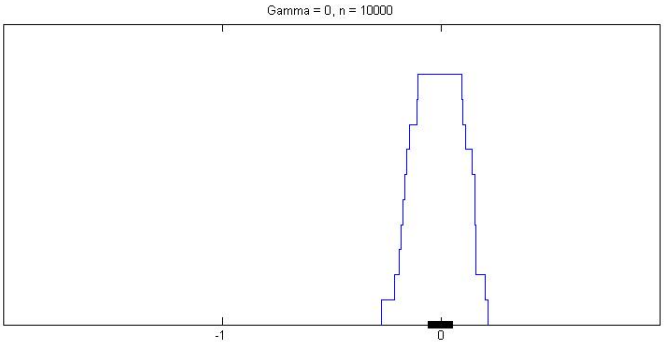
$\gamma = -0.3,\ n = 10000$



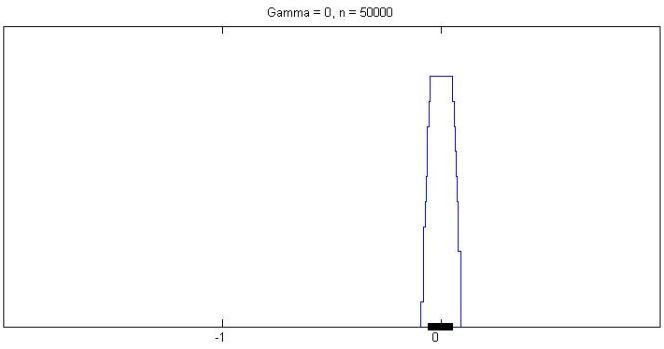
$\gamma = -0.3,\ n = 50000$



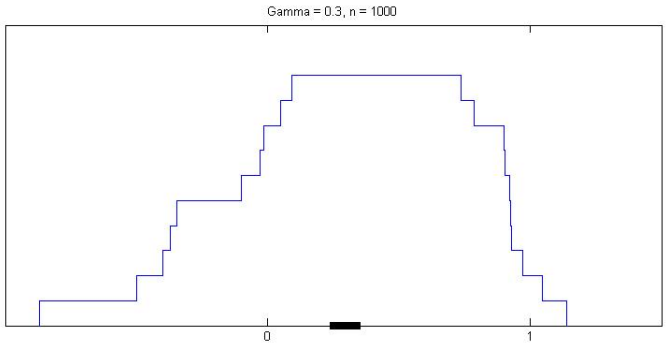
$\gamma = 0,\ n = 1000$



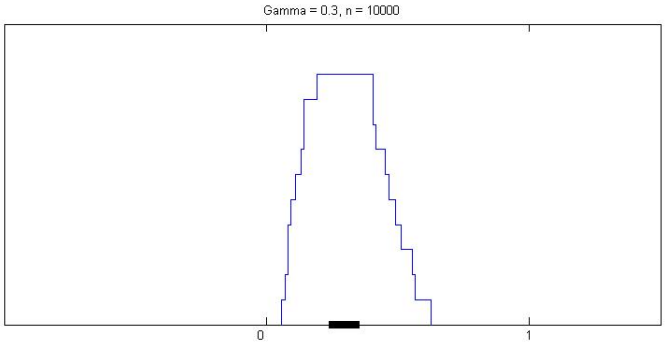
$\gamma = 0,\ n = 10000$



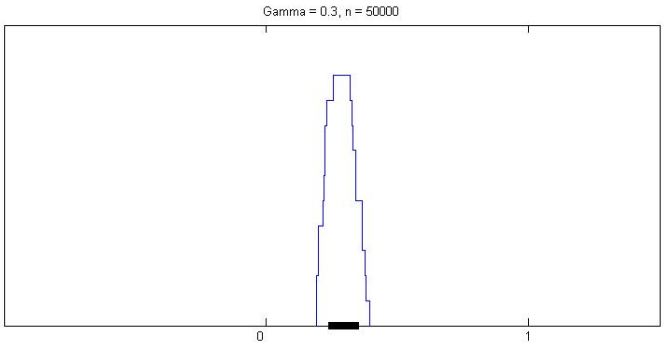
$\gamma = 0,\ n = 50000$



$\gamma = 0.3,\ n = 1000$



$\gamma = 0.3,\ n = 10000$



$\gamma = -0.3,\ n = 50000$

FIGURE 2.1[5]

One observation from the simulation is that almost all estimated sets are intervals. Another observation is that the proposed set estimators are rough when sample size is not large ($n = 1000$, or 10000). Specifically, the sizes of the estimated sets are noticeably

[5]The identification set is marked as a bold interval on the horizontal axis.

larger than the size of the identification set. However, I find that the identification sets lie entirely in the estimated sets mostly when sample size is not large ($n = 1000$, or 10000). This is a desirable feature for applications in which one wants to avoid missing any part of the identification set in estimation. The size of the estimated set (or, the length when the estimated set is an interval) can be smaller than the size of the identification set sometimes when sample size is large ($n = 50000$). And, when this happens, it is not always the case that the estimated set is a subset of the identification set.

The precision of the estimation are possibly affected by: (1) the dimension of the parameter to estimate, and (2) the distributions of the error term and the individual effect. Therefore, that the proposed estimators are rough with sample of moderate size might be because: (1) the dimension of the parameter θ is too high;[6] (2) the distributions of the individual effect α_i and the error term ε_{it} have similar magnitudes of fluctuation[7]. The following alternative designs aim at investigating the effects of these two factors on the precision of the proposed estimators.

Alternative Design 1:

In this design, I change the true distribution of α_i, while keeping everything else in the basic design unchanged. Now α_i has support $\{-3.0, -2.5, ..., 2.5, 3.0\}$, with

$$P(\alpha_i = a_j) = \Phi\left((a_j + a_{j+1})/2\right) \qquad \text{for } a_j = -3.0,$$

[6] The dimension of θ is 63 in the basic design.

[7] The variance of α_i and ε_{it} are close. In addition, the support of α_i lies in $[-3, 3]$, while ε_{it} clusters in $[-3, 3]$.

$P(\alpha_i = a_j) = \Phi\left(\left(a_j + a_{j+1}\right)/2\right) - \Phi\left(\left(a_j + a_{j-1}\right)/2\right)$ for $a_j = -2.5, ..., 2.5$,

$P(\alpha_i = a_j) = 1 - \Phi\left(\left(a_j + a_{j-1}\right)/2\right)$ for $a_j = 3.0$.

Compared with the basic design, α_i has fewer supporting points now. Effectively, the dimension of θ is reduced from 63 in the basic design to 27. The following table shows the identification sets and their average Hausdorff distances from the estimated sets for different true parameter values and sample sizes for this design.

True Value	*Identification Set*	*Average Hausdorff Distance*		
		$n = 1000$	$n = 10000$	$n = 50000$
$\gamma = -0.3$	$[-0.322, -0.248]$	1.156	0.225	0.033
$\gamma = 0$	$[-0.059, 0.060]$	0.7044	0.1126	0.0248
$\gamma = 0.3$	$[0.231, 0.341]$	0.8312	0.1572	0.0354

TABLE 2.2

Compared with Table 2.1, the identification set for each different true value of γ in Table 2.2 stays approximately unchanged in both size and location after changing the distribution of α_i. Also, compared with Table 2.1, when the sample size is 1000, neither do the average Hausdorff distances for different true values of γ decrease uniformly, nor they change very much. Therefore, reducing the dimension of θ does not improve the precision of the estimators when the sample size is 1000. However, Table 2.2 does show slight decrease in the average Hausdorff distances for all different true values of γ when the sample size is 10000. The average Hausdorff distances remain approximately unchanged when the sample size is 50000.

Alternative Design 2:

Like in the previous design, in this design only the true distribution of α_i changes, while everything else remains the same as in the basic design. Now α_i has support $\{-3.0, -2.0, ..., 2.0, 3.0\}$, with

$$P(\alpha_i = a_j) = \Phi\left(\left(a_j + a_{j+1}\right)/2\right) \qquad \text{for } a_j = -3.0,$$

$$P(\alpha_i = a_j) = \Phi\left(\left(a_j + a_{j+1}\right)/2\right) - \Phi\left(\left(a_j + a_{j-1}\right)/2\right) \text{ for } a_j = -2.0, ..., 2.0,$$

$$P(\alpha_i = a_j) = 1 - \Phi\left(\left(a_j + a_{j-1}\right)/2\right) \qquad \text{for } a_j = 3.0.$$

Compared with Alternative Design 1, α_i has even fewer supporting points now. And the dimension of θ is reduced from 27 in Alternative Design 1 to 15. The following table shows the identification sets and their average Hausdorff distance from the estimated sets for different true parameter values and sample sizes for this design.

True Value	*Identification Set*	*Average Hausdorff Distance*		
		$n = 1000$	$n = 10000$	$n = 50000$
$\gamma = -0.3$	$[-0.346, -0.248]$	0.8046	0.1234	0.0366
$\gamma = 0$	$[-0.056, 0.056]$	0.777	0.12	0.029
$\gamma = 0.3$	$[0.241, 0.346]$	0.616	0.1642	0.0386

TABLE 2.3

Compared with Table 2.1, the identification set for each different true value of γ in Table 2.3 still stays approximately unchanged in both size and location. Table 2.3 shows noticeable decrease in the average Hausdorff distances for all different true values of γ when the sample size is 1000. Also, Table 2.3 shows slightly decrease in the average Hausdorff distances when the sample size is 10000. Therefore, further reducing the

dimension of θ does improve the precision of the estimators when the sample size is 1000 and 10000. The average Hausdorff distances remain approximately unchanged when the sample size is 50000.

Alternative Design 3:

In the basic design, the distributions of the individual effect α_i and the error term ε_{it} have similar magnitudes of fluctuation. In general, the phenomenon that the fluctuation magnitude of the error term is comparable with that of some explanatory variable could potentially compromise the precision of the estimators. In this design and the next design, I exam whether this is the case here. In this design, I change the distribution of ε_{it} so that its variance is significantly smaller than the variance of α_i, while keeping everything else in the basic design unchanged. Now, $\varepsilon_{it} \sim N(0,\, 0.09)$.

Table 2.4 below shows the identification sets and their average Hausdorff distances from the estimated sets for different true parameter values and sample sizes for this design.

True Value	*Identification Set*	*Average Hausdorff Distance*		
		$n = 1000$	$n = 10000$	$n = 50000$
$\gamma = -0.3$	$[-2.657,\, -0.232]$	0.1453	0.0787	0.0143
$\gamma = 0$	$[-0.032,\, 0.028]$	2.5847	0.104	0.014
$\gamma = 0.3$	$[0.247,\, 0.373]$	3.1203	0.147	0.0337

TABLE 2.4

Accordingly, the size of the identification sets change significantly as the distribu-

tion of the error term changes. More specifically, as the variance of ε_{it} decreases, the identification set of γ when $\gamma = -0.3$ extends dramatically. The identification set for $\gamma = 0$ shrinks to about half the size as the original one. The identification set for $\gamma = 0.3$ extends slightly. Compared with Table 2.1, the precision of the estimators is improved noticeable for each sample size when $\gamma = -0.3$. The precision is also improved for sample size 10000 and 50000 when $\gamma = 0$ or 0.3. However, the estimators are much less precise for sample size 1000 when $\gamma = 0$ or 0.3.

Alternative Design 4:

In this design, $\varepsilon_{it} \sim N(0, 9)$, while everything else is kept unchanged. Opposite to the previous design, the variance of ε_{it} is significantly larger than that of α_i. Table 5 below shows the identification sets and the average Hausdorff distances for this design.

True Value	*Identification Set*	*Average Hausdorff Distance*		
		$n = 1000$	$n = 10000$	$n = 50000$
$\gamma = -0.3$	$[-0.394, -0.197]$	0.933	0.2963	0.073
$\gamma = 0$	$[-0.126, 0.123]$	1.0917	0.2493	0.056
$\gamma = 0.3$	$[0.202, 0.466]$	0.7473	0.3287	0.0446

TABLE 2.5

After increasing the variance of ε_{it}, the identification sets for all different values of γ extend slightly. Compared with the basic design, all the estimators are less precise except the one for sample size 1000 when $\gamma = -0.3$.

Summary of Simulation Results

The simulation results suggest not to use the set estimator when sample size is 1000 or less in general. Comparisons of simulation results in the 5 different designs show/suggest: In the discrete choice panel data model, (1) size and location of the identification set of γ is not sensitive to the dimension of the unknown parameter; (2) size of the identification set of γ is sensitive to the distribution of the error term; (3) when sample size is not very large, the estimators get more precise as the dimension of the unknown parameter decreases. Yet, this improvement in precision is only noticeable when the decrease in the dimension of the unknown parameter is large enough; (4) relation between the fluctuation magnitudes of α_i and ε_{it} affects the precision of the estimators in a complicated way, especially when moving from the similar fluctuation magnitudes case to the case in which ε_{it} has a much smaller variance than α_i; (5) when moving from the similar fluctuation magnitudes case to the case in which ε_{it} has a much larger variance than α_i, the estimators are less precise; (6) the Hausdorff distances decrease rapidly as the sample size increases in all different designs, which confirms Theorem 2.3.1.

2.6 Conclusions

In this paper, I consider estimation in dynamic discrete choice panel data models. Point-identification can fail due to unobservable heterogeneity among individuals

and the initial condition problem. Therefore, the focus is on estimating the identification set by set estimators. The proposed estimators are the contour sets of some level c of the function $nQ_n(\theta)$, where Q_n is the sample criterion function. These are special cases of the contour set constructed in CHT (2007). Starting with the simple discrete choice model $y_{it} = \mathbf{1}\{y_{i\,t-1}\gamma + \alpha_i \geq \varepsilon_{it}\}$, I show that for any sequence$\{b_n\}$ s.t. $b_n \to +\infty$ and $b_n/n \to 0$, $C_n(b_n) = \{\theta \in \Theta : nQ_n(\theta) \leq b_n\}$ is a consistent estimator under regularity condition (C2.3.1). And the rate of convergence can be made arbitrarily close to $O_p(1/\sqrt{n})$ by letting b_n approach infinity at an arbitrarily slow rate.The identification and estimation approach is easily extended to the models $y_{it} = \mathbf{1}\{x_{it}'\beta + y_{i\,t-1}\gamma + \alpha_i + \varepsilon_{it} \geq 0\}$, where x_{it} are exogenous discrete variables with finite support points.

For models with continuous x_{it}, the identification set is characterized by conditional moment equalities. As shown by Andrews and Shi (2011), by properly choosing instrument functions, one can transform conditional moment equalities into unconditional ones without losing identification power. For any arbitrary number of instrument functions of x_{it}, I estimate the subset of the parameter space Θ that is characterized by finitely many unconditional moment equalities transformed from the conditional moment equalities via these instrument functions. The estimators still take the form $C_n(b_n) = \{\theta \in \Theta : nQ_n(\theta) \leq b_n\}$. And similar consistency results hold for $\{b_n\}$ s.t. $b_n \to +\infty$ and $b_n/n \to 0$ under regularity condition (C2.4.1).

The simulation results confirm the consistency. Besides that, the simulation re-

sults suggest that the size and location of the identification set of γ is not sensitive to the dimension of the unknown parameter, but very sensitive to the distribution of the error term. And the estimation precision increases as the dimension of the unknown parameter decreases. Finally, both the identification set and the estimation precision are affected by the distribution of the error term in a complicated way, which requires further study to understand.

Future work includes the construction of the confidence region for the identification set, developing criteria for picking the instrument functions, as well as the identification and estimation in models where the distribution of α_i is nonparametric.

References

AI, C., AND X. CHEN (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1843.

ANDREWS, D. W. (1994): "Empirical process methods in econometrics," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. McFadden, vol. 4 of *Handbook of Econometrics*, chap. 37, pp. 2247–2294. Elsevier.

ANDREWS, D. W., AND X. SHI (2011): "Inference Based on Conditional Moment Inequalities," *Econometrica*, forthcoming.

ANDREWS, D. W. K., AND G. SOARES (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78(1), 119–157.

ARADILLAS-LOPEZ, A. (2010): "Semiparametric estimation of a simultaneous game with incomplete information," *Journal of Econometrics*, 157(2), 409–431.

ARELLANO, M., AND B. E. HONORÉ (2001): "Panel data models: some recent developments," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 5 of *Handbook of Econometrics*, chap. 53, pp. 3229–3296. Elsevier.

BERESTEANU, A., AND F. MOLINARI (2008): "Asymptotic Properties for a Class of Partially Identified Models," *Econometrica*, 76(4), 763–814.

BIERENS, H. J. (1982): "Consistent model specification tests," *Journal of Econometrics*, 20(1), 105–134.

——— (1990): "A Consistent Conditional Moment Test of Functional Form," *Econometrica*, 58(6), 1443–58.

BLUNDELL, R., X. CHEN, AND D. KRISTENSEN (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 75(6), 1613–1669.

BUGNI, F. A. (2010): "Bootstrap Inference in Partially Identified Models Defined by Moment Inequalities: Coverage of the Identified Set," *Econometrica*, 78(2), 735–753.

CANAY, I. A. (2010): "EL inference for partially identified models: Large deviations optimality and bootstrap validity," *Journal of Econometrics*, 156(2), 408–425.

CANAY, I. A., A. SANTOS, AND A. M. SHAIKH (2011): "On the Testability of Identification in Some Nonparametric Models with Endogeneity," Working paper.

CHAMBERLAIN, G. (1984): "Panel data," in *Handbook of Econometrics*, ed. by Z. Griliches, and M. D. Intriligator, vol. 2 of *Handbook of Econometrics*, chap. 22, pp. 1247–1318. Elsevier.

——— (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34(3), 305–334.

CHEN, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6 of *Handbook of Econometrics*, chap. 76. Elsevier.

CHEN, X., V. CHERNOZHUKOV, S. S. LEE, AND W. NEWEY (2011): "Local identification of nonparametric and semiparametric models," CeMMAP working papers CWP17/11, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

CHEN, X., AND D. POUZO (2009): "Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals," *Journal of Econometrics*, 152(1), 46–60.

——— (2011): "Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Generalized Residuals," *Econometrica*, forthcoming.

CHERNOZHUKOV, V., H. HONG, AND E. TAMER (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75(5), 1243–1284.

CHINTAGUNTA, P., E. KYRIAZIDOU, AND J. PERKTOLD (2001): "Panel data analysis of household brand choices," *Journal of Econometrics*, 103(1-2), 111–153.

DAVYDOV, Y. A.AND LIFSHITS, M. A., AND N. V. SMORODINA (1998): *Local Properties of Distributions of Stochastic Functionals*. Providence, RI: American Mathematical Society.

DOMINGUEZ, M. A., AND I. N. LOBATO (2004): "Consistent Estimation of Models Defined by Conditional Moment Restrictions," *Econometrica*, 72(5), 1601–1615.

DUDLEY, R. M. (1985): "An Extended Wichura Theorem, Definitions of Donsker Class, and Weighted Empirical Distribution," in *Probability in Banach Space*, vol. V of *Lecture Notes in Mathematics*, pp. 141–178. Berlin: Springer.

GALLANT, A. R., AND D. W. NYCHKA (1987): "Semi-nonparametric Maximum Likelihood Estimation," *Econometrica*, 55(2), 363–90.

GINE, E., AND J. ZINN (1990): "Bootstrapping general empirical measures," *Annal of Probab*, 18, 851–869.

HAHN, J., AND G. RIDDER (2009): "Partial Identification and Confidence Intervals," Working paper.

HANSEN, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029–54.

HONORÉ, B. E., AND E. KYRIAZIDOU (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68(4), 839–874.

HONORÉ, B. E., AND E. TAMER (2006): "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 74(3), 611–629.

HOROWITZ, J. L. (2009): "Specification Testing in Nonparametric Instrumental Variables Estimation," Working paper.

IMBENS, G. W., AND C. F. MANSKI (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72(6), 1845–1857.

KIM, K. I. (2009): "Set Estimation and Inference with Models Characterized by Conditional Moment Inequalities," Working paper.

KRESS, R. (1999): *Linear Integral Equations*. Springer, New York.

LESER, C. E. V. (1963): "Forms of Engel Functions," *Econometrica*, 31, 694–703.

MANSKI, C. F. (1990): "Nonparametric Bounds on Treatment Effects," *American Economic Review*, 80(2), 319–23.

——— (2003): *Partial Identification of Probability Distributions*. Springer, New York.

MANSKI, C. F., AND E. TAMER (2002): "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70(2), 519–546.

NEWEY, W. K. (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica*, 58(4), 809–37.

NEWEY, W. K., AND J. L. POWELL (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71(5), 1565–1578.

PRAIS, S. J., AND H. S. HOUTHAKKER (1955): *The Analysis of Family Budgets*. Cambridge: Cambridge University Press.

ROMANO, J. P., AND A. M. SHAIKH (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78(1), 169–211.

SANTOS, A. (2011): "Inference in Nonparametric Instrumental Variables with Partial Identification," *Econometrica*, forthcoming.

VAN DER VAART, A. W., AND J. WELLNER (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.

WORKING, H. (1943): "Statistical Laws of Family Expenditure," *Journal of the American Statistical Association*, 38, 43–56.

Appendix A

Proofs of the Theorems in Chapter 1

In Appendix A, we prove Theorems 1.3.1-1.3.3, 1.4.1-1.4.2. We start with listing the lemmas that we will use in the proofs. Proofs of these lemmas are given in Appendix B.

The following lemmas are used in the proofs of the theorems:

Lemma A1. Let Assumption 1.3.1(ii) and 1.3.3(ii) hold. Then,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{m\left(y_i,\,\theta,\,h\left(x_i\right)\right)\cdot g\left(t,\,z_i\right)-E\left[m\left(Y,\,\theta,\,h\left(X\right)\right)\cdot g\left(t,\,Z\right)\right]\right\}\xrightarrow{\mathcal{L}}G\left(\alpha,\,t\right)$$

where $G\left(\alpha,\,t\right)$ is a tight Gaussian process on $\mathcal{A}\times T$.

Lemma A2. (convergence rate in $\|\cdot\|_w$) Let Assumption 1.2.1, 1.2.2, 1.3.1, 1.3.2,

and 1.3.3(i)-(iv) hold. For any

$$\hat{\alpha}_n \in \underset{\alpha \in \mathcal{A}_n \cap R}{\arg\min} \left\{ \underset{t \in T}{\sup} n \left[J_n(\alpha,\, t) \right]' W_n(\alpha,\, t) \left[J_n(\alpha,\, t) \right] \right\}$$

as in (1.7)

$$d_w(\hat{\alpha}_n,\, \mathcal{A}_I \cap R) = O_p\left(\max\left\{ \delta_{w,n},\, n^{-1/2} \right\} \right).$$

Lemma A3. (convergence rate in $\|\cdot\|_{L_2}$) Let Assumption 1.2.1, 1.2.2, 1.3.1-1.3.3 hold. Then

$$d_{L_2}(\hat{\alpha}_n,\, \mathcal{A}_I \cap R) = O_p\left(\delta_{s,n} + \psi_n \cdot \max\left\{ \delta_{w,n},\, n^{-1/2} \right\} \right).$$

Lemma A4. Let A, B be sets of two Hilbert spaces, respectively, and $\mathcal{F}_n^b \subseteq \mathcal{F}_{n+1}^b \subseteq L^\infty(A)$ with $\mathcal{F}_\infty^b$ being the closure of $\cup \mathcal{F}_n^b$ under $||\cdot||_\infty$ for each $b \in B$. If $g \in L^\infty(A \times B)$, and there exists a sequence $\{g_n\} \in L^\infty(A \times B)$ s.t. $||g_n - g||_\infty = o(1)$, then:

$$\inf_{b \in B} \inf_{f \in \mathcal{F}_n^b} \sup_{a \in A} |g_n(a,\, b) - f(a)| \longrightarrow \inf_{b \in B} \inf_{f \in \mathcal{F}_\infty^b} \sup_{a \in A} |g(a,\, b) - f(a)|.$$

Lemma A5. Let A_n, B be sets with norms $||\cdot||_a$, $||\cdot||_b$ and let $G_n : A_n \times B \to \mathbb{R}$ and $F_n : A_n \times B \to \mathbb{R}$ be random functions. Assume: (1) G_n and F_n are continuous on $A_n \times B$

almost sure; (2) $\sup_{a\in A_n}\sup_{b\in B}\left[G_n\left(a,\,b\right)-F_n\left(a,\,b\right)\right]^2=o_p\left(1\right)$; and (3) $\inf_{a\in A_n}\sup_{b\in B}F_n^2\left(a,\,b\right)=O_p\left(1\right)$. It then follows:

$$\inf_{a\in A_n}\sup_{b\in B}G_n^2\left(a,\,b\right)=\inf_{a\in A_n}\sup_{b\in B}F_n^2\left(a,\,b\right)+o_p\left(1\right).$$

Lemma A6. For any $c>0$, let $\mathcal{H}_{nR,c}\equiv\left\{\left(r_\theta,\,r_h\right)\in\mathcal{H}_{nR}:\,\|r_h'p^{k_n}\|_s^2\leq c,\,\|r_h\|_E^2\leq c\right\}$. Define the following family of functions:

$$V_{k_n,c}^{\alpha_0}=\{v:\,T\to\mathbb{R}^{d_m}\,s.t.\,v(t)=\frac{\partial\rho\left(\theta_0,\,h_0,\,t\right)}{\partial\theta'}\cdot r_\theta+\frac{d\,\rho\left(\theta_0,\,h_0,\,t\right)}{d\,h}\left[p^{k_n}\right]'r_h,\,\left(r_\theta,\,r_h\right)\in\mathcal{H}_{nR,c}\}.$$

For any real number sequence $c_n\to\infty$ and $\alpha_0\in\mathcal{A}_I$, we have

$$\cup V_{k_n,c_n}^{\alpha_0}=\cup V_{k_n}^{\alpha_0}.$$

(Proofs of Lemmas A1-A6 above are given in Appendix B.)

Proof of Theorem 1.3.1

Without loss of generality, assume $m\left(\cdot\right)$ and $h\left(\cdot\right)$ to be real-valued in the proof for simplicity of notation.

Define $u_i(\alpha, t) = m(y_i, \theta, h(x_i)) \cdot g(t, z_i)$.

First consider the case where W_n is chosen to be the identity matrix. Lemma A3, Assumption 1.3.3(v), and Assumption 1.3.5(ii) imply that for

$$\hat{\alpha}_n \in \underset{\alpha \in \mathcal{A}_n \cap R}{\arg\min} \sup_{t \in T} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_i(\alpha, t) \right]^2$$

$$d_{L_2}(\hat{\alpha}_n, \mathcal{A}_I \cap R) = o_p\left(n^{-1/4}\right). \tag{A.1}$$

Let $\delta_n = o_p\left(n^{-1/4}\right)$ such that $n^{-1/2} = o(\delta_n)$. Define the neighborhoods

$$B^{\delta_n}(\alpha_0) = \{\alpha \in \mathcal{A}_n : ||\alpha - \alpha_0||_{L^2} \leq \delta_n\}$$

for all $\alpha_0 \in \mathcal{A}_I$. Then

$$\min_{\alpha \in \mathcal{A}_n \cap R} \sup_{t \in T_n} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_i(\alpha, t) \right]^2 = \inf_{\alpha_0 \in \mathcal{A}_I \cap R} \min_{\alpha \in B^{\delta_n}(\alpha_0)} \sup_{t \in T} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_i(\alpha, t) \right]^2 + o_p(1). \tag{A.2}$$

By Lemma A1, the empirical process

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{m(y_i, \theta, h(x_i)) \cdot g(t, z_i) - E[m(Y, \theta, h(X)) \cdot g(t, Z)]\}, (\alpha, t) \in \mathcal{A} \times T$$

is asymptotically $|| \cdot ||_{L^2}$-equicontinuous in probability w.r.t. α, uniformly on T. So,

since $\delta_n \downarrow 0$,

$$\sup_{||\alpha_1-\alpha_2||_{L^2}\leq\delta_n} \sup_T |\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{[m(y_i,\,\theta_1,\,h_1(x_i)) - m(y_i,\,\theta_2,\,h_2(x_i))]\, g(t,\,z_i) - E[[m(Y,\,\alpha_1) - m(Y,\,\alpha_2)]\, g(t,\,Z)]\}| = o_p(1). \tag{A.3}$$

Therefore,

$$\begin{aligned}
&\min_{\alpha\in\mathcal{A}_n\cap R}\sup_{t\in T_n}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i(\alpha,\,t)\right]^2 \\
=\;& \inf_{\alpha_0\in\mathcal{A}_I\cap R}\min_{\alpha\in B^{\delta_n}(\alpha_0)}\sup_{t\in T}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i(\alpha,\,t)\right]^2 + o_p(1) \\
=\;& \inf_{\alpha_0\in\mathcal{A}_I\cap R}\min_{\alpha\in B^{\delta_n}(\alpha_0)}\sup_{t\in T}\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i(\alpha_0,\,t) + \\
& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[m(y_i,\,\theta,\,h(x_i)) - m(y_i,\,\theta_0,\,h_0(x_i))]\cdot g(t,\,z_i)\}^2 + o_p(1) \\
=\;& \inf_{\alpha_0\in\mathcal{A}_I\cap R}\min_{\alpha\in B^{\delta_n}(\alpha_0)}\sup_{t\in T}\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i(\alpha_0,\,t) + \sqrt{n}E[[m(Y,\,\alpha) - m(Y,\,\alpha_0)]\cdot g(t,\,Z)]\}^2 \\
& + o_p(1).
\end{aligned} \tag{A.4}$$

where the last equality in (A.4) follows from (A.3) and Lemma A5.

Notice that for any $\alpha_0 \in \mathcal{A}_I$, $\alpha \in B^{\delta_n}(\alpha_0)$, we have $||\alpha - \alpha_0||_{L^2} = o_p\left(n^{-1/4}\right)$. Conse-

quently,

$$\begin{aligned}
&\sqrt{n}E\left[\left[m\left(Y,\,\alpha\right)-m\left(Y,\,\alpha_0\right)\right]\cdot g\left(t,\,Z\right)\right] \\
&=\sqrt{n}\frac{d\,E\left[m\left(Y,\,\alpha_0\right)\cdot g\left(t,\,Z\right)\right]}{d\,\alpha}\left[\alpha-\alpha_0\right]+\sqrt{n}\frac{d^2\,E\left[m\left(Y,\,\tilde{\alpha}\right)\cdot g\left(t,\,Z\right)\right]}{d\,\alpha^2}\left[\alpha-\alpha_0,\,\alpha-\alpha_0\right] \\
&\leq\sqrt{n}\frac{d\,E\left[m\left(Y,\,\alpha_0\right)\cdot g\left(t,\,Z\right)\right]}{d\,\alpha}\left[\alpha-\alpha_0\right]+\sqrt{n}\cdot C\cdot\left|\left|\alpha-\alpha_0\right|\right|_{L^2}^2\quad\text{for some } C>0 \\
&=\sqrt{n}\frac{d\,E\left[m\left(Y,\,\alpha_0\right)\cdot g\left(t,\,Z\right)\right]}{d\,\alpha}\left[\alpha-\alpha_0\right]+o_p\left(1\right), \qquad \text{(A.5)}
\end{aligned}$$

where $\tilde{\alpha}\equiv\alpha_0+s\left(\alpha-\alpha_0\right)$ for some $s\in\left(0,\,1\right)$, and the inequality in (A.5) follows from Assumption 1.3.6 (*ii*) and (*iii*).

Continuing with (A.4) and reapplying Lemma A5 yield the first equality in (A.6) below.

$$\begin{aligned}
&\min_{\alpha\in\mathcal{A}_n\cap R}\sup_{t\in T_n}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i\left(\alpha,\,t\right)\right]^2 \\
=\;&\inf_{\alpha_0\in\mathcal{A}_I\cap R}\min_{\alpha\in B^{\delta_n}(\alpha_0)}\sup_{t\in T}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i\left(\alpha_0,\,t\right)+\sqrt{n}\frac{d\,E\left[m\left(Y,\,\alpha_0\right)\cdot g\left(t,\,Z\right)\right]}{d\,\alpha}\left[\alpha-\alpha_0\right]\right\}^2 \\
&+o_p(1) \\
=\;&\inf_{\alpha_0\in\mathcal{A}_I\cap R}\min_{\alpha\in B^{\delta_n}(\alpha_0)}\sup_{t\in T}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i\left(\alpha_0,\,t\right)+\sqrt{n}\frac{d\,E\left[m\left(Y,\,\alpha_0\right)\cdot g\left(t,\,Z\right)\right]}{d\,\alpha}\left[\alpha-\Pi_n\alpha_0\right]\right\}^2 \\
&+o_p(1) \qquad \text{(A.6)}
\end{aligned}$$

where $\Pi_n\alpha_0 \equiv (\theta_0,\, \Pi_n h_0)$ and the second equality is because

$$\begin{aligned}
& \sqrt{n}\frac{d\,E\left[m\left(Y,\,\alpha_0\right)\cdot g\left(t,\,Z\right)\right]}{d\,\alpha}\left[\alpha-\alpha_0\right] \\
= \;& \sqrt{n}\frac{d\,E\left[m\left(Y,\,\alpha_0\right)\cdot g\left(t,\,Z\right)\right]}{d\,\theta}\left(\theta-\theta_0\right)+\sqrt{n}\frac{d\,E\left[m\left(Y,\,\alpha_0\right)\cdot g\left(t,\,Z\right)\right]}{d\,h}\left[h-h_0\right] \\
= \;& \sqrt{n}\frac{d\,E\left[m\left(Y,\,\alpha_0\right)\cdot g\left(t,\,Z\right)\right]}{d\,\theta}\left(\theta-\theta_0\right)+\sqrt{n}\frac{d\,E\left[m\left(Y,\,\alpha_0\right)\cdot g\left(t,\,Z\right)\right]}{d\,h}\left[h-\Pi_n h_0\right]+o\left(1\right) \\
= \;& \sqrt{n}\frac{d\,E\left[m\left(Y,\,\alpha_0\right)\cdot g\left(t,\,Z\right)\right]}{d\,\alpha}\left[\alpha-\Pi_n\alpha_0\right]+o\left(1\right) \qquad \text{(A.7)}
\end{aligned}$$

where the second equality in (A.7) is because $\sqrt{n}\,|\frac{d\,E\left[m\left(Y,\,h_0\right)\cdot w\left(t,\,Z\right)\right]}{d\,h}\left[\Pi_n h_0-h_0\right]|\leq \sqrt{n}\|\Pi_n h_0-h_0\|_w=o\left(1\right)$ uniformly on $\mathcal{H}_I$ according to Assumption 1.3.5 (ii).

Next, we focus on the local parameters of $\mathcal{H}_I$ of the form

$$h=\Pi_n h_0+\frac{p^{k_n'}r}{\sqrt{n}},\;\; h_0\in\mathcal{H}_I,\, r\in\mathbb{R}^{k_n}.$$

Note that for any $h\in B^{\delta_n}\left(h_0\right)$, there exists $r\in\mathbb{R}^{k_n}$, s.t. $h=\Pi_n h_0+\frac{p^{k_n'}r}{\sqrt{n}},\;\; h_0\in\mathcal{H}_I$. Therefore

$$B^{\delta_n}\left(h_0\right)\subseteq\left\{h\in\mathcal{H}:\; h=\Pi_n h_0+\frac{p^{k_n'}r}{\sqrt{n}},\, r\in\mathbb{R}^{k_n}\right\},\,\forall h_0\in\mathcal{H}_I. \qquad \text{(A.8)}$$

Now we proceed to show that there exists a real number sequence $c_n\nearrow\infty$ s.t.

$$h=\Pi_n h_0+\frac{p^{k_n'}r}{\sqrt{n}}\in B^{\delta_n}\left(h_0\right), \qquad \text{(A.9)}$$

for any $r\in\mathbb{R}^{k_n}$ that satisfies $\|p^{k_n'}r\|_E^2\leq c_n^2,\, ||r||_E\leq c_n$.

To show (A.9), first note that

$$\begin{aligned}\sup_{h_0\in\mathcal{H}_I}||h_0-\left(\Pi_n h_0+\frac{p^{k'_n}r}{\sqrt{n}}\right)||_{L^2} &\leq \sup_{h_0\in\mathcal{H}_I}||h_0-\Pi_n h_0||_{L^2}+||\frac{p^{k'_n}r}{\sqrt{n}}||_{L^2}\\ &\leq \frac{\delta_n}{2}+\sqrt{J}\frac{c_n}{\sqrt{n}}.\end{aligned} \tag{A.10}$$

The first inequality in (A.10) follows from the triangle inequality. And the second inequality follows from Assumption 1.3.4(ii), Assumption 1.3.5(ii), and the fact that $||r||\leq c_n$.

Also note that

$$\begin{aligned}\sup_{h_0\in\mathcal{H}_I}||\Pi_n h_0+\frac{p^{k'_n}r}{\sqrt{n}}||_s &\leq \sup_{h_0\in\mathcal{H}_I}||\Pi_n h_0||_s+||\frac{p^{k'_n}r}{\sqrt{n}}||_s\\ &\leq B-\gamma_n+\frac{c_n}{\sqrt{n}}.\end{aligned} \tag{A.11}$$

The second inequality in (A.11) follows from Assumption 1.3.5(iii) and the fact that $\|p^{k'_n}r\|_E^2\leq c_n^2$.

According to (A.10) and (A.11), as long as $\frac{c_n}{\sqrt{n}}=o(\delta_n)$ and $\frac{c_n}{\sqrt{n}}=o(\gamma_n)$, for n large enough, we have:

$$\sup_{h_0\in\mathcal{H}_I}||h_0-\left(\Pi_n h_0+\frac{p^{k'_n}r}{\sqrt{n}}\right)||_{L^2}\leq\delta_n \tag{A.12}$$

and

$$\sup_{h_0 \in \mathcal{H}_I} ||\Pi_n h_0 + \frac{p^{k'_n} r}{\sqrt{n}}||_s \leq B. \tag{A.13}$$

Such $c_n \nearrow \infty$ exists because $n^{-1/2} = o(\delta_n)$ and $n^{-1/2} = o(\gamma_n)$ by Assumption 1.3.5(iii).

Therefore, we can pick up $c_n \nearrow \infty$ s.t. $\frac{c_n}{\sqrt{n}} = o(\delta_n)$ and $\frac{c_n}{\sqrt{n}} = o(\gamma_n)$ (e.g. $c_n = \ln n$ might work). As long as $\|p^{k'_n} r\|_E^2 \leq c_n^2$, $||r||_E \leq c_n$, for $h = \Pi_n h_0 + \frac{p^{k'_n} r}{\sqrt{n}}$, (A.12) guarantees that

$$||h_0 - h||_{L^2} \leq \delta_n. \tag{A.14}$$

And (A.13) guarantees that

$$||h||_s \leq B. \tag{A.15}$$

(A.14) and (A.15) imply that $h = \Pi_n h_0 + \frac{p^{k'_n} r}{\sqrt{n}} \in B^{\delta_n}(h_0)$. So we prove (A.9). It follows that

$$\left\{ h \in \mathcal{H} : \, h = \Pi_n h_0 + \frac{p^{k'_n} r}{\sqrt{n}}, \, r \in \mathbb{R}^{k_n}, \, r' \Lambda_n r \leq c_n^2, \, ||r|| \leq c_n \right\} \subseteq B^{\delta_n}(h_0) \tag{A.16}$$

for all $h_0 \in \mathcal{H}_I$.

The first inequality in (A.17) is implied by (A.8). The equality repeats (A.6). And

the final inequality is implied by (A.16)

$$\begin{aligned}
&\inf_{h_0\in\mathcal{H}_I, r\in\mathbb{R}^{k_n}} \max_{t\in T}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i\left(t,\, h_0\right)+\frac{d\,E\left[m\left(Y,\, h_0\right)\cdot g\left(t,\, Z\right)\right]}{d\,h}\left[p^{k_n}\right]' r\right\}^2 \\
&\leq \min_{h_0\in\mathcal{H}_I}\min_{h\in B^{\delta_n}(h_0)}\max_{t\in T}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i\left(t,\, h_0\right)+\sqrt{n}\frac{d\,E\left[m\left(Y,\, h_0\right)\cdot g\left(t,\, Z\right)\right]}{d\,h}\left[h-h_0\right]\right]^2 \\
&= \min_{h\in\mathcal{H}_n}\max_{t\in T}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i\left(t,\, h\right)\right]^2+o_p\left(1\right) \\
&\leq \inf_{\substack{h_0\in\mathcal{H}_I, r\in\mathbb{R}^{k_n}\, s.t.\\ r'\Lambda_n r\leq c_n^2,\, ||r||\leq c_n}}\max_{t\in T}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i\left(t,\, h_0\right)+\frac{d\,E\left[m\left(Y,\, h_0\right)\cdot g\left(t,\, Z\right)\right]}{d\,h}\left[p^{k_n}\right]' r\right\}^2
\end{aligned} \tag{A.17}$$

Define the families of functions:

$$\begin{aligned}
V_{k_n,c}^{\alpha_0} \;=\; &\{v:T\to\mathbb{R}\, s.t.\, v(t)=\frac{\partial\rho\left(\alpha_0,\, t\right)}{\partial\theta'}\cdot r_\theta+\frac{d\,\rho\left(\alpha_0,\, t\right)}{d\,h}\left[p^{k_n}\right]' r_h,\, r\in\mathbb{R}^{d_\theta+d_{k_n}}, \\
&\|p^{k_n'}r\|_E^2\leq c^2,\, ||r||\leq c\},
\end{aligned}$$

for all $h\in\mathcal{H}_0$ and $c>0$. Consequently,

$$\begin{aligned}
\inf_{\alpha_0\in\mathcal{A}_I\cap R}\inf_{v\in V_{k_n}^{\alpha_0}}\sup_{t\in T}\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i\left(\alpha_0,\, t\right)+v\}^2 \;&\leq\; \min_{\alpha\in\mathcal{A}_n\cap R}\sup_{t\in T_n}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i\left(\alpha,\, t\right)\right]^2+o_p\left(1\right) \\
&\leq\; \inf_{\alpha_0\in\mathcal{A}_I\cap R}\inf_{v\in V_{k_n,c_n}^{\alpha_0}}\sup_{t\in T}\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i\left(\alpha_0,\, t\right)+v\}^2
\end{aligned} \tag{A.18}$$

Lemma 1 implies that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i\left(\alpha,\,t\right)\xrightarrow{\mathcal{L}}G\left(\alpha,\,t\right) \tag{A.19}$$

on $\mathcal{A}_I \times T$.

(A.19) in turn implies that

$$||\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i\left(\alpha,\,t\right)-G\left(\alpha,\,t\right)||_{\infty}=o_p\left(1\right). \tag{A.20}$$

(A.21) is implied by Lemma A4 and Theorem 1.11.1 (Extended continuous mapping) in van der Vaart and Wellner (1996) whose conditions are satisfied by (A.20).

$$\inf_{\alpha_0\in\mathcal{A}_I\cap R}\inf_{v\in V_{kn}^{\alpha_0}}\sup_{t\in T}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i\left(\alpha_0,\,t\right)+v\right\}^2\xrightarrow{\mathcal{L}}\inf_{\alpha\in\mathcal{A}_I\cap R}\inf_{v\in V_{\infty}^{\alpha}}\max_{t\in T}\left[G\left(\alpha,\,t\right)+v\left(t\right)\right]^2. \tag{A.21}$$

(A.22) is implied by Lemma A4 and Theorem 1.11.1 in van der Vaart and Wellner (1996)whose conditions are satisfied by (A.20) and Lemma A6.

$$\inf_{\alpha_0\in\mathcal{A}_I\cap R}\inf_{v\in V_{kn,cn}^{\alpha_0}}\sup_{t\in T}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i\left(\alpha_0,\,t\right)+v\right\}^2\xrightarrow{\mathcal{L}}\inf_{\alpha\in\mathcal{A}_I\cap R}\inf_{v\in V_{\infty}^{\alpha}}\max_{t\in T}\left[G\left(\alpha,\,t\right)+v\left(t\right)\right]^2. \tag{A.22}$$

(A.18), (A.21) and (A.22) imply that

$$\min_{\alpha\in\mathcal{A}_n\cap R}\sup_{t\in T_n}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i\left(\alpha,\,t\right)\right]^2\xrightarrow{\mathcal{L}}\inf_{\alpha\in\mathcal{A}_I\cap R}\inf_{v\in V_{\infty}^{\alpha}}\max_{t\in T}\left[G\left(\alpha,\,t\right)+v\left(t\right)\right]^2. \tag{A.23}$$

For other choices of weighting matrix $W_n(\alpha,\, t)$ that satisfy Assumption 1.3.2, it can be shown via similar steps that

$$\begin{aligned} & \min_{\alpha\in\mathcal{A}_n\cap R}\sup_{t\in T_n}\left[W_n^{1/2}(\alpha,\, t)\,\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_i(\alpha,\, t)\right]^2 \\ \xrightarrow{\mathcal{L}} \quad & \inf_{\alpha\in\mathcal{A}_I\cap R}\inf_{v\in V_\infty^\alpha}\max_{t\in T}\left\{W^{1/2}(\alpha,\, t)\left[G(\alpha,\, t)+v(t)\right]\right\}^2. \end{aligned} \tag{A.24}$$

■

Proof of Theorem 1.3.2

Let $\delta_n = o_p\left(n^{-1/4}\right)$ as in the proof of Theorem 1. According to Theorem 2.4.1 in van der Vaart and Wellner (1996),

$$\frac{1}{n}\sum_{i=1}^{n}m(y_i,\, \theta,\, h(x_i))\,g(t,\, z_i)\xrightarrow{a.s.}E\left[m(Y,\, \theta,\, h(X))\,g(t,\, Z)\right] \tag{A.25}$$

as a process on $\mathcal{A}\times T$. Applying the theorem of maximum and the continuous mapping theorem yields

$$\min_{\alpha\in\mathcal{A}}\sup_{t\in T}\left[\frac{1}{n}\sum_{i=1}^{n}m(y_i,\, \theta,\, h(x_i))\,g(t,\, z_i)\right]^2\xrightarrow{a.s.}\min_{\alpha\in\mathcal{A}}\sup_{t\in T}\left[E\left[m(Y,\, \theta,\, h(X))\,g(t,\, Z)\right]\right]^2. \tag{A.26}$$

(A.25) also implies that

$$\sup_{\alpha\in\mathcal{A},\, t\in T} |\frac{1}{n}\sum_{i=1}^{n} u_i\left(\alpha,\, t\right) - E\left[u\left(\alpha,\, t\right)\right]| \xrightarrow{a.s.} 0. \tag{A.27}$$

Let $\hat{\alpha}_n \in \underset{\alpha\in\mathcal{A}_n\cap R}{\arg\min}\sup_{t\in T}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} u_i\left(\alpha,\, t\right)\right]^2$ and $\Pi_n\hat{\alpha}_n = \left(\hat{\theta}_n,\, \Pi_n\hat{h}_n\right)$ where $\Pi_n\hat{h}_n$ is as in Assumption 1.3.5. In addition, let $\hat{t}_n \in \underset{t\in T}{\arg\max}\left[\frac{1}{n}\sum_{i=1}^{n} u_i\left(\Pi_n\hat{\alpha}_n,\, t\right)\right]^2$. For n large enough, we have

$$\begin{aligned}
&\min_{\alpha\in\mathcal{A}_n}\sup_{t\in T}\left[\frac{1}{n}\sum_{i=1}^{n} u_i\left(\alpha,\, t\right)\right]^2 - \min_{\alpha\in\mathcal{A}}\sup_{t\in T}\left[\frac{1}{n}\sum_{i=1}^{n} u_i\left(\alpha,\, t\right)\right]^2 \\
&\leq \left[\frac{1}{n}\sum_{i=1}^{n} u_i\left(\Pi_n\hat{\alpha}_n,\, \hat{t}_n\right)\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n} u_i\left(\hat{\alpha}_n,\, \hat{t}_n\right)\right]^2 \\
&\leq \sup_{t\in T,\, ||\alpha_1-\alpha_2||_{L^2}\leq\delta_n} |\left[\frac{1}{n}\sum_{i=1}^{n} u_i\left(\alpha_1,\, t\right)\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n} u_i\left(\alpha_2,\, t\right)\right]^2|.
\end{aligned} \tag{A.28}$$

The first inequality in (A.28) follows from

$$\min_{\alpha\in\mathcal{A}_n}\sup_{t\in T}\left[\frac{1}{n}\sum_{i=1}^{n} u_i\left(\alpha,\, t\right)\right]^2 \leq \sup_{t\in T}\left[\frac{1}{n}\sum_{i=1}^{n} u_i\left(\Pi_n\hat{\alpha}_n,\, t\right)\right]^2 = \left[\frac{1}{n}\sum_{i=1}^{n} u_i\left(\Pi_n\hat{\alpha}_n,\, \hat{t}_n\right)\right]^2 \tag{A.29}$$

and

$$\min_{\alpha\in\mathcal{A}}\sup_{t\in T}\left[\frac{1}{n}\sum_{i=1}^{n} u_i\left(\alpha,\, t\right)\right]^2 = \sup_{t\in T}\left[\frac{1}{n}\sum_{i=1}^{n} u_i\left(\hat{\alpha}_n,\, t\right)\right]^2 \geq \left[\frac{1}{n}\sum_{i=1}^{n} u_i\left(\hat{\alpha}_n,\, \hat{t}_n\right)\right]^2. \tag{A.30}$$

The second inequality in (A.28) holds for n large enough because $||\hat{h}_n - \Pi_n \hat{h}_n||_{L^2} = o(\delta_n)$ due to Assumption 1.3.5(ii).

Similarly, it can be shown that

$$\min_{\alpha \in \mathcal{A}} \sup_{t \in T} \left[\frac{1}{n} \sum_{i=1}^{n} u_i(\alpha, t) \right]^2 - \min_{\alpha \in \mathcal{A}_n} \sup_{t \in T} \left[\frac{1}{n} \sum_{i=1}^{n} u_i(\alpha, t) \right]^2$$
$$\leq \sup_{t \in T, ||\alpha_1 - \alpha_2||_{L^2} \leq \delta_n} | \left[\frac{1}{n} \sum_{i=1}^{n} u_i(\alpha_1, t) \right]^2 - \left[\frac{1}{n} \sum_{i=1}^{n} u_i(\alpha_2, t) \right]^2 |. \tag{A.31}$$

(A.28) and (A.31) imply that

$$| \min_{\alpha \in \mathcal{A}_n} \sup_{t \in T} \left[\frac{1}{n} \sum_{i=1}^{n} u_i(\alpha, t) \right]^2 - \min_{\alpha \in \mathcal{A}} \sup_{t \in T} \left[\frac{1}{n} \sum_{i=1}^{n} u_i(\alpha, t) \right]^2 |$$
$$\leq \sup_{t \in T, ||\alpha_1 - \alpha_2||_{L^2} \leq \delta_n} | \left[\frac{1}{n} \sum_{i=1}^{n} u_i(\alpha_1, t) \right]^2 - \left[\frac{1}{n} \sum_{i=1}^{n} u_i(\alpha_2, t) \right]^2 |. \tag{A.32}$$

Note that

$$
\begin{aligned}
& \left| \left[\frac{1}{n}\sum_{i=1}^{n} u_i(\alpha_1,\, t)\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n} u_i(\alpha_2,\, t)\right]^2 \right| \\
\leq\ & \left|\frac{1}{n}\sum_{i=1}^{n} u_i(\alpha_1,\, t) - \frac{1}{n}\sum_{i=1}^{n} u_i(\alpha_2,\, t)\right| \left[\left|\frac{1}{n}\sum_{i=1}^{n} u_i(\alpha_1,\, t)\right| + \left|\frac{1}{n}\sum_{i=1}^{n} u_i(\alpha_2,\, t)\right|\right] \\
\leq\ & \left|\frac{1}{n}\sum_{i=1}^{n} u_i(\alpha_1,\, t) - \frac{1}{n}\sum_{i=1}^{n} u_i(\alpha_2,\, t)\right| \cdot 2 \sup_{\alpha\in\mathcal{A},\, t\in T} \left|\frac{1}{n}\sum_{i=1}^{n} u_i(\alpha,\, t)\right| \\
\leq\ & \left|\frac{1}{n}\sum_{i=1}^{n} u_i(\alpha_1,\, t) - E\left[u(\alpha_1,\, t)\right]\right| \cdot 2 \sup_{\alpha\in\mathcal{A},\, t\in T} \left|\frac{1}{n}\sum_{i=1}^{n} u_i(\alpha,\, t)\right| \\
& + \left|\frac{1}{n}\sum_{i=1}^{n} u_i(\alpha_2,\, t) - E\left[u(\alpha_2,\, t)\right]\right| \cdot 2 \sup_{\alpha\in\mathcal{A},\, t\in T} \left|\frac{1}{n}\sum_{i=1}^{n} u_i(\alpha,\, t)\right| \\
& + \left|E\left[u(\alpha_1,\, t)\right] - E\left[u(\alpha_2,\, t)\right]\right| \cdot 2 \sup_{\alpha\in\mathcal{A},\, t\in T} \left|\frac{1}{n}\sum_{i=1}^{n} u_i(\alpha,\, t)\right| \\
\leq\ & 4 \sup_{\alpha\in\mathcal{A},\, t\in T} \left|\frac{1}{n}\sum_{i=1}^{n} u_i(\alpha,\, t) - E\left[u(\alpha,\, t)\right]\right| \cdot \sup_{\alpha\in\mathcal{A},\, t\in T} \left|\frac{1}{n}\sum_{i=1}^{n} u_i(\alpha,\, t)\right| \\
& + 2 \cdot \left|E\left[u(\alpha_1,\, t)\right] - E\left[u(\alpha_2,\, t)\right]\right| \cdot \sup_{\alpha\in\mathcal{A},\, t\in T} \left|\frac{1}{n}\sum_{i=1}^{n} u_i(\alpha,\, t)\right| \\
\leq\ & 4 \sup_{\alpha\in\mathcal{A},\, t\in T} \left|\frac{1}{n}\sum_{i=1}^{n} u_i(\alpha,\, t) - E\left[u(\alpha,\, t)\right]\right| \cdot \sup_{\alpha\in\mathcal{A},\, t\in T} \left|\frac{1}{n}\sum_{i=1}^{n} u_i(\alpha,\, t)\right| \\
& + 2 \cdot \left[\max_{y\in\mathcal{Y},\, |l|\leq B} \left|m_2^1(y,\, l)\right|\right] \cdot P \cdot ||h_1 - h_2||_{L^2} \cdot \sup_{h\in\mathcal{H},\, t\in T} \left|\frac{1}{n}\sum_{i=1}^{n} u_i(t,\, h)\right|.
\end{aligned}
\tag{A.33}
$$

(A.27) and (A.33) together imply

$$
\sup_{t\in T,\, ||\alpha_1-\alpha_2||_{L^2}\leq\delta_n} \left| \left[\frac{1}{n}\sum_{i=1}^{n} u_i(\alpha_1,\, t)\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n} u_i(\alpha_2,\, t)\right]^2 \right| \xrightarrow{a.s.} 0. \tag{A.34}
$$

(A.32) and (A.34) together imply

$$\left| \min_{\alpha \in \mathcal{A}_n \cap R} \sup_{t \in T_n} \left[\frac{1}{n} \sum_{i=1}^{n} u_i (\alpha, t) \right]^2 - \min_{\alpha \in \mathcal{A} \cap R} \sup_{t \in T} \left[\frac{1}{n} \sum_{i=1}^{n} u_i (\alpha, t) \right]^2 \right| \xrightarrow{a.s.} 0. \tag{A.35}$$

(A.26) and (A.35) imply that

$$\min_{\alpha \in \mathcal{A}_n \cap R} \sup_{t \in T_n} \left[\frac{1}{n} \sum_{i=1}^{n} u_i (\alpha, t) \right]^2 \xrightarrow{a.s.} \min_{\alpha \in \mathcal{A} \cap R} \sup_{t \in T} \left[E \left[m (Y, \theta, h (X)) g (t, Z) \right] \right]^2. \tag{A.36}$$

Assumption 1.3.2 and (A.36) imply that

$$\min_{\alpha \in \mathcal{A}_n \cap R} \sup_{t \in T_n} \left[W_n^{1/2} (\alpha, t) \frac{1}{n} \sum_{i=1}^{n} u_i (\alpha, t) \right]^2 \xrightarrow{a.s.}$$
$$\min_{\alpha \in \mathcal{A} \cap R} \sup_{t \in T} \left[W^{1/2} (\alpha, t) E \left[m (Y, \theta, h (X)) g (t, Z) \right] \right]^2,$$

which completes the proof of Theorem 1.3.2. ■

Proof of Theorem 1.3.3

Recall that the bootstrap test statistic takes the form

$$S_n^* = \min_{\alpha \in \mathcal{A}_n \cap R} \left\{ \sup_{t \in T_n} n \left[J_n^* (\alpha, t) \right]' W_n^* (\alpha, t) \left[J_n^* (\alpha, t) \right] + \lambda_n P_n (\alpha, t) \right\} \tag{A.37}$$

where

$$J_n^*(\alpha,\, t) \equiv \frac{1}{n}\sum_{i=1}^{n}\left[m\left(y_i^*,\, \theta,\, h\left(x_i^*\right)\right)\cdot g\left(t,\, z_i^*\right) - \frac{1}{n}\sum_{i=1}^{n} m\left(y_i,\, \theta,\, h\left(x_i\right)\right) g\left(t,\, z_i\right)\right]. \tag{A.38}$$

According to Lemma A1,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{m\left(y_i,\, h\left(x_i\right)\right)\cdot g\left(t,\, z_i\right) - E\left[m\left(Y,\, h\left(X\right)\right)\cdot g\left(t,\, Z\right)\right]\right\} \xrightarrow{\mathcal{L}} G\left(\alpha,\, t\right).$$

Consequently,

$$J_n^*(\alpha,\, t) \xrightarrow{\mathcal{L}^*} G\left(\alpha,\, t\right) \text{ a.s.} \tag{A.39}$$

according to Theorem 3.6.3 in van der Vaart and Wellner (1996).

It can be shown through essentially the same argument as in the proof of Lemma .17 in Santos (2011) that

$$S_n^* = \min_{\alpha\in\mathcal{A}_n\cap\mathcal{A}_I\cap R}\left\{\sup_{t\in T_n} n\left[J_n^*\left(\alpha,\, t\right)\right]' W_n^*\left(\alpha,\, t\right)\left[J_n^*\left(\alpha,\, t\right)\right]\right\} + o_{p^*}\left(1\right), \tag{A.40}$$

which, together with (A.39), imply that

$$S_n^* \Rightarrow D' \equiv \inf_{\alpha\in\mathcal{A}_I\cap R}\sup_{t\in T}\left\{G\left(\alpha,\, t\right)' W\left(\alpha,\, t\right) G\left(\alpha,\, t\right)\right\}. \tag{A.41}$$

Recall that the null asymptotic distribution of S_n is

$$D = \inf_{\alpha\in\mathcal{A}_I\cap R}\inf_{v\in V_\infty^\alpha}\sup_{t\in T}\left\{\left[G\left(\alpha,\, t\right) + v\left(t\right)\right]' W\left(\alpha,\, t\right)\left[G\left(\alpha,\, t\right) + v\left(t\right)\right]\right\}. \tag{A.42}$$

According to Definition 1.3.4, the function $v(t) \equiv 0$ belongs to V_{∞}^{α} for all $\alpha \in \mathcal{A}_I \cap R$. Therefore, D' stochastically dominates D (in a nonstrict sense), which proves the first part of Theorem 1.3.3. The second part follows from the same argument as in the proof of Theorem 1.3.2. ■

Proof of Theorem 1.3.4

Define by $C(1-\rho)$ the $(1-\rho)$th quantile of D'. Then $C_n^*(1-\rho) \xrightarrow{\text{a.s.}} C(1-\rho)$ according to Theorem 1.3.3.

The first equality in (A.43) below follows from Theorem 1.3.1. The inequality in (A.43) follows from Theorem 1.3.3.

$$\begin{aligned} & \limsup_{n\to\infty} Pr\left(S_n > C_n^*(1-\rho) \mid \mathbf{H}_0\right) \\ = \;& Pr\left(D > C(1-\rho)\right) \\ \leq \;& Pr\left(D' > C(1-\rho)\right) \\ = \;& \rho \end{aligned} \tag{A.43}$$

The first part of Theorem 3.4 follows from (A.43).

When $\mathcal{A}_I \cap R = \varnothing$, it follows from Theorem 1.3.2 that $S_n = O_p(n)$, and it follows from Theorem 3.3 that $S_n^* = O_p(\lambda_n)$. The second part of Theorem 1.3.4 follows from the

fact that $\lambda_n = o(n)$ which is guaranteed by Assumption 1.3.7. ■

Proof of Theorem 1.4.1

According to (1.17) which defines our confidence sets, for any $\theta \in \Theta_I$,

$$Pr\left(\theta \in \mathbf{CS}_n\left(1-\rho\right)\right) = Pr\left(S_n\left(\theta\right) \leq C_n^*\left(1-\rho,\, \theta\right)\right). \tag{A.44}$$

Taking $\liminf$ on both sides of (A.44) yields

$$\begin{aligned} & \liminf_{n\to\infty} Pr\left(\theta \in \mathbf{CS}_n\left(1-\rho\right)\right) \\ = \; & \liminf_{n\to\infty} Pr\left(S_n\left(\theta\right) \leq C_n^*\left(1-\rho,\, \theta\right)\right) \\ \geq \; & 1-\rho \end{aligned} \tag{A.45}$$

where the inequality follows directly from Theorem 1.3.4.

Suppose $\liminf_{n\to\infty} \inf_{\theta\in\Theta_I} Pr\left(\theta \in \mathbf{CS}_n\left(1-\rho\right)\right) < 1-\rho$. Then $\exists \epsilon > 0$ s.t.

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta_I} Pr\left(\theta \in \mathbf{CS}_n\left(1-\rho\right)\right) = 1-\rho-\epsilon. \tag{A.46}$$

Consequently, $\exists$ sequence $\{\theta_n\}_{n=1}^{\infty} \subseteq \Theta_I$ s.t.

$$\liminf_{n\to\infty} Pr\left(\theta_n \in \mathbf{CS}_n\left(1-\rho\right)\right) = 1-\rho-\epsilon. \tag{A.47}$$

According to the definition of Θ_I, $\exists\ \{h_n\}_{n=1}^{\infty} \subseteq \mathcal{H}$ s.t. $\{(\theta_n,\, h_n)\}_{n=1}^{\infty} \subseteq \mathcal{A}_I$.

Because Θ is compact under $\|\cdot\|_E$, and $\mathcal{H}$ is compact under $\|\cdot\|_c$, there is subsequence $\{(\theta_{l_n},\, h_{l_n})\}_{n=1}^{\infty}$ and $\tilde{\alpha} = \left(\tilde{\theta},\, \tilde{h}\right)$s.t. $(\theta_{l_n},\, h_{l_n}) \to \tilde{\alpha}$ under $\|\cdot\|_E + \|\cdot\|_c$, which implies that $(\theta_{l_n},\, h_{l_n}) \to \tilde{\alpha}$ under $\|\cdot\|_w$ because $\|\cdot\|_w \preceq \|\cdot\|_E + \|\cdot\|_c$. Therefore, $d_w\left(\tilde{\alpha},\, \mathcal{A}_I\right) = 0$.

It follows from Assumption 1.3.3 (iv) that

$$\max_{t\in T} \|E\left[m(Y,\, \theta,\, h\left(X\right)) \cdot g\left(t,\, Z\right)\right]\|_E \leq c_2 d_w\left(\tilde{\alpha},\, \mathcal{A}_I\right). \tag{A.48}$$

Therefore, $\tilde{\alpha} \in \mathcal{A}_I$ because $\max_{t\in T} \|E\left[m(Y,\, \theta,\, h\left(X\right)) \cdot g\left(t,\, Z\right)\right]\|_E = 0$. Consequently, $\tilde{\theta} \in \Theta_I$ and $\theta_{l_n} \to \tilde{\theta}$ under $\|\cdot\|_E$.

Lemma A1, (A.47), and stochastic equicontinuity imply that there is $N > 0$, s.t.

$$Pr\left(\tilde{\theta} \in \mathrm{CS}_n\left(1-\rho\right)\right) \leq 1 - \rho - \frac{\epsilon}{2}. \tag{A.49}$$

(A.49) implies that $\liminf_{n\to\infty} Pr\left(\tilde{\theta} \in \mathrm{CS}_n\left(1-\rho\right)\right) < 1-\rho$, which contradicts with (A.45). Therefore,

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta_I} Pr\left(\theta \in \mathrm{CS}_n\left(1-\rho\right)\right) \geq 1-\rho. \tag{A.50}$$

For any $\theta \notin \Theta_I$,

$$\begin{aligned} & \lim_{n\to\infty} Pr\left(\theta \in \mathrm{CS}_n\left(1-\rho\right)\right) \\ = & \lim_{n\to\infty} Pr\left(S_n\left(\theta\right) \leq C_n^*\left(1-\rho,\, \theta\right)\right) \\ = & \ 0 \end{aligned} \tag{A.51}$$

where the second equality follows from Theorem 1.3.4. ■

Proof of Theorem 1.4.2 follows the same reasoning and very similar steps as in the proof of Theorem 1.4.1. Therefore, it is skipped here.

Appendix B

Supplementary Appendix to Chapter 1

In this appendix we prove Lemmas A1-A6.

Proof of Lemma A1:

If Assumption 1.3.1(ii) (a) is satisfied, $g(t, z) = 1\{z \leq t\}$ for $t \in \mathcal{Z}$. $\{\{z : z \leq t\} : t \in \mathcal{Z}\}$is a a *Vapnik–Chervonenkis* (VC) class of sets. Consequently, being a family of indicator functions of the VC class of sets, $\{g(t, \cdot) : t \in \mathcal{Z}\}$ forms a *Donsker* class according to Thm 2.6.4 of van der Vaart and Wellner (1996). (Or, for a simpler exhibition, see the extension of the type I classes of functions discussed in Andrews (1994).) Also note that in this case $|g(t, \cdot)| \leq 1$. Therefore, $\{g(t, \cdot) : t \in \mathcal{Z}\}$ is a uniformly bounded *Donsker* class.

If Assumption 1.3.1(ii) (b) is satisfied, $\{g(t, \cdot) : t \in T\}$ becomes a type II class defined in Andrews (1994). According to Thm 2 in Andrews (1994) and the compactness of T,

$\{g(t,\, z) :\, t \in T\}$ is a uniformly bounded *Donsker* class.

Therefore, Assumption 1.3.1(ii) guarantees that $\{g(t,\, z) :\, t \in T\}$ forms a uniformly bounded *Donsker* class.

According to Assumption 1.3.3(i), $\{m(\cdot,\, \alpha) :\, \alpha \in \mathcal{A}\}$ is a uniformly bounded *Donsker* class.

Therefore, it follows directly from Example 2.10.8 of van der Vaart and Wellner (1996) that $\{m(\cdot,\, \alpha)\, g(t,\, \cdot) :\, (\alpha,\, t) \in \mathcal{A} \times T\}$, being the pairwise products $\{m(\cdot,\, \alpha) :\, \alpha \in \mathcal{A}\} \cdot \{g(t,\, z) :\, t \in T\}$, forms a *Donsker* class. ■

Proof of Lemma A2:

The proof of Lemma A2 uses the following lemma (stated in general terms):

Lemma A2.1. Suppose the following conditions hold: (1) $Q(\theta) \geq 0$ and $\Theta_I = \{\theta \in \Theta :\, Q(\theta) = 0\}$; (2) $\Theta_n \subseteq \Theta$ are closed and $\exists\, \Pi_n\theta \in \Theta_n$ for each $\theta \in \Theta$ s.t.$||\Pi_n\theta - \theta|| = o(1)$ and $\sigma_n \equiv \sup_{\theta_0 \in \Theta_I} ||\Pi_n\theta_0 - \theta_0|| = o(1)$; (3) $\sup_{\theta \in \Theta_n} |Q_n(\theta) - Q(\theta)| = O_p\left(n^{-1/2}\right)$; (4) $\exists$ positive constants c_1, c_2 and δ s.t. $c_1 \left(\inf_{\theta_0 \in \Theta_I} \|\theta - \theta_0\| \wedge \delta\right) \leq Q(\theta) \leq c_2 \inf_{\theta_0 \in \Theta_I} \|\theta - \theta_0\|$ for all $\theta \in \Theta$. Then, for $\hat{\theta}_n \in \arg\min_{\theta \in \Theta_n} Q_n(\theta)$ it follows that $\inf_{\theta_0 \in \Theta_0} ||\hat{\theta}_n - \theta_0|| = O_p\left(\max\left\{\sigma_n,\, n^{-1/2}\right\}\right)$.

(Proof: Assumption (3) implies that

$$\sup_{\theta \in \Theta_n} |Q_n(\theta) - Q(\theta)| \leq L \cdot n^{-1/2} \tag{B.1}$$

for some positive constant $L < \infty$.

For any fixed $\theta_0 \in \Theta_0$, $\exists$ sequence $\{\theta_{0n}\}$ with $\theta_{0n} \in \Theta_n$ s.t. $||\theta_{0n} - \theta_0|| \leq K \cdot \sigma_n$ for some positive $K < \infty$ (which does not depend on θ_0 because of Assumption (2)).

Let $\delta_n = \max\left\{\frac{4L}{c_1}, \frac{2c_2K}{c_1}\right\} \cdot \max\left\{\sigma_n, n^{-1/2}\right\}$. Let $\Theta_0^{\delta_n}$ denote an open δ_n enlargement of Θ_0 under $||\cdot||$. By Assumption (4), we have:

$\Delta_n \equiv \inf_{\theta \in \left(\Theta_0^{\delta_n}\right)^c \cap \Theta} Q(\theta) \geq c_1 \cdot \delta_n > 0$ for n large enough.

$$\begin{aligned} Q\left(\hat{\theta}_n\right) - Q(\theta_{0n}) &= \left[Q\left(\hat{\theta}_n\right) - Q_n\left(\hat{\theta}_n\right)\right] + \left[Q_n\left(\hat{\theta}_n\right) - Q_n(\theta_{0n})\right] \\ &\quad + \left[Q_n(\theta_{0n}) - Q(\theta_{0n})\right] \\ &\leqq \left[Q\left(\hat{\theta}_n\right) - Q_n\left(\hat{\theta}_n\right)\right] + \left[Q_n(\theta_{0n}) - Q(\theta_{0n})\right] \\ &\leq \left|Q\left(\hat{\theta}_n\right) - Q_n\left(\hat{\theta}_n\right)\right| + \left|Q(\theta_{0n}) - Q_n(\theta_{0n})\right| \end{aligned} \tag{B.2}$$

It follows that

$$
\begin{aligned}
& Pr\left(Q\left(\hat{\theta}_n\right) < Q\left(\theta_{0n}\right) + \frac{\Delta_n}{2}\right) \\
= \quad & Pr\left(Q\left(\hat{\theta}_n\right) - Q\left(\theta_{0n}\right) < \frac{\Delta_n}{2}\right) \\
\geq \quad & Pr\left(Q\left(\hat{\theta}_n\right) - Q\left(\theta_{0n}\right) < \frac{1}{2}c_1\delta_n\right) \\
\geq \quad & Pr\left(|Q\left(\hat{\theta}_n\right) - Q_n\left(\hat{\theta}_n\right)| + |Q\left(\theta_{0n}\right) - Q_n\left(\theta_{0n}\right)| < 2L \cdot n^{-1/2}\right) \\
\longrightarrow \quad & 1 \qquad \text{(B.3)}
\end{aligned}
$$

where the last step is implied by (B.1).

By Assumption (4) and $||\theta_{0n} - \theta_0|| = O\left(\delta_n\right)$, we have $Q\left(\theta_{0n}\right) \leq c_2.||\theta_{0n} - \theta_0|| < c_2 \cdot K \cdot \sigma_n < \frac{\Delta_n}{2}$ for n large enough. Therefore:

$$
Pr\left(Q\left(\hat{\theta}_n\right) < \Delta_n\right) \longrightarrow 1. \qquad \text{(B.4)}
$$

This implies that $Pr\left(d\left(\hat{\theta}_n,\, \Theta_0\right) \leq \delta_n\right) \longrightarrow 1$. Therefore, $\inf_{\theta_0 \in \Theta_0} ||\hat{\theta}_n - \theta_0|| \leq O_p\left(\delta_n\right) = O_p\left(\max\left\{\sigma_n,\, n^{-1/2}\right\}\right)$, which completes our proof.)

Define

$$
Q\left(\alpha\right) = \sup_{t \in T} \| W^{1/2}\left(\alpha,\, t\right) \cdot E\left[m\left(Y,\, \theta\,, h\left(X\right)\right) \cdot g\left(t,\, Z\right)\right] \|_E,
$$

and

$$
Q_n\left(\alpha\right) = \sup_{t \in T} \| W_n^{1/2}\left(\alpha,\, t\right) \cdot \frac{1}{n} \sum_{i=1}^{n} m\left(y_i,\, \theta,\, h\left(x_i\right)\right) \cdot g\left(t,\, z_i\right) \|_E.
$$

Obviously, $\arg\min_{\alpha\in\mathcal{A}_n\cap R}\left\{\sup_{t\in T} n\left[J_n\left(\alpha,\, t\right)\right]' W_n\left(\alpha,\, t\right)\left[J_n\left(\alpha,\, t\right)\right]\right\} = \arg\min_{\alpha\in\mathcal{A}_n\cap R} Q_n\left(\alpha\right)$.

Assumption 1.3.1(ii) implies condition (1) in Lemma A2.1 holds for $\|\cdot\|_w$.

Assumption 1.3.4 (iii) and (iv) implies condition (2) holds for $\mathcal{A}_n$.

Lemma A1 implies that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{m\left(y_i,\,\theta,\,h\left(x_i\right)\right)\cdot g\left(t,\,z_i\right) - E\left[m\left(Y,\,\theta,\,h\left(X\right)\right)\cdot g\left(t,\,Z\right)\right]\right\} \xrightarrow{\mathcal{L}} G\left(\alpha,\,t\right) \tag{B.5}$$

on $T\times\mathcal{H}$.

Consequently,

$$\begin{aligned}
& \sqrt{n}\left|Q_n\left(\alpha\right) - Q\left(\alpha\right)\right| \\
= \;& \sqrt{n}\Big|\sup_{t\in T}\|W_n^{1/2}\left(\alpha,\,t\right)\frac{1}{n}\sum_{i=1}^{n} m\left(y_i,\,\theta,\,h\left(x_i\right)\right)\cdot g\left(t,\,z_i\right)\|_E \\
& -\sup_{t\in T}\|W^{1/2}\left(\alpha,\,t\right)E\left[m\left(Y,\,\theta,\,h\left(X\right)\right)\cdot g\left(t,\,Z\right)\right]\|_E\Big| \\
\leq \;& \sqrt{n}\sup_{t\in T}\|W_n^{1/2}\left(\alpha,\,t\right)\frac{1}{n}\sum_{i=1}^{n} m\left(y_i,\,\theta,\,h\left(x_i\right)\right)\cdot g\left(t,\,z_i\right) \\
& -W^{1/2}\left(\alpha,\,t\right)E\left[m\left(Y,\,\theta,\,h\left(X\right)\right)\cdot g\left(t,\,Z\right)\right]\|_E \\
\leq \;& \sqrt{n}\sup_{t\in T}\|W^{1/2}\left(\alpha,\,t\right)\{\frac{1}{n}\sum_{i=1}^{n} m\left(y_i,\,\theta,\,h\left(x_i\right)\right)\cdot g\left(t,\,z_i\right) \\
& -E\left[m\left(Y,\,\theta,\,h\left(X\right)\right)\cdot g\left(t,\,Z\right)\right]\}\|_E \\
& +\sqrt{n}\sup_{t\in T}\|\left[W_n^{1/2}\left(\alpha,\,t\right) - W^{1/2}\left(\alpha,\,t\right)\right]\cdot\frac{1}{n}\sum_{i=1}^{n} m\left(y_i,\,\theta,\,h\left(x_i\right)\right)\cdot g\left(t,\,z_i\right)\|_E \\
\leq \;& \sqrt{n}\overline{W}^{1/2}\sup_{t\in T}\|\frac{1}{n}\sum_{i=1}^{n} m\left(y_i,\,\theta,\,h\left(x_i\right)\right)\cdot g\left(t,\,z_i\right) - E\left[m\left(Y,\,\theta,\,h\left(X\right)\right)\cdot g\left(t,\,Z\right)\right]\|_E \\
& +I\cdot\sup_{t\in T}\|\frac{1}{n}\sum_{i=1}^{n} m\left(y_i,\,\theta,\,h\left(x_i\right)\right)\cdot g\left(t,\,z_i\right)\|_E \quad \text{for some } I>0
\end{aligned}$$

$$\xrightarrow{\mathcal{L}} \quad \overline{W}^{1/2} \cdot \sup_{t \in T} \| G(t,\, \alpha) \|_E + I \cdot \sup_{t \in T} \| E\left[m\left(Y,\, \theta,\, h\left(X\right)\right) \cdot g\left(t,\, Z\right) \right] \|_E \tag{B.6}$$

Therefore, w.p.$\to 1$,

$$\begin{aligned} & \sqrt{n} \sup_{\alpha \in \mathcal{A}_n \cap R} \left| Q_n(\alpha) - Q(\alpha) \right| \\ \leq \quad & \sup_{\alpha \in \mathcal{A}_n \cap R} \sup_{t \in T} \| G(t,\, \alpha) \| + I \cdot \sup_{\alpha \in \mathcal{A}_n \cap R} \cdot \sup_{t \in T} \| E\left[m\left(Y,\, \theta,\, h\left(X\right)\right) \cdot g\left(t,\, Z\right) \right] \|_E \end{aligned} \tag{B.7}$$

(B.7) implies that $\sqrt{n} \sup_{\alpha \in \mathcal{A}_n \cap R} \left| Q_n(\alpha) - Q(\alpha) \right| = O_p(1)$, which implies condition (3) in Lemma A2.1.

Assumption 1.3.3 (iv) states that condition (4) holds.

With condition (1) - (4) of Lemma A2.1 being verified, Lemma 2 is valid according to Lemma A2.1. ■

Proof of Lemma A3:

For any $\alpha_0 \in \mathcal{A}_I$,

$$\begin{aligned} & \|\hat{\alpha}_n - \alpha_0\|_{L_2} \\ \leq\ & \|\hat{\alpha}_n - P(\alpha_0)\|_{L_2} + \|P(\alpha_0) - \alpha_0\|_{L_2} \\ \leq\ & \psi_n \cdot \|\hat{\alpha}_n - P(\alpha_0)\|_w + \delta_{s,n} \\ \leq\ & \psi_n \cdot (\|\hat{\alpha}_n - \alpha_0\|_w + \|\alpha_0 - P(\alpha_0)\|_w) + \delta_{s,n} \\ \leq\ & \psi_n \cdot (\|\hat{\alpha}_n - \alpha_0\|_w + \delta_{w,n}). \end{aligned} \tag{B.8}$$

According to (B.8),

$$\|\hat{\alpha}_n - \alpha_0\|_{L_2} \leq \psi_n \cdot (\|\hat{\alpha}_n - \alpha_0\|_w + \delta_{w,n}). \tag{B.9}$$

Taking $\inf_{\alpha \in \mathcal{A}_n \cap R}$ on both sides of (B.9) yields

$$d_{L_2}(\hat{\alpha}_n, \mathcal{A}_I \cap R) \leq \psi_n \cdot (d_w(\hat{\alpha}_n, \mathcal{A}_I \cap R) + \delta_{w,n}), \tag{B.10}$$

which, together with Lemma A2 implies lemma A3. ■

Proof of Lemma A4:

The triangle inequality implies that :

$$
\begin{aligned}
& |\inf_{b\in B}\inf_{f\in\mathcal{F}_{\infty}^{b}}\sup_{a\in A}|g\left(a,\,b\right)-f\left(a\right)|-\inf_{b\in B}\inf_{f\in\mathcal{F}_{n}^{b}}\sup_{a\in A}|g_{n}\left(a,\,b\right)-f\left(a\right)||\\
\leq\quad & |\inf_{b\in B}\inf_{f\in\mathcal{F}_{\infty}^{b}}\sup_{a\in A}|g\left(a,\,b\right)-f\left(a\right)|-\inf_{b\in B}\inf_{f\in\mathcal{F}_{n}^{b}}\sup_{a\in A}|g\left(a,\,b\right)-f\left(a\right)||\\
+\quad & |\inf_{b\in B}\inf_{f\in\mathcal{F}_{n}^{b}}\sup_{a\in A}|g\left(a,\,b\right)-f\left(a\right)|-\inf_{b\in B}\inf_{f\in\mathcal{F}_{n}^{b}}\sup_{a\in A}|g_{n}\left(a,\,b\right)-f\left(a\right)||\quad . \qquad \text{(B.11)}
\end{aligned}
$$

Fix $\varepsilon>0$ and there exists $\left\{\hat{f}_{b,\infty}\right\}_{b\in B}$ with $\hat{f}_{b,\infty}\in\mathcal{F}_{\infty}^{b}$ for all $b\in B$ s.t.:

$$
\inf_{b\in B}\inf_{f\in\mathcal{F}_{\infty}^{b}}\sup_{a\in A}|g\left(a,\,b\right)-f\left(a\right)|\geq\inf_{b\in B}\sup_{a\in A}|g\left(a,\,b\right)-\hat{f}_{b,\infty}|-\varepsilon.
$$

Also there exits $\hat{f}_{b,n}\in\mathcal{F}_{n}^{b}$ s.t. $\inf_{f\in\mathcal{F}_{n}^{b}}||f-\hat{f}_{b,\infty}||_{\infty}\geq||\hat{f}_{b,n}-\hat{f}_{b,\infty}||_{\infty}-\varepsilon$, for each $b\in B$ and $n=1,2,3,\ldots$.

Finally, there exits $\tilde{b}\in B$ s.t.

$$
\inf_{b\in B}\sup_{a\in A}|g\left(a,\,b\right)-\hat{f}_{b,\infty}\left(a\right)|\geq\sup_{a\in A}|g\left(a,\,\tilde{b}\right)-\hat{f}_{\tilde{b},\infty}\left(a\right)|-\varepsilon.
$$

By the definition of $\mathcal{F}_{\infty}^{b}$, $\inf_{f\in\mathcal{F}_{n}^{b}}||f-\hat{f}_{b,\infty}||_{\infty}=o\left(1\right)$, for each $b\in B$.

Therefore,

$$
\begin{aligned}
& \inf_{b\in B}\inf_{f\in\mathcal{F}_n^b}\sup_{a\in A}|g\left(a,\,b\right)-f\left(a\right)| \\
\leq\quad & \inf_{b\in B}\sup_{a\in A}|g\left(a,\,b\right)-\hat{f}_{b,n}\left(a\right)| \\
\leq\quad & \inf_{b\in B}\left\{\sup_{a\in A}|g\left(a,\,b\right)-\hat{f}_{b,\infty}\left(a\right)|+\sup_{a\in A}|\hat{f}_{b,\infty}\left(a\right)-\hat{f}_{b,n}\left(a\right)|\right\} \\
=\quad & \inf_{b\in B}\left\{\sup_{a\in A}|g\left(a,\,b\right)-\hat{f}_{b,\infty}\left(a\right)|+||\hat{f}_{b,\infty}-\hat{f}_{b,n}||_{\infty}\right\} \\
\leq\quad & \inf_{b\in B}\left\{\sup_{a\in A}|g\left(a,\,b\right)-\hat{f}_{b,\infty}\left(a\right)|+\inf_{f\in\mathcal{F}_n^b}||f-\hat{f}_{b,\infty}||_{\infty}+\varepsilon\right\} \\
\leq\quad & \sup_{a\in A}|g\left(a,\,\tilde{b}\right)-\hat{f}_{\tilde{b},\infty}\left(a\right)|+\inf_{f\in\mathcal{F}_n^{\tilde{b}}}||f-\hat{f}_{\tilde{b},\infty}||_{\infty}+\varepsilon \\
\leq\quad & \inf_{b\in B}\sup_{a\in A}|g\left(a,\,b\right)-\hat{f}_{b,\infty}\left(a\right)|+o\left(1\right)+2\varepsilon \\
\leq\quad & \inf_{b\in B}\inf_{f\in\mathcal{F}_\infty^b}\sup_{a\in A}|g\left(a,\,b\right)-f\left(a\right)|+o\left(1\right)+3\varepsilon
\end{aligned}
\tag{B.12}
$$

Since ε is arbitrary, we have:

$$
\begin{aligned}
\inf_{b\in B}\inf_{f\in\mathcal{F}_\infty^b}\sup_{a\in A}|g\left(a,\,b\right)-f\left(a\right)| \quad &\leq \quad \inf_{b\in B}\inf_{f\in\mathcal{F}_n^b}\sup_{a\in A}|g\left(a,\,b\right)-f\left(a\right)| \\
&\leq \quad \inf_{b\in B}\inf_{f\in\mathcal{F}_\infty^b}\sup_{a\in A}|g\left(a,\,b\right)-f\left(a\right)|+o\left(1\right)
\end{aligned}
\tag{B.13}
$$

where the first inequality follows from $\mathcal{F}_n^b\subseteq\mathcal{F}_\infty^b$.

(B.13) implies that,

$$\inf_{b\in B}\inf_{f\in\mathcal{F}_n^b}\sup_{a\in A}|g\left(a,\,b\right)-f\left(a\right)|=\inf_{b\in B}\inf_{f\in\mathcal{F}_\infty^b}\sup_{a\in A}|g\left(a,\,b\right)-f\left(a\right)|+o\left(1\right). \tag{B.14}$$

Next, $||g_n-g||_\infty=o\left(1\right)$ implies that:

$$\begin{aligned}
&\inf_{b\in B}\inf_{f\in\mathcal{F}_n^b}\sup_{a\in A}|g\left(a,\,b\right)-f\left(a\right)|\\
\leq\;&\inf_{b\in B}\inf_{f\in\mathcal{F}_n^b}\left[\sup_{a\in A}|g\left(a,\,b\right)-g_n\left(a,\,b\right)|+\sup_{a\in A}|g_n\left(a,\,b\right)-f\left(a\right)|\right]\\
\leq\;&o\left(1\right)+\inf_{b\in B}\inf_{f\in\mathcal{F}_n^b}\sup_{a\in A}|g_n\left(a,\,b\right)-f\left(a\right)|
\end{aligned} \tag{B.15}$$

Similar, we can show that:

$$\begin{aligned}
&\inf_{b\in B}\inf_{f\in\mathcal{F}_n^b}\sup_{a\in A}|g_n\left(a,\,b\right)-f\left(a\right)|\\
\leq\;&\inf_{b\in B}\inf_{f\in\mathcal{F}_n^b}\sup_{a\in A}|g\left(a,\,b\right)-f\left(a\right)|+o\left(1\right)
\end{aligned} \tag{B.16}$$

Therefore, by (B.15) and (B.16), we have

$$\inf_{b\in B}\inf_{f\in\mathcal{F}_n^b}\sup_{a\in A}|g\left(a,\,b\right)-f\left(a\right)|=\inf_{b\in B}\inf_{f\in\mathcal{F}_n^b}\sup_{a\in A}|g_n\left(a,\,b\right)-f\left(a\right)|+o\left(1\right). \tag{B.17}$$

(B.11), (B.14), and (B.17) complete the proof. ■

Proof of Lemma A5:

Fix $\varepsilon_n \searrow 0$. There exists a sequence $\{\hat{a}_n\}$ s.t. $\hat{a}_n \in A_n$ and the following inequality holds:

$$\inf_{a\in A_n}\max_{b\in B} F_n^2(a,\, b) \geq \max_{b\in B} F_n^2(\hat{a}_n,\, b) + \varepsilon_n. \tag{B.18}$$

Then step 5 in (B.19) follows from (B.18) and $\sup_{a\in A_n}\sup_{b\in B}\left[G_n(a,\, b) - F_n(a,\, b)\right]^2 = o_p(1)$:

$$\begin{aligned}
&\inf_{a\in A_n}\max_{b\in B} G_n^2(a,\, b) \\
&= \inf_{a\in A_n}\max_{b\in B}\left\{\left[G_n(a,\, b) - F_n(a,\, b)\right] + F_n(a,\, b)\right\}^2 \\
&\leq \inf_{a\in A_n}\max_{b\in B}\left\{\left[G_n(a,\, b) - F_n(a,\, b)\right]^2 + F_n^2(a,\, b) + 2|F_n(a,\, b)| \cdot |G_n(a,\, b) - F_n(a,\, b)|\right\} \\
&\leq \max_{b\in B}\left\{\left[G_n(\hat{a}_n,\, b) - F_n(\hat{a}_n,\, b)\right]^2 + F_n^2(\hat{a}_n,\, b) + 2|F_n(\hat{a}_n,\, b)| \cdot |G_n(\hat{a}_n,\, b) - F_n(\hat{a}_n,\, b)|\right\} \\
&\leq \max_{b\in B}\left[G_n(\hat{a}_n,\, b) - F_n(\hat{a}_n,\, b)\right]^2 + \max_{b\in B} F_n^2(\hat{a}_n,\, b) \\
&+2\max_{b\in B}|F_n(\hat{a}_n,\, b)| \cdot \max_{b\in B}|G_n(\hat{a}_n,\, b) - F_n(\hat{a}_n,\, b)| \\
&\leq o_p(1) + \inf_{a\in A_n}\max_{b\in B} F_n^2(a,\, b) - \varepsilon_n + 2\sqrt{\inf_{a\in A_n}\max_{b\in B} F_n^2(a,\, b) - \varepsilon_n}\cdot o_p(1) \\
&= \inf_{a\in A_n}\max_{b\in B} F_n^2(a,\, b) + o_p(1)
\end{aligned} \tag{B.19}$$

Similarly, it can be shown that

$$\inf_{a\in A_n}\max_{b\in B} F_n^2(a,\, b) = \inf_{a\in A_n}\max_{b\in B} G_n^2(a,\, b) + o_p(1). \tag{B.20}$$

(B.19) and (B.20) together complete the proof. ■

Proof of Lemma A6:

According to Definition 1.3.4, $V^{\alpha_0}_{k_n,c_n} \subseteq V^{\alpha_0}_{k_n}$. Therefore

$$\cup V^{\alpha_0}_{k_n,c_n} \subseteq \cup V^{\alpha_0}_{k_n}. \tag{B.21}$$

Next, for any positive integer l and $v \in V^{\alpha_0}_{k_l}$, there is $r \in \mathcal{H}_{lR}$ s.t.

$$v(t) = \frac{\partial \rho\,(\alpha_0,\, t)}{\partial \theta'} \cdot r_\theta + \frac{d\,\rho\,(\alpha_0,\, t)}{d\,h} \left[p^{k_l}\right]' r_h.$$

Since $c_n \to \infty$, there is $N > n$ s.t.

$$c_N \geq max \left\{ \|r'_h p^{k_n}\|^2_s \,,\, ||r||_E \right\}. \tag{B.22}$$

Let $r_N \in \mathcal{H}_{NR}$ be $r'_N = (r', 0, ..., 0)$. Then (B.22) implies that

$$v \in \cup V^{\alpha_0}_{k_N,c_N}. \tag{B.23}$$

Since (B.23) holds for any $v \in V^{\alpha_0}_{k_l}$, we have

$$V^{\alpha_0}_{k_l} \subseteq V^{\alpha_0}_{k_N,c_N}. \tag{B.24}$$

Since for any positive integer l there is an appropriate N s.t. (B.24) holds, we have

$$\cup V_{k_n}^{\alpha_0} \subseteq \cup V_{k_n,c_n}^{\alpha_0}. \tag{B.25}$$

(B.21) and (B.25) implies that $\cup V_{k_n,c_n}^{\alpha_0} = \cup V_{k_n}^{\alpha_0}$. ■

Proof of Assumption 1.3.3(ii) Being Satisfied When $m(\cdot)$ Is Pointwise Lipschitz Continuously in α and Continuous in y with Compact $\mathcal{X} \times \mathcal{Y}$:

The following lemmas are used in the proof:

Lemma B1 Let Assumption 2.1(ii). There exists a constant $K > 0$ s.t. for all ε sufficiently small

$$\log N(\mathcal{H}, ||\cdot||_\infty, \varepsilon) \leq K\left(\frac{1}{\varepsilon}\right)^{\frac{d_x}{m}}.$$

(Proof: Lemma B1 is the same as the Lemma .3 in Santos (2011). And a proof is given there.)

Lemma B2 Let $\mathcal{F} \equiv \{f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} : f(x, y) = m(y, \theta, h(x)) \text{ for some } (\theta, h) \in \mathcal{A}\}$. There exist $B' < \infty$ s.t. $F(y) \equiv \max_{\theta\in\Theta, \|l\|_E \leq B'} \|m(y, \theta, l)\|_E$ is an envelop for $\mathcal{F}$. More-

over, there exists constants $K_0,\, K > 0$ s.t. for all norm with $||F|| < \infty$ and ε sufficiently small:

$$N_{[]}\left(\mathcal{F},\, ||\cdot||,\, \varepsilon||F||\right) \leq K_0 \cdot \left(\frac{\text{diam}\Theta}{\varepsilon}\right)^{d_\theta} \cdot \exp\left[K \cdot \left(\frac{4}{\varepsilon}\right)\frac{d_x}{m}\right].$$

(Proof: The compactness of $\mathcal{H}$, which is implied by Assumption 1.2.1(ii), implies that for any $h \in \mathcal{H}$ there exist $B' < \infty$ s.t. $||h||_c \leq B'$. In turn, $||h||_\infty \leq ||h||_c \leq B'$ for any $h \in \mathcal{H}$. Therefore, for any $f\left(x,\, y\right) = m\left(y,\, \theta,\, h\left(x\right)\right) \in \mathcal{F}$,

$$\begin{aligned} \|f\left(x,\, y\right)\|_E &\leq \max_{\theta\in\Theta, \|l\|_E \leq B'} \|m(y,\, \theta,\, l)\|_E \\ &= F\left(y\right). \end{aligned}$$

The pointwise Lipschitz continuity of m implies

$$\begin{aligned} & \|m\left(y,\, \theta_1,\, h_1\left(x\right)\right) - m\left(y,\, \theta_2,\, h_2\left(x\right)\right)\|_E \\ \leq\; & M\left(y\right) \|\left(\theta_1,\, h_1\left(x\right)\right)' - \left(\theta_2,\, h_2\left(x\right)\right)'\|_E \\ \leq\; & M\left(y\right)\left(\|\theta_1 - \theta_2\|_E + \|h_1 - h_2\|_\infty\right). \end{aligned} \tag{B.26}$$

(B.26) shows that the class $\mathcal{F}$ is Lipschitz in $\mathcal{A}$ w.r.t. the norm $\|\cdot\|_E + \|\cdot\|_\infty$, which in turn implies the first inequality in (B.27) below. For ε small enough, and some $K' < \infty$, the third inequality in (B.27) follows from Lemma B1 and the compactness of Θ.

$$
\begin{aligned}
N_{[]}\left(\mathcal{F},\, ||\cdot||,\, \varepsilon||F||\right) &\leq N\left(\mathcal{A},\, \|\cdot\|_E + \|\cdot\|_\infty,\, \frac{\varepsilon}{2}\right) \\
&\leq N\left(\Theta,\, ||\cdot||_E,\, \frac{\varepsilon}{4}\right) \cdot N\left(\mathcal{H},\, ||\cdot||_\infty,\, \frac{\varepsilon}{4}\right) \\
&\leq \left(\frac{K' \cdot \mathrm{diam}\Theta}{\frac{\varepsilon}{4}}\right)^{d_\theta} \cdot \exp\left(K \cdot \left(\frac{4}{\varepsilon}\right)^{\frac{d_x}{m}}\right) \\
&= K_0 \cdot \left(\frac{\mathrm{diam}\Theta}{\varepsilon}\right)^{d_\theta} \cdot \exp\left[K \cdot \left(\frac{4}{\varepsilon}\right)^{\frac{d_x}{m}}\right].
\end{aligned}
\tag{B.27}
$$

Setting $K_0 = (4K')^{d_\theta}$ justifies the final step in (B.27), which completes the proof of Lemma B2.)

Equipped with Lemma B1 and Lemma B2, the proof proceeds as follows:

The boundedness of $\mathcal{A}$ w.r.t. $\|\cdot\|_E + \|\cdot\|_\infty$, the Lipschitz continuity of m in α, the compactness of $\mathcal{Y}$, and the continuity of m in y together imply the uniform boundedness of m, which in turn implies that, for any $(\theta,\, h) \in \mathcal{A}$,

$$
E\left[\|m\left(Y,\, \theta,\, h\left(X\right)\right)\|_E^2\right] < \infty. \tag{B.28}
$$

Therefore, the central limit theorem implies the convergence in distribution pointwise on $\mathcal{A}$. To verify uniform asymptotic equicontinuity, let $\mathcal{F}$ and $F\left(y\right)$ be as in Lemma

B2. $F(\cdot)$ is continuous according to the theorem of maximum. Then compactness of $\mathcal{Y}$ implies boundedness of $F(\cdot)$, which implies that

$$||F||_{L^2}^2 = E\left[\left[F(Y)\right]^2\right] < \infty. \tag{B.29}$$

(B.29) guarantees that

$$N_{[]}(\mathcal{F}, ||\cdot||_{L^2}, \varepsilon) = 1, \tag{B.30}$$

for some $D > 0$ and any $\varepsilon \geq D$.

(B.30) implies the first step in (B.31) below. The second step in (B.31) is implied by Lemma B2 whose condition is satisfied by (B.29). The change of variable $u = \dfrac{\varepsilon}{||F||_{L^2}}$ yields the third step. The final step is justified by $\dfrac{d_x}{m} < 2$, which is implied by Assumption 1.2.1(ii), and that $q^{\frac{d_x}{m}} > \log q$ for q large enough.

$$\begin{aligned}
& \int_0^\infty \sqrt{\log N_{[]}(\mathcal{F}, ||\cdot||_{L^2}, \varepsilon)}\, d\varepsilon \\
= & \int_0^D \sqrt{\log N_{[]}(\mathcal{F}, ||\cdot||_{L^2}, \varepsilon)}\, d\varepsilon \\
\leq & \int_0^D \sqrt{K \cdot \left(\frac{4||F||_{L^2}}{\varepsilon}\right)^{\frac{d_x}{m}} + \log K_0 + d_\theta \log(\text{diam}\Theta) - d_\theta \log \frac{\varepsilon}{||F||_{L^2}}}\, d\varepsilon \\
= & \int_0^{\frac{D}{||F||_{L^2}}} \sqrt{K \cdot \left(\frac{4}{u}\right)^{\frac{d_x}{m}} - d_\theta \log u + \log K_0 + d_\theta \log(\text{diam}\Theta)}\, du \\
\leq & \infty.
\end{aligned} \tag{B.31}$$

According to Theorem 2.5.6 in van der Vaart and Wellner (1996), whose condition is satisfied by (B.31), $\mathcal{F}$ is a Donsker class. ■

Appendix C

Proofs of the Theorems in Chapter 2

Proof of Lemma 2.3.1

Proof of Part (1):

By definition,

$$E_n\left[m_j\left(\theta\right)\right]=\pi\left(A_j;\,\theta\right)-\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left(y_i=A_j\right),$$

$$E_p\left[m_j\left(\theta\right)\right]=\pi\left(A_j;\,\theta\right)-E_P\left[\mathbf{1}\left(y_i=A_j\right)\right].$$

Therefore,

$$E_n\left[m_j\left(\theta\right)\right]-E_p\left[m_j\left(\theta\right)\right]=(-\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left(y_i=A_j\right)+E_p\left[\mathbf{1}\left(y_i=A_j\right)\right]$$

is not a function of θ. Neither does

$$E_n[m(\theta)] - E_p[m(\theta)] = (E_n[m_1(\theta)] - E_p[m_1(\theta)], ..., E_n[m_J(\theta)] - E_p[m_J(\theta)])'.$$

By the multivariate central limit theorem, $\sqrt{n}(E_n[m(\theta)] - E_p[m(\theta)]) \to_d \Delta$ uniformly on Θ, where $\Delta \sim N(\mathbf{0}, \Sigma)$, $\Sigma_{ii} = P(A_i)[1 - P(A_i)]$ and $\Sigma_{ij} = -P(A_i)P(A_j)$ when $i \neq j$.

Proof of Part (2):

Since $E_p[m(\theta)] = \mathbf{0}$ for all $\theta \in \Theta_I$, $\sqrt{n}E_n[m(\theta)] = \sqrt{n}(E_n[m(\theta)] - E_p[m(\theta)])$ for all $\theta \in \Theta_I$. Moreover, $\sqrt{n}E_n[m(\theta)]$ does not change over θ on Θ_I, and $\sqrt{n}E_n[m(\theta)] \to_d \Delta$ on Θ_I.

By continuous mapping theorem, $nQ_n(\theta) = ||\sqrt{n}E_n[m(\theta)]||^2 \to_d ||\Delta||^2$ on Θ_I. Hence $\mathcal{C}_n \to_d ||\Delta||^2$ on Θ_I.

$\forall b_n \to \infty$, $\forall \varepsilon > 0$, $\exists L > 0$, s.t. $Pr\left(||\Delta||^2 \leq L\right) > 1 - \varepsilon$.

$\exists N > 0$, s.t. $b_n > L$ for all $n \geq N$.

Consequently, $Pr(\mathcal{C}_n \leq b_n) \geq Pr(\mathcal{C}_n \leq L)$ for all $n \geq N$.

Therefore, $\liminf Pr(\mathcal{C}_n \leq b_n) \geq \liminf Pr(\mathcal{C}_n \leq L) = Pr\left(||\Delta||^2 \leq L\right) > 1 - \varepsilon$.

Since the above is true for arbitrary ε, $\liminf Pr(\mathcal{C}_n \leq b_n) \geq 1$.

Since $\limsup Pr(\mathcal{C}_n \leq b_n) \leq 1$, $\lim Pr(\mathcal{C}_n \leq b_n) = 1$. ■

Q.E.D.

Proof of Lemma 2.3.2

Proof:

(a) By equation (3), $f_t(y_{it}|y_{i\,t-1},\,\alpha_i;\,\gamma)$ is continuous w.r.t. γ and α_i. Consequently, by equation (5), $\pi\left(A_j;\,\theta\right)$ is continuous w.r.t. θ.

$\pi\left(A_j;\,\theta\right) - E_p\left[\mathbf{1}\left(y_i = A_j\right)\right]$ being continuous w.r.t. θ is immediate. So does $\pi\left(A_j;\,\theta\right) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left(y_i = A_j\right)$ being continuous w.r.t. θ.

The above is true for all $j = 1, ..., J$.

Therefore, by definition, Q and Q_n are continuous w.r.t. θ.

(b) Since $\sqrt{n}\left(E_n\left[m\left(\theta\right)\right] - E_p\left[m\left(\theta\right)\right]\right) \to_d \Delta$ on Θ,

$$\sqrt{n}\left(E_n\left[m\left(\theta\right)\right] - E_p\left[m\left(\theta\right)\right]\right) = O_p\left(1\right).$$

Hence $\left(E_n\left[m\left(\theta\right)\right] - E_p\left[m\left(\theta\right)\right]\right) = O_p\left(1/\sqrt{n}\right)$.[1]

Since $0 \leq \pi\left(A_j;\,\theta\right) \leq 1$, $0 \leq \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left(y_i = A_j\right) \leq 1$, and $0 \leq E_P\left[\mathbf{1}\left(y_i = A_j\right)\right] \leq 1$ for

[1] For a sequence of random vector $X_n = (X_{n1},\, ...\, X_{nJ})$, $X_n = O_p\left(1\right)$ means that $X_{nj} = O_p\left(1\right)$ for each $j = 1,\, ...,\, J$.

each $\{A_j\}$, $|E_n[m_j(\theta)]| \leq 1$ and $|E_p[m_j(\theta)]| \leq 1$ on Θ for each $j = 1, ..., J$. Consequently,

$$|E_n[m(\theta)] + E_p[m(\theta)]| \leq 2 \cdot \mathbf{1}.$$

Therefore

$$\begin{aligned} Q_n(\theta) - Q_p(\theta) &= (E_n[m(\theta)] - E_p[m(\theta)])'(E_n[m(\theta)] + E_p[m(\theta)]) \\ &= O_p\left(1/\sqrt{n}\right) \end{aligned}$$

on Θ. Recall that $E_n[m(\theta)] - E_p[m(\theta)]$ does not change over θ on Θ.

So

$$\begin{aligned} sup_\Theta(Q_n - Q) &= sup_\Theta\left\{(E_n[m(\theta)] - E_p[m(\theta)])'(E_n[m(\theta)] + E_p[m(\theta)])\right\} \\ &\leq 2sup_\Theta||E_n[m(\theta)] - E_p[m(\theta)]|| \\ &= 2||E_n[m(\theta)] - E_p[m(\theta)]|| \\ &= O_p\left(1/\sqrt{n}\right) \end{aligned}$$

Similarly, $sup_\Theta(Q - Q_n) = O_p(1/\sqrt{n})$.

Therefore, $sup_\Theta|Q - Q_n| = O_p(1/\sqrt{n})$

(c) Since $nQ_n(\theta) = ||\sqrt{n}E_n[m(\theta)]||^2 \rightarrow_d ||\Delta||^2$ on Θ_I, $nQ_n(\theta) = O_p(1)$ on Θ_I. Consequently, $Q_n(\theta) = O_p(1/n)$ on Θ_I.

Recall that $Q_n(\theta)$ does not change over θ on Θ_I. So $sup_{\Theta_I} Q_n = O_p(1/n)$. ■

Proof of Theorem 2.3.1

Define the $\epsilon-$expansion of Θ_I in $\Theta$by $\Theta_I^\epsilon = \{\theta \in \Theta : d(\theta, \Theta_I) \leq \epsilon\}$.

Proof:

Let $\widehat{\Theta}_I = C_n(b_n)$. Then $\Theta_I \subseteq \widehat{\Theta}_I \iff \mathcal{C}_n \leq b_n$.Therefore:

$$Pr\left(\Theta_I \subseteq \widehat{\Theta}_I\right) = Pr(\mathcal{C}_n \leq b_n) \to 1$$

by Lemma 2.3.1.

This implies that $sup_{\theta \in \Theta_I} d\left(\theta, \widehat{\Theta}_I\right) \to_p 0$.

By definition, $Q \geq 0$. For any $\epsilon > 0$, suppose $inf_{\Theta \setminus \Theta_I^\epsilon} Q = 0$. Then there exists $\theta_0 \in \Theta$, s.t. $Q(\theta_0) = 0$, and θ_0 is a limit point of $\Theta \setminus \Theta_I^\epsilon$. $Q(\theta_0) = 0$ implies that $\theta_0 \in \Theta_I$. However, this contradicts the fact that θ_0 is a limit point of $\Theta \setminus \Theta_I^\epsilon$. Therefore, it must be the case that $inf_{\Theta \setminus \Theta_I^\epsilon} Q > 0$. Define $\delta(\epsilon) = inf_{\Theta \setminus \Theta_I^\epsilon} Q$. Then $\delta(\epsilon) > 0$.

Note that $sup_\Theta |Q - Q_n| = O_p(1/\sqrt{n})$, and$Q \leq Q_n + o_p(1)$ uniformly on Θ. Therefore $sup_{\widehat{\Theta}_I} Q \leq sup_{\widehat{\Theta}_I} Q_n + o_p(1)$. By definition, $\widehat{\Theta}_I = \{\theta \in \Theta : nQ_n(\theta) \leq b_n\}$. Therefore, $\forall \theta \in \widehat{\Theta}_I$, $Q_n(\theta) \leq b_n/n$. So $sup_{\widehat{\Theta}_I} Q_n \leq b_n/n$. Consequently, $sup_{\widehat{\Theta}_I} Q \leq b_n/n + o_p(1)$. Since $b_n/n \to 0$, $sup_{\widehat{\Theta}_I} Q \leq o_p(1)$. Therefore, $sup_{\widehat{\Theta}_I} Q < \delta(\epsilon) = inf_{\Theta \setminus \Theta_I^\epsilon} Q$ w.p.$\to 1$. So $\widehat{\Theta}_I \cap$

$(\Theta \setminus \Theta_I^\epsilon)$ is empty wp$\to 1$, which implies $\widehat{\Theta}_I \subseteq \Theta_I^\epsilon$ wp$\to 1$. Therefore, $sup_{\theta \in \widehat{\Theta}_I} d(\theta, \Theta_I) \leq \epsilon$, wp$\to 1$. Since ϵ is arbitrary, $sup_{\theta \in \widehat{\Theta}_I} d(\theta, \Theta_I) \to_p 0$.

$sup_{\theta \in \Theta_I} d\left(\theta, \widehat{\Theta}_I\right) \to_p 0$ and $sup_{\theta \in \widehat{\Theta}_I} d(\theta, \Theta_I) \to_p 0$ together imply that $d_H\left(\widehat{\Theta}_I, \Theta_I\right) \to_p 0$. ■

Proof of Lemma 2.3.3

Proof:

Note that wp$\to 1$, uniformly on Θ,

$$\begin{aligned} nQ_n(\theta) &= ||\sqrt{n}\left(E_n[m(\theta)] - E_p[m(\theta)]\right) + \sqrt{n}E_p[m(\theta)]||^2 \\ &\geq |\sqrt{n}||E_p[m(\theta)]|| - ||\sqrt{n}\left(E_n[m(\theta)] - E_p[m(\theta)]\right)||^2 \\ &\geq |K \cdot \sqrt{n}\left(d(\theta, \Theta_I) \wedge \delta\right) - O_p(1)|^2 \end{aligned}$$

The first inequality follows from the triangle inequality. And the second inequality follows from Condition C2.3.1 and Lemma 2.3.1. Therefore, for any $\varepsilon > 0$, one can choose n_ε large enough s.t. $\forall n \geq n_\varepsilon$, wp at least $1 - \varepsilon$,

$$nQ_n(\theta) \geq \frac{1}{2} \cdot K^2 \cdot n \cdot [d(\theta, \Theta_I) \wedge \delta]^2$$

uniformly on Θ, and, hence, uniformly on $\left\{\theta \in \Theta : d(\theta, \Theta_I) \geq (\kappa_\varepsilon/n)^{1/2}\right\}$ for any $\kappa_\varepsilon > 0$.

■

Proof of Theorem 2.3.2

Proof[2]:

For any $\varepsilon > 0$, let the positive constants $(\kappa, \delta, n_\varepsilon, \kappa_\varepsilon)$ be as specified in Lemma 2.3.3. Let $c := (\kappa \cdot \kappa_\varepsilon) \vee b_n$. There exists $n'_\varepsilon > n_\varepsilon$ s.t. for all $n > n'_\varepsilon$ wp at least $1 - \varepsilon$, $\epsilon_n := [c/(n\kappa)]^{1/2} \leq \delta$ and $\epsilon_n \geq (\kappa_\varepsilon/n)^{1/2}$. Hence, by Lemma 2.3.3,

$$inf_{\Theta \setminus \Theta_I^{\epsilon_n}} nQ_n \geq \kappa \cdot n \cdot (\epsilon_n \wedge \delta)^2 \geq \kappa \cdot n \cdot \epsilon_n^2 = c.$$

By definition, $\widehat{\Theta}_I = \{\theta \in \Theta : nQ_n(\theta) \leq b_n\}$. So $sup_{\widehat{\Theta}_I} nQ_n \leq b_n \leq c$. Therefore, $\widehat{\Theta}_I \subseteq \Theta_I^{\epsilon_n}$, which, associated by Theorem 2.3.1, implies that $d\left(\widehat{\Theta}_I, \Theta_I\right) \leq \epsilon_n$. Therefore, $d\left(\widehat{\Theta}_I, \Theta_I\right) = O_p\left([c/n]^{1/2}\right)$. ■

[2]This proof and the proof in A.10 follows the same logic as in CHT (2007) A.2. Part (2).

Proof of Lemma 4.1

Proof of Part (1):

$M(\theta) = (M_1(\theta), ..., M_{H\times J}(\theta))'$.

By definition, $\forall k \in \{1, 2, ..., H \times J\}$,

$$\begin{aligned} M_k(\theta) &= g_j^h(x_i)\, m_j(\theta) \\ &= g_j^h(x_i)\left[\pi(A_j; x_i, \theta) - \mathbf{1}(y_i = A_j)\right] \end{aligned}$$

for some $j = 1, ..., J,\ and\ h = 1, ..., H$.

Recall that $\theta = (\gamma, , \beta, \theta_1, ..., \theta_{2M})$.

(a)

$$\begin{aligned} \partial M_k(\theta)/\partial\gamma &= \left[\partial\pi(A_j; x_i, \theta)/\partial\gamma\right] g_j^h(x_i) \\ &= g_j^h(x_i)\sum_{m=1}^{M}\left[\sum_{t=1}^{T}\left[y_{it}y_{i\,t-1}f_{\varepsilon t} - (1-y_{i\,t})\cdot y_{i\,t-1}f_{\varepsilon t}\right]\prod_{t'\neq t} f_{t'}\theta_m\right] \\ &\quad + g_j^h(x_i)\sum_{m=M+1}^{2M}\left[\sum_{t=1}^{T}\left[y_{it}y_{i\,t-1}f_{\varepsilon t} - (1-y_{i\,t})\cdot y_{i\,t-1}f_{\varepsilon t}\right]\prod_{t'\neq t} f_{t'}\theta_m\right]. \end{aligned}$$

By equation (2.14), which defines f_t for $t = 1, ..., T$, $|f_{t'}| \leq 1$.

By assumptions, $f_{\varepsilon t}$ is bounded, which means $\exists W > 0$, s.t. $|f_{\varepsilon t}| \leq W$.

Also, $\theta_m \in [0, 1]$. And $y_{it} \in \{0, 1\}$.

Therefore, $\left|\left[y_{it}y_{i\,t-1}f_{\varepsilon t} - (1-y_{i\,t})\cdot y_{i\,t-1}f_{\varepsilon t}\right]\prod_{t'\neq t} f_{t'}\theta_m\right| \leq 2W$.

Consequently, $|\partial M_k(\theta)/\partial\gamma| \leq |g_j^h(x_i)| \cdot (2M) \cdot T \cdot (2W)$.

(b)

$$\begin{aligned}
\partial M_k(\theta)/\partial\beta &= [\partial\pi(A_j;\, x,\, \theta)/\partial\beta]\, g_j^h(x) \\
&= g_j^h(x_i) \sum_{m=1}^{M} \left[\sum_{t=1}^{T} [y_{it} x_{it} f_{\varepsilon t} - (1 - y_{i\,t}) \cdot x_{it} f_{\varepsilon t}] \prod_{t' \neq t} f_{t'} \theta_m \right] \\
&\quad + g_j^h(x_i) \sum_{m=M+1}^{2M} \left[\sum_{t=1}^{T} [y_{it} x_{it} f_{\varepsilon t} - (1 - y_{i\,t}) \cdot x_{it} f_{\varepsilon t}] \prod_{t' \neq t} f_{t'} \theta_m \right].
\end{aligned}$$

Note that x_{it} is bounded by assumption.

Then, similarly, $\exists U > 0$, s.t. $|\partial M_k(\theta)/\partial\beta| \leq |g_j^h(x_i)| \cdot U$.

(c)

$$\begin{aligned}
\partial M_k(\theta)/\partial\theta_m &= [\partial\pi(A_j;\, x_i,\, \theta)/\partial\theta_m]\, g_j^h(x) \\
&= g_j^h(x) \prod_{t=1}^{T} f_t(y_{it} | y_{i\,t-1},\, x_{it},\, \alpha_i = a_m;\, \gamma,\, \beta)
\end{aligned}$$

for $m = 1, ..., M$.

Therefore, $|\partial M_k(\theta)/\partial\theta_m| \leq |g_j^h(x)|$.

Similarly, $|\partial M_k(\theta)/\partial\theta_m| \leq |g_j^h(x)|$, for $m = M+1, ..., 2M$.

According to (a), (b), and (c), $\exists V_k > 0$, s.t. $||\nabla M_k(\theta)/\nabla\theta|| \leq |g_j^h(x_i)| \cdot V_k$.

So $\forall \theta_1,\, \theta_2 \in \Theta'$, $|M_k(\theta_1) - M_k(\theta_2)| \leq |g_j^h(x_i)| \cdot V_k \cdot ||\theta_1 - \theta_2||$.

The above is true for all $k = 1, ..., H \times J$.

Recall that $M(\theta) = (M_1(\theta), ..., M_{H\times J}(\theta))'$.

Therefore, $||M(\theta_1) - M(\theta_2)|| \leq \left[sup_{h,j}|g_j^h(x_i)|\right] \cdot \left(\sum_{k=1}^{H\times J} V_k^2\right)^{1/2} \cdot ||\theta_1 - \theta_2||$.

Since $g^h(x_i) \in \mathcal{G} \subseteq L^d(S)$, $E_p|g_j^h(x_i)|^d < \infty$ for all finite h and j.

Consequently $E_p\left[sup_{h,j}|g_j^h(x_i)|\right]^d < \infty$.

Therefore, $\{M(\theta) : \theta \in \Theta\}$ is P-Donsker.[3]

Equivalently, $\sqrt{n}(E_n[M(\theta)] - E_p[M(\theta)]) \Rightarrow \Delta(\theta)$ in $L^\infty(\Theta')$.

(d) $\Delta(\theta)$ having continuous paths a.s. follows from the functional central limit theorem stated in Pollard (1990)[4].

Define $f_{ni}(w, \theta) = \{M(\theta; x_i(w), y_i(w)) - E_P[M(\theta)]\}/\sqrt{n}$.

Define $X_n(w, \theta) = \sum_{i=1}^n f_{ni}(w, \theta)$.

By definition, $X_n(w, \theta) = \sqrt{n}(E_n[M(\theta)] - E_p[M(\theta)])$.

Define $\rho_n(\theta_1, \theta_2) = \left(\sum_i E_P||f_{ni}(\cdot, \theta_1 - f_{ni}(\cdot, \theta_2)||^2\right)^{1/2}$.

And, define $\rho(\theta_1, \theta_2) = \lim \rho_n(\theta_1, \theta_2)$, when the limit exists.

The five regularity conditions in the functional central limit theorem are:

(1) $\{f_{ni}\}$ is manageable;

(2) $H(\theta_1, \theta_2) = \lim E_P[X_n(w, \theta_1) \cdot X_n(w, \theta_2)]$ exists for every θ_1, θ_2 in Θ;

(3) $\limsup \sum_i E_P\left(F_{ni}^2\right) < \infty$ for some function F_{ni}, s.t. $||f_{ni}|| \leq F_{ni}$, almost sure.

[3] $\{M(\theta) : \theta \in \Theta\}$ being P-Donsker follows directly from van der Vaart (1998) Theorem 19.4 and Example 19.6 in $p270 - 271$.

[4] The theorem is formally presented on page 53 of David Pollard's book *Empirical Processes Theory and Applications.*

(4) For the same F_{ni}, $\sum_i E_P\left[\left(F_{ni}^2\right) \cdot \mathbf{1}\left(F_{ni} > \epsilon\right)\right] \to 0$ for each $\epsilon > 0$.

(5) The limit $\rho(\cdot, \cdot)$ is well defined, and for all deterministic sequences $\{s_n\}$ and $\{t_n\}$, if $\rho(s_n, t_n) \to 0$, then $\rho_n(s_n, t_n) \to 0$.

The functional central limit theorem is applicable for the general triangle arrays $\{f_{ni}\}$. As stated in Pollard (1990), for the special case where $f_{ni}(w, \theta) = f_i(w, \theta)/\sqrt{n} = \{M(\theta; x_i(w), y_i(w)) - E_P[M(\theta)]\}/\sqrt{n}$, with the $\{f_i\}$ i.i.d., $\rho = \rho_n$, and condition (5) is trivially satisfied.

Condition (1) is needed to prove the convergence of $X_n(w, \theta)$, but is not related to showing that the limit has continuous paths almost sure. As already proven, $X_n(w, \theta) \Rightarrow \Delta(\theta)$. Therefore, to prove that $\Delta(\theta)$ has continuous paths a.s.,all that needed is to show that condition (2), (3), and (4) hold.

Since $\lim E_P[X_n(w, \theta_1) \cdot X_n(w, \theta_2)] = cov(\Delta(\theta_1), \Delta(\theta_2))$, which exists for arbitrary θ_1, θ_2 in Θ by the definition of Gaussian process, condition (2) is satisfied.

$f_{ni}(w, \theta) = \{M(\theta; x_i(w), y_i(w)) - E_P[M(\theta)]\}/\sqrt{n}$.

And $M(\theta) = (M_1(\theta), ..., M_{H\times J}(\theta))'$.

By definition, $\forall k \in \{1, 2, ..., H \times J\}$,

$$M_k(\theta) = g_j^h(x_i)\, m_j(\theta)$$
$$= g_j^h(x_i)\left[\pi(A_j; x_i, \theta) - \mathbf{1}(y_i = A_j)\right]$$

A.2. has shown that $|\pi(A_j; x_i, \theta) - \mathbf{1}(y_i = A_j)| \leq 1$. Therefore,

$|M_k(\theta; x_i(w), y_i(w)) - E_P[M_k(\theta)]| \leq 2|g_j^h(x_i) - E_P\left[g_j^h(x_i)\right]|$.

Consequently, $||f_{ni}(w, \theta)|| \leq \left[2 \cdot \sqrt{H \times J} \cdot \sup_{j,h} |g_j^h(x_i) - E_P\left[g_j^h(x_i)\right]|\right]/\sqrt{n}$.

Define $F_{ni}(w) = \left[2 \cdot \sqrt{H \times J} \cdot \sup_{j,h} |g_j^h(x_i(w)) - E_P\left[g_j^h(x_i(w))\right]|\right] / \sqrt{n}$.

In turn, $[F_{ni}(w)]^2 = \left[4 \cdot H \cdot J \cdot \sup_{j,h} |g_j^h(x_i(w)) - E_P\left[g_j^h(x_i(w))\right]|^2\right] / n$.

Then $\sum_i E_P\left(F_{ni}^2\right) = E_p\left[\sup_{j,h} |g_j^h(x_i(w)) - E_P\left[g_j^h(x_i(w))\right]|^2\right] < \infty$. This equation follows from the fact the $g_j^h(x_i(w)) \in L^d(S)$, $d > 1$, implies that $g_j^h(x_i(w)) \in L^1(S)$, and that there are only finitely many g_j^h. Therefore, condition (3) is satisfied. Moreover, $\sum_i E_P\left(F_{ni}^2\right)$ is uniformly bounded. This is because $E_p\left[\sup_{j,h} |g_j^h - E_P\left(g_j^h\right)|^2\right]$ is not a function of n.

$F_{ni}(w) \to 0$, almost sure. Then, $F_{ni}(w) \to_p 0$. Therefore for each $\epsilon > 0$, $P(F_{ni}(w) > \epsilon) = E_P[\mathbf{1}(F_{ni} > \epsilon)] \to 0$. This combined with condition (3) implies condition (4).

Proof of Part (2):

Since $E_p[M(\theta)] = 0$ for all $\theta \in \Theta_{\mathcal{G}}$, $\sqrt{n}(E_n[M(\theta)]) \Rightarrow \Delta(\theta)$ in $L^\infty(\Theta_{\mathcal{G}})$.

By continuous mapping theorem, $nQ_n(\theta) = ||\sqrt{n}E_n[m(\theta)]||^2 \Rightarrow ||\Delta(\theta)||^2$ in $L^\infty(\Theta_{\mathcal{G}})$.

Therefore $\mathcal{C}_n = sup_{\theta \in \Theta_{\mathcal{G}}} nQ_n(\theta) \to_d sup_{\theta \in \Theta_{\mathcal{G}}} ||\Delta(\theta)||^2$.

By assumption, $\Delta(\theta)$ have a.s. continuous paths.

This implies that $||\Delta(\theta)||^2$ have a.s. continuous paths.

Thus, $sup_{\theta \in \Theta_{\mathcal{G}}} ||\Delta(\theta)||^2 \leq sup_{\theta \in \Theta} ||\Delta(\theta)||^2 < \infty$, a.s..

$sup_{\theta \in \Theta} ||\Delta(\theta)||^2 < \infty$, a.s. follows from the compactness of Θ and the fact that a

continuous function is bounded on a compact set.

Therefore, $sup_{\theta\in\Theta_{\mathcal{G}}}||\Delta(\theta)||^2 < \infty$, a.s..[5]

Thus, $\forall b_n \to \infty$, $\forall \varepsilon > 0$, $\exists L > 0$, s.t. $Pr\left(sup_{\theta\in\Theta_{\mathcal{G}}}||\Delta(\theta)||^2 \leq L\right) > 1-\varepsilon$, and the cdf of $sup_{\theta\in\Theta_{\mathcal{G}}}||\Delta(\theta)||^2$ is continuous at L[6].

$\exists N > 0$, s.t. $b_n > L$ for all $n \geq N$.

Consequently, $Pr(\mathcal{C}_n \leq b_n) \geq Pr(\mathcal{C}_n \leq L)$ for all $n \geq N$.

Therefore, $\liminf Pr(\mathcal{C}_n \leq b_n) \geq \liminf Pr(\mathcal{C}_n \leq L) = Pr\left(sup_{\theta\in\Theta_{\mathcal{G}}}||\Delta(\theta)||^2 \leq L\right) > 1-\varepsilon$.

Since the above is true for arbitrary ε, $\liminf Pr(\mathcal{C}_n \leq b_n) \geq 1$.

Since $\limsup Pr(\mathcal{C}_n \leq b_n) \leq 1$, $\lim Pr(\mathcal{C}_n \leq b_n) = 1$. ■

A.7. Proof of Lemma 4.2

Proof:

(a) By equation (2.14), $f_t(y_{it}|y_{i\,t-1},\, x_{it},\, \alpha_i;\, \gamma,\, \beta)$ is continuous w.r.t. γ, β and α_i. Consequently, by equation (16), $\pi(A_j;\, x_i,\, \theta)$ is continuous w.r.t. θ.

$E_p\left\{g_j^h(x)\left[\pi(A_j;\, x_i,\, \theta) - \mathbf{1}(y_i = A_j)\right]\right\}$ being continuous w.r.t. θ is immediate.

So does $\frac{1}{n}\sum_{i=1}^n \left\{g_j^h(x)\left[\pi(A_j;\, x_i,\, \theta) - \mathbf{1}(y_i = A_j)\right]\right\}$ being continuous w.r.t. θ.

[5] Indeed, the assumption that $\Delta(\theta)$ have a.s. continuous paths is needed to prevent the random variable $sup_{\theta\in\Theta_{\mathcal{G}}}||\Delta(\theta)||^2$ from having strictly positive probability mass at ∞. In other words, the assumption guarantees that the cdf of $sup_{\theta\in\Theta_{\mathcal{G}}}||\Delta(\theta)||^2$ is left continuous at ∞.

[6] Note that the continuous points of a cdf is dense in the real line.

The above is true for all $h = 1, ..., H$ and $j = 1, ..., J$.

Therefore, by definition, Q and Q_n are continuous w.r.t. θ.

(b)$Q_n(\theta) - Q(\theta) = (E_n[M(\theta)] - E_p[M(\theta)])'(E_n[M(\theta)] + E_p[M(\theta)])$

So $|Q_n(\theta) - Q(\theta)| = ||E_n[M(\theta)] - E_p[M(\theta)]|| \cdot ||E_n[M(\theta)] + E_p[M(\theta)]||$.

$M(\theta) = (M_1(\theta), ..., M_{H\times J}(\theta))'$.

Therefore, $\forall k \in \{1, 2, ..., H \times J\}$,

$$\begin{aligned} M_k(\theta) &= g_j^h(x_i) m_j(\theta) \\ &= g_j^h(x_i)[\pi(A_j; x_i, \theta) - \mathbf{1}(y_i = A_j)] \end{aligned}$$

for some $j = 1, ..., J, and\, h = 1, ..., H$.

As shown in A.3.(b), $|\pi(A_j; x_i, \theta) - \mathbf{1}(y_i = A_j)| \leq 1$. Therefore $E_p|M_k(\theta)| \leq E_p|g_j^h(x_i)|$.

And $E_n|M_k(\theta)| \leq \frac{1}{n}\sum_{i=1}^n |g_j^h(x_i)| < E_p|g_j^h(x_i)| + 1$, for n large enough.

The last inequality follows from the weak law of large number.

By assumption, $E_p|g_j^h(x_i)|^d < \infty$ for all h and j.

By Jensen's inequality, $E_p|g_j^h(x_i)| \leq \left[E_p|g_j^h(x_i)|^d\right]^{1/d} < \infty$.

Since both h and j take finitely many different numbers, $\exists \Gamma > 0$, s.t. $E_p|g_j^h(x_i)| \leq \Gamma - 1$ for all h and j.

Consequently, $E_p|M_k(\theta)| \leq \Gamma$, and $E_n|M_k(\theta)| \leq \Gamma$ for all h and j.

Therefore, $|E_n[M(\theta)] + E_p[M(\theta)]| \leq 2\Gamma \cdot \mathbf{1}$.

So $||E_n[M(\theta)] + E_p[M(\theta)]|| \leq 2\Gamma\sqrt{H \times J}$.

Since $\sqrt{n}(E_n[M(\theta)] - E_p[M(\theta)]) \Rightarrow \Delta(\theta)$ in $L^\infty(\Theta')$,

$\sqrt{n}||E_n[M(\theta)] - E_p[M(\theta)]|| \Rightarrow ||\Delta(\theta)||$ in $L^\infty(\Theta')$.

$sup_{\theta\in\Theta}\sqrt{n}||E_n[M(\theta)] - E_p[M(\theta)]|| \rightarrow_d sup_{\theta\in\Theta}||\Delta(\theta)||^2$.

Since $sup_{\theta\in\Theta}||\Delta(\theta)||^2 < \infty$ a.s., which is shown in A.6., the above weak convergence implies that

$sup_{\theta\in\Theta}||E_n[M(\theta)] - E_p[M(\theta)]|| = O_p(1/\sqrt{n})$.

Thus, $sup_{\theta\in\Theta}|Q_n(\theta) - Q(\theta)| \leq \left(2\Gamma\sqrt{H \times J}\right) \cdot sup_{\theta\in\Theta}||E_n[M(\theta)] - E_p[M(\theta)]|| = O_p(1/\sqrt{n})$.

So $sup_{\theta\in\Theta}|Q_n(\theta) - Q(\theta)| = O_p(1/\sqrt{n})$.

(c) As shown in A.6. Part (2), $nQ_n(\theta) = ||\sqrt{n}E_n[m(\theta)]||^2 \Rightarrow ||\Delta(\theta)||^2$ in $L^\infty(\Theta_{\mathcal{G}})$.

So $sup_{\theta\in\Theta_{\mathcal{G}}}nQ_n(\theta) \rightarrow_d sup_{\theta\in\Theta_{\mathcal{G}}}||\Delta(\theta)||^2$.

Since $sup_{\theta\in\Theta_{\mathcal{G}}}||\Delta(\theta)||^2 < \infty$ a.s., the above weak convergence implies that $sup_{\theta\in\Theta_{\mathcal{G}}}Q_n(\theta) = O_p(1/n)$. ■

A.8. Proof of Theorem 2.4.1

Proof:

Let $\widehat{\Theta}_{\mathcal{G}} = C_n(b_n)$. Then $\Theta_{\mathcal{G}} \subseteq \widehat{\Theta}_{\mathcal{G}} \iff \mathcal{C}_n \leq b_n$.Therefore:

$$Pr\left(\Theta_{\mathcal{G}} \subseteq \widehat{\Theta}_{\mathcal{G}}\right) = Pr\left(\mathcal{C}_n \leq b_n\right) \to 1$$

by Lemma 2.4.1.

This implies that $sup_{\theta \in \Theta_I} d\left(\theta,\, \widehat{\Theta}_{\mathcal{G}}\right) \to_p 0$.

By definition, $Q \geq 0$. For any $\epsilon > 0$, suppose $inf_{\Theta \setminus \Theta^{\epsilon}_{\mathcal{G}}} Q = 0$. Then there exists $\theta_0 \in \Theta$, s.t. $Q(\theta_0) = 0$, and θ_0 is a limit point of $\Theta \setminus \Theta^{\epsilon}_{\mathcal{G}}$.[7] $Q(\theta_0) = 0$ implies that $\theta_0 \in \Theta_{\mathcal{G}}$. However, this contradicts the fact that θ_0 is a limit point of $\Theta \setminus \Theta^{\epsilon}_{\mathcal{G}}$. Therefore, it must be the case that $inf_{\Theta \setminus \Theta^{\epsilon}_{\mathcal{G}}} Q > 0$. Define $\delta(\epsilon) = inf_{\Theta \setminus \Theta^{\epsilon}_{\mathcal{G}}} Q$. Then $\delta(\epsilon) > 0$.

By RC(b) in Lemma 2.4.2, $sup_{\Theta} |Q - Q_n| = O_p\left(1/\sqrt{n}\right)$,

$Q \leq Q_n + o_p(1)$ uniformly on Θ. Therefore $sup_{\widehat{\Theta}_{\mathcal{G}}} Q \leq sup_{\widehat{\Theta}_{\mathcal{G}}} Q_n + o_p(1)$. By definition, $\widehat{\Theta}_{\mathcal{G}} = \{\theta \in \Theta : nQ_n(\theta) \leq b_n\}$. Therefore, $\forall \theta \in \widehat{\Theta}_{\mathcal{G}}$, $Q_n(\theta) \leq b_n/n$. So $sup_{\widehat{\Theta}_{\mathcal{G}}} Q_n \leq b_n/n$. Consequently, $sup_{\widehat{\Theta}_{\mathcal{G}}} Q \leq b_n/n + o_p(1)$. Since $b_n/n \to 0$, $sup_{\widehat{\Theta}_{\mathcal{G}}} Q \leq o_p(1)$. Therefore, $sup_{\widehat{\Theta}_{\mathcal{G}}} Q < \delta(\epsilon) = inf_{\Theta \setminus \Theta^{\epsilon}_{\mathcal{G}}} Q$ wp$\to 1$. So $\widehat{\Theta}_{\mathcal{G}} \cap (\Theta \setminus \Theta^{\epsilon}_{I})$ is empty wp$\to 1$, which implies $\widehat{\Theta}_{\mathcal{G}} \subseteq \Theta^{\epsilon}_{\mathcal{G}}$ wp$\to 1$. Therefore, $sup_{\theta \in \widehat{\Theta}_{\mathcal{G}}} d(\theta,\, \Theta_{\mathcal{G}}) \leq \epsilon$, wp$\to 1$. Since ϵ is arbitrary, $sup_{\theta \in \widehat{\Theta}_{\mathcal{G}}} d(\theta,\, \Theta_{\mathcal{G}}) \to_p 0$.

$sup_{\theta \in \Theta_{\mathcal{G}}} d\left(\theta,\, \widehat{\Theta}_{\mathcal{G}}\right) \to_p 0$ and $sup_{\theta \in \widehat{\Theta}_{\mathcal{G}}} d(\theta,\, \Theta_{\mathcal{G}}) \to_p 0$ together imply that $d_H\left(\widehat{\Theta}_{\mathcal{G}},\, \Theta_{\mathcal{G}}\right) \to_p 0$. ■

[7] Recall that $\Theta^{\epsilon}_{\mathcal{G}} = \{\theta \in \Theta : d(\theta,\, \Theta_{\mathcal{G}}) \leq \epsilon\}$, which is by definition the $\epsilon-$expansion of $\Theta_{\mathcal{G}}$ in Θ.

Proof of Lemma 2.4.3

Proof:

Note that wp$\to 1$, uniformly on Θ,

$$\begin{aligned} nQ_n(\theta) &= ||\sqrt{n}\left(E_n\left[M(\theta)\right] - E_p\left[M(\theta)\right]\right) + \sqrt{n}E_p\left[M(\theta)\right]||^2 \\ &\geq |\sqrt{n}||E_p\left[M(\theta)\right]|| - ||\sqrt{n}\left(E_n\left[M(\theta)\right] - E_p\left[M(\theta)\right]\right)||^2 \\ &\geq |K \cdot \sqrt{n}\left(d(\theta,\, \Theta_{\mathcal{G}}) \wedge \delta\right) - O_p(1)|^2 \end{aligned}$$

The first inequality follows from the triangle inequality. And the second inequality follows from Condition C.4.1 and Lemma 4.1. Therefore, for any $\varepsilon > 0$, one can choose n_ε large enough s.t. $\forall n \geq n_\varepsilon$, wp at least $1 - \varepsilon$,

$$nQ_n(\theta) \geq \tfrac{1}{2} \cdot K^2 \cdot n \cdot \left[d(\theta,\, \Theta_{\mathcal{G}}) \wedge \delta\right]^2$$

uniformly on Θ, and, hence, uniformly on $\left\{\theta \in \Theta :\, d(\theta,\, \Theta_{\mathcal{G}}) \geq (\kappa_\varepsilon / n)^{1/2}\right\}$ for any $\kappa_\varepsilon > 0$. ■

Proof of Theorem 2.4.2

Proof:

For any $\varepsilon > 0$, let the positive constants $(\kappa,\, \delta,\, n_\varepsilon, \kappa_\varepsilon)$ be as specified in Lemma 2.4.3. Let $c := (\kappa \cdot \kappa_\varepsilon) \vee b_n$. There exists $n'_\varepsilon > n_\varepsilon$ s.t. for all $n > n'_\varepsilon$ wp at least $1 - \varepsilon$, $\epsilon_n :=$

$[c/(n\kappa)]^{1/2} \leq \delta$ and $\epsilon_n \geq (\kappa_\varepsilon/n)^{1/2}$. Hence, by Lemma 2.4.3,

$$inf_{\Theta \setminus \Theta_{\mathcal{G}}^{\epsilon_n}} nQ_n \geq \kappa \cdot n \cdot (\epsilon_n \wedge \delta)^2 \geq \kappa \cdot n \cdot \epsilon_n^2 = c.$$

By definition, $\widehat{\Theta}_{\mathcal{G}} = \{\theta \in \Theta : nQ_n(\theta) \leq b_n\}$. So $sup_{\widehat{\Theta}_{\mathcal{G}}} nQ_n \leq b_n \leq c$. Therefore, $\widehat{\Theta}_{\mathcal{G}} \subseteq \Theta_{\mathcal{G}}^{\epsilon_n}$, which, associated with Theorem 2.4.1, implies that $d\left(\widehat{\Theta}_{\mathcal{G}}, \Theta_{\mathcal{G}}\right) \leq \epsilon_n$. Therefore, $d\left(\widehat{\Theta}_{\mathcal{G}}, \Theta_{\mathcal{G}}\right) = O_p\left([c/n]^{1/2}\right)$. ■