Subvector Inference for Varying Coefficient Models with Partial Identification*

Shengjie Hong[†] Yu-Chin Hsu[‡] Yuanyuan Wan[§]

Abstract

This paper considers a general class of varying coefficient models defined by a set of moment equalities and/or inequalities, where unknown functional parameters are not necessarily point-identified. We propose an inferential procedure for a subvector of the varying parameters and establish the asymptotic validity of the resulting confidence sets uniformly over a broad family of data-generating processes. We also propose a practical specification test for a set of necessary conditions of models considered in this paper. Monte Carlo studies show that the proposed methods have good finite sample properties. We apply our method to estimate the return to education using part of the year 2005 1%-population census data from China.

Keywords: Varying coefficient; Moment inequalities; Partial-identification

JEL classification: C12, C14, C15

*This version: October 5, 2024. The authors are grateful to the Co-Editor Xiaoxia Shi, two anonymous referees, Mathieu Marcoux and Erhao Xie, for valuable comments and suggestions, which have considerably improved the presentation of the paper. Shengjie Hong gratefully acknowledges the support from the National Natural Science Foundation of China (NSFC) under Grant number 72373175. Yu-Chin Hsu gratefully acknowledges the research support from the National Science and Technology Council of Taiwan (NSTC112-2628-H-001-001), the Academia Sinica Investigator Award of the Academia Sinica, Taiwan (AS-IA-110-H01), and from the Center for Research in Econometric Theory and Applications of National Taiwan University (Grant No. 113L8601). Yuanyuan Wan gratefully acknowledges the support from the SSHRC Insight Grant #43520190500 and #43520240418.

[†]School of Economics, Renmin University of China. E-mail: hongshj@ruc.edu.cn. 59 Zhongguancun St., Haidian Dist., Beijing, 100872, China.

[‡]Institute of Economics, Academia Sinica; Department of Finance, National Central University; Department of Economics, National Chengchi University. E-mail: ychsu@econ.sinica.edu.tw. 128 Academia Rd., Section 2, Nankang, Taipei, 115, Taiwan.

[§]Department of Economics, University of Toronto. E-mail: yuanyuan.wan@utoronto.ca. 232, Max Gluskin House, 150 St. George St., Toronto, Ontario, M5S 3G7, Canada.

1 Introduction

Since the seminal paper of Hastie and Tibshirani (1993), varying coefficient models have been widely adopted in empirical research in economics and finance for their balance of providing both dimension reduction and flexible modeling of heterogeneous effects. For example, Li, Huang, Li, and Fu (2002) proposed a semiparametric varying coefficient model to estimate production functions in which the elasticity of inputs varies with the intermediate production and management expenses. Ang and Liu (2004) studied how to discount cash flows with time-varying expected returns based on varying coefficient models. Cai, Ren, and Yang (2015) used varying coefficient models to estimate time-varying betas and alpha in the conditional capital asset pricing model. And Cai, Chen, and Fang (2018) used varying coefficient models to estimate the growth effect of FDI. See Cai and Hong (2009) and Cai (2010) for more references on applications of varying coefficient models.

Motivated by empirical applications, the econometric theory of varying coefficient models has been developed and extended to various modeling environments. For instance, Chen and Tsay (1993) considered the time series setting and developed varying coefficient autoregressive models. Fan and Zhang (1999), Cai, Fan, and Li (2000), and Ahmad, Leelahanon, and Li (2005) discussed efficient estimation. Fan and Zhang (2000) and Fan and Li (2004) considered the panel data setting. Cai and Xu (2008) proposed quantile regression methods for a class of smooth coefficient models. Cai, Das, Xiong, and Wu (2006) and Cai, Fang, Lin, and Su (2019) studied a class of instrumental variable regression functional-coefficient representation for the regression function. Su, Murtazashvili, and Ullah (2013) proposed a consistent inference procedure for the testing constancy of varying coefficients.

Our paper contributes to the literature on varying coefficient models. We consider making inferences in a general class of varying coefficient models defined by a set of conditional moment equalities and/or inequalities. The notable difference from the existing literature is that the unknown functional parameters can be partially identified in our setup. In practice, the assumptions that deliver point-identification of the parameters may not necessarily hold. For example, in a varying coefficient linear regression or quantile regression model, the slope parameter is not point-identified if the outcome variable is interval-observed or censored, as is quite common in many survey data. In a varying coefficient instrumental regression model, the structural parameter may not be point-identified if the instrumental variable is imperfect (e.g. not independent of the structural error). In an oligopoly market entry model, the profit function with varying coefficients is typically not point-identified if there are multiple equilibria and the equilibrium selection mechanism is unknown to researchers. As we will discuss these examples in detail in Section 2 and revisit them in our empirical and simulation studies (Sections 4 and 5, respectively), we hope to emphasize here that it is useful to develop inferential procedures for varying coefficients that are robust to partial identification.

Our approach is built upon and extends Andrews and Shi (2014, AS hereafter), who considered a class of conditional moment inequality models in which the parameter is also a function of a subset of covariates. AS focuses on confidence sets for the whole varying parameter vector evaluated at a given point; however, motivated by some empirical applications of varying coefficient models, we instead focus on constructing confidence sets for a subvector of the parameters. For this purpose, we use a different test statistic from that in AS. Specifically, we extend the profiling-based method of Bugni, Canay, and Shi (2017), which was initially designed for subvector inference in unconditional moment inequality models with finite-dimensional parameters, to the current setup of conditional moment inequality with functional parameters.

Our paper also contributes to the literature of conditional moment inequality models.[1] Recently, a line of work studies partially identified conditional moment models; an incomplete list includes Kim (2008), Andrews and Shi (2013, 2017), Chernozhukov, Lee, and Rosen (2013), Lee, Song, and Whang (2013), Armstrong (2014, 2015, 2018),

[1]There has been a large literature on unconditional moment inequality models under partial identification, see, for example, Andrews, Berry, and Jia (2004), Imbens and Manski (2004), Chernozhukov, Hong, and Tamer (2007), Andrews and Guggenberger (2009), Romano and Shaikh (2008, 2010), Andrews and Soares (2010), Wan (2013), Menzel (2014), Bugni, Canay, and Shi (2015, 2017), Pakes, Porter, Ho, and Ishii (2015), Andrews and Kwon (2019), and Belloni, Bugni, and Chernozhukov (2019) among others. For a more thorough review, please see Canay and Shaikh (2017) and references therein.

Bontemps and Magnac (2017), and Hsu and Shi (2017), among others. All these papers consider finite-dimensional parameters and hence do not accommodate varying coefficients. There are a small number of papers that allow the parameter vector to contain an infinite-dimensional component, for example, Santos (2012), Tao (2015), Hong (2017), and Chernozhukov, Newey, and Santos (2023), but they consider only conditional moment equalities.

We propose a specification test for the necessary implications of the model, which was not considered in AS. We show that our test controls the size uniformly over a set of DGPs and is consistent against any violation of the necessary implication of the model. Our paper, therefore, also contributes to the literature of specification tests for conditional moment inequalities with infinite-dimensional parameters, and it complements the existing work of Andrews and Shi (2013), Bugni, Canay, and Shi (2015), and Marcoux, Russell, and Wan (2024), where the parameters are finite-dimensional.

To illustrate our method, we estimate the varying returns to education in different areas of China using the mother's education as the IV. Local development factors, such as the quality of the local labor market and the local infrastructure development, can affect the return to education. Therefore, we construct the model such that the coefficient of the education level varies with a measure of the local development level. Instead of assuming IV independence, we assume the mother's education positively correlates with children's talent, which leads to a set of moment inequalities. Our estimation results show that the confidence interval for the return to education varies substantially across local development levels in both its width (reflecting the identification power) and location (reflecting the magnitude of the education effect). These features can not be captured by either a point-identified varying coefficient model or moment inequality models with non-varying coefficients.

The rest of the paper is organized as follows. We present the model and a few motivating examples in Section 2. In Section 3, we construct the uniformly valid confidence set and propose the model specification test. In Section 4, we use Monte Carlo simulations to illustrate the finite sample performance of the proposed methods. Section 5 reports results from our empirical application and Section 6 concludes. For ease of exposition, we collect all the proofs and additional empirical and simulation results in the Appendix.

2 Model and Examples

We consider a general class of varying coefficient models defined by a set of conditional moment inequalities and/or equalities. Specifically, for any $z \in \mathcal{Z}$, let

$$\begin{aligned} &E_P[m_j(W, \theta_0(z))|X, Z = z] \geq 0 \text{ a.s. } X, \text{ for } j = 1, \ldots, p \text{ and} \\ &E_P[m_j(W, \theta_0(z)|X, Z = z] = 0 \text{ a.s. } X, \text{ for } j = p+1, \ldots, k. \end{aligned} \tag{2.1}$$

In this model, $m_j(\cdot, \theta)$ for $j = 1, \ldots, k$ are known real-valued moment functions. Let the conditioning variables $X \in \mathcal{X} \subseteq R^{d_x}$ and $Z \in \mathcal{Z} \subseteq R^{d_z}$. The varying coefficient $\theta_0(\cdot) : \mathcal{Z} \to \Theta \subseteq \mathbb{R}^{d_\theta}$ varies with z and takes value in a compact set Θ. The random vector W contains some other random variables $Y \in \mathcal{Y} \subseteq R^{d_y}$ and possibly (X, Z), so that $W = (X', Y', Z')' \in R^{d_w}$ with $d_w = d_y + d_x + d_z$. In empirical applications, Y is often the outcome variable of interest. Without loss of generality, we assume that X and Z do not overlap. We use P for the probability measure that generates the data and E_P for the expectation under the distribution P. The main departure of our paper from the classical varying coefficient models is that we allow $\theta_0(z)$ to be partially identified in the sense that its identified set

$$\Theta_P(z) = \{\theta \in \Theta : \text{ (2.1) holds with } \theta \text{ in place of } \theta_0(z).\} \tag{2.2}$$

may contain more than one element.

Model (2.1) encompasses a broad class of models and applies to many empirical contexts, including those mentioned in the introduction, including the conventional point-identified varying coefficient models as special cases. Here, we discuss four detailed examples relating to imperfect IV, interval data, entry games, and a firm-level gravity model, respectively. The first example is followed by an empirical study in Section 5, and the second and third ones are followed by simulation studies. We provide a few additional examples in Appendix E, including ones relating to quantile regression with interval-outcome, quantile regression with censoring, and test of local average treatment effect (LATE) assumptions, respectively.

Example 2.1 (Imperfect IV) *Consider an example of estimating return to education using a linear model,*

$$Y = X_1\theta_{01}(Z) + X_2'\theta_{02}(Z) + \varepsilon, \tag{2.3}$$

where Y is the logarithm of wage, X_1 is the key explanatory variable education, X_2 is a vector of exogenous demographic variables (may include an intercept term), ε is the unobserved talent or ability, and Z is the variable that drives the varying coefficients $\theta_0(z) \equiv (\theta_{01}(z), \theta_{02}(z))$. The choice of Z depends on the research goals. For example, some literature argues that the return to education depends on experiences (see discussions in Card, 2001; Schultz, 2003; Su, Murtazashvili, and Ullah, 2013), and it can be restrictive to impose a parametric assumption on $\theta(z)$ without additional information. So, in this case, Z is the experience.[2] *In our empirical illustration in Section 5, we highlight that the return to education depends on the quality of the local labor market and infrastructure. There, Z is a proxy of the local development level.*

Regardless of the research goals, if the education is correlated with the structural error ε, one may need to use the instrumental variable approach to identify parameters. The model becomes a version of the IV-varying coefficient model studied by Cai, Fang, Lin, and Su (2019). On the other hand, the (mean)-independence assumption of many popular instrumental variables, such as distance to college or parent's education, can be controversial in some applications.[3] *In such cases, as discussed in Nevo and Rosen (2012), it may be more reasonable to assume that the children's talent is positively correlated with their parent's education conditioning on Z, that is, $E_P[\varepsilon X_{IV}|X_2, Z = z] \geq 0$ for all z. Such an*

[2] These discussions were confirmed by the empirical study in Cai, Fang, Lin, and Su (2019, see Figure 5), who found that the effect of schooling on earning (logarithm of hourly wage) increases monotonically in experiences using an index of labor market attitudes as the instrument.

[3] See the recent literature on testing on IV-validity, e.g., Kitagawa (2015), Huber and Mellace (2015), Mourifié and Wan (2017),and Kédagni and Mourifié (2020), Sun (2023), among others.

imperfect instrument leads to the following moment inequality model:

$$E_P[X_{IV}(Y - X_1\theta_{01}(Z) - X_2'\theta_{02}(Z))|X_2, Z = z] \geq 0 \quad a.s.\ X_2. \tag{2.4}$$

Together with the unconditional (with respect to X_1 *and* X_2*) mean restriction* $E[\varepsilon|Z = z] = 0$*, this forms a special case of our model in Equation (2.1) with* $X = (X_1, X_2', X_{IV}')'$ *and* $\theta_0(z) = (\theta_{01}(z), \theta_{02}'(z))'$*,* $p = 1$*, and* $k = 2$*. The parameter of interest is the partial effect of education on wage at a particularly given experience level* z_0*, which is the subvector* $\theta_{01}(z_0)$ *of* $\theta_0(z_0)$*.*

Example 2.2 (Interval Data) *Even if all the right-hand side variables* (X_1, X_2, Z) *in Equation (2.3) are exogenous and there is no endogeneity issue in estimating return to education, we may still not be able to point-identify the parameters if researchers only observe the wage bracket but not the wage itself. Interval-observed data is common in household-level datasets such as the Current Population Survey (CPS), and its implication on identification and inference in constant-coefficient models are well studied in the literature; see, for instance, Manski and Tamer (2002), Imbens and Manski (2004), and Kaido (2017). In this scenario, the following varying coefficient moment inequalities hold for any fixed* $Z = z_0 \in \mathcal{Z}$*, of* Z*:*

$$E_P[Y_u - X_1\theta_{10}(Z) - X_2\theta_{20}(Z)|X, Z = z_0] \geq 0 \quad a.s.\ X \text{ and} \tag{2.5}$$

$$E_P[X_1\theta_{10}(Z) + X_2\theta_{20}(Z) - Y_\ell|X, Z = z_0] \geq 0 \quad a.s.\ X. \tag{2.6}$$

We will offer a simulation study using this example to illustrate the use of our method in Section 4.

Example 2.3 (Entry Game) *In the literature on industrial organization, researchers often use discrete choice games to model firms' entry and exit behavior and study the competition effect. These models are often point-identified if researchers know a priori that the data are generated from the same equilibrium or covariates satisfy certain support conditions. However, if researchers prefer to be more robust on the equilibrium selection* *mechanism or the support conditions do not hold, the moment inequality approach offers an alternative (Ciliberto and Tamer, 2009). Meanwhile, the key parameter — the strength of the strategic interaction — can differ in different markets. For instance, Aradillas-López and Gandhi (2016, Section 6.3.4 and Figures 5-6) found that in the U.S. retail drugstore industry, the competition effect among Walgreens, CVS, and Rite Aid decreases with the market size (population). Our model can be useful in these applications, which we illustrate in a simulation study in Appendix D.2.*

Example 2.4 *Consider a gravity model where exporting firm i, for $i = 1, ..., N$, chooses between L destination countries in each period. Assuming away any inter-temporal dependence in export profits, firm's exporting decision can be characterized by a (simplified) static version of Morales, Sheu, and Zahler (2019)'s conditional moment inequality model, constructed based on the revealed preference principle, as follows*

$$E[(\pi_{il} - \pi_{il'})V_{il}(1 - V_{il'})|X_i, Z_i] \geq 0 \text{ for all pairs } (l, l') \in \{1, 2, ..., L\}^2 \text{ s.t. } l \neq l', \quad (2.7)$$

where π_{il} is the profits of exporting to an actually chosen destination l by firm i, while $\pi_{il'}$ is the potential profits of alternatively exporting to destination l'; V_{il} is a dummy variable with $V_{il} = 1$ indicating l is actually chosen by i, and $V_{il'}$ is defined similarly; Z_i represents firm size, and X_i is the vector of other firm characteristics. More specifically, the profits π_{il} equals the revenue r_{il} minus the costs c_{il} as usual, i.e., $\pi_{il} = r_{il} - c_{il}$. As shown by Chaney (2018), on firm-level, the distance elasticity of trade varies with firm size Z_i. This suggests the following specification for the revenue r_{il}:

$$r_{il} = \exp\left[\alpha_l + X_i'\beta + \rho(Z_i)D_{il}\right] + \epsilon_{il}, \text{ and } E(\epsilon_i|X_i, Z_i) = 0, \quad (2.8)$$

where D_{il} is a proxy of the distance between firm i and destination country l, with its varying coefficient $\rho(Z_i)$ representing the (varying) distance elasticity.[4] The costs c_{il} typi-

[4] D_{il} is commonly calculated as the distance of the shipment, which represents the geodetic distance between the population center of the city where firm i is located and the population center of its export destination l (Mayer and Zignago, 2011; Dingel, 2017; Almunia, Antràs, Lopez-Rodriguez, and Morales, 2021).

cally consist of "iceberg" trade costs and fixed costs. For its detailed specification, refer to Morales, Sheu, and Zahler (2019). For each (l, l') pair s.t. $l \neq l'$, substituting Equation (2.8) into Equation (2.7) yields a conditional moment inequality with varying coefficient $\rho(Z_i)$. And there are $L \times (L-1)$ such inequalities in total, the set of which constitute a special case of Model (2.1).

In practice, the pre-specified value z_0 is chosen by empirical needs. For example, researchers may be interested in the effects of competition in the drug industry in the median-size market or the average return to education in low-income provinces. Researchers may also be interested in making a joint inference on $\theta_{01}(z)$ over a collection of z: $\mathcal{Z}^T \equiv \{z_1, z_2, \cdots, z_T\}$.[5] We will analyze the statistical properties of these confidence sets in Section 3 and construct both types of confidence sets in our empirical application in Section 5.

3 Confidence Set

In this section, we propose a profiled test statistic for constructing confidence sets (CS) of subvectors of $\theta_0(z_0)$, for instance, the first component $\theta_{01}(z_0)$.[6] $z_0 \in \mathcal{Z}$ is a pre-specified value. A valid CS, denoted by $\widehat{CS}_n$, with confidence level $1-\alpha$ for $\theta_{01}(z_0)$ should satisfy that

$$\liminf_{n\to\infty} \inf_{(\theta_1, P)\in\mathcal{H}_0} Pr(\theta_1 \in \widehat{CS}_n) \geq 1-\alpha. \tag{3.1}$$

where $\mathcal{H}_0$ is a collection of (θ_1, P) and will be made specific later in Equation (3.9).

[5]In some empirical contexts, there are other natural choices of z_0. For example, let Z be the running variable of a fuzzy regression discontinuity design (FRD) and z_0 be the known cutoff. Under the local monotonicity and local continuity, the local average treatment effect (LATE) is identified at the cutoff z_0, see Imbens and Lemieux (2008). In this case, LATE is the key parameter, and the cutoff point z_0 is the natural choice of interest. If the FRD has multiple cutoffs $\mathcal{Z}^T \equiv \{z_1, z_2, \cdots, z_T\}$, then $\mathcal{Z}^T$ is the natural collection of interest. However, when the FRD assumptions are rejected, it is possible to partially identify the LATE at z_0 by relaxing the local continuity condition to the first-order stochastic dominance between the distributions of potential outcomes on either side of the cutoff.

[6]We can extend our method to the case in which researchers are interested in $\lambda(z_0) \equiv \lambda(\theta(z_0))$ for some function $\lambda : \Theta \to \Lambda \subseteq \mathbb{R}^{d_\lambda}$, as Bugni, Canay, and Shi (2017) for unconditional moment inequalities.

We first define a set of instrument functions to transform the conditional inequalities (in X) into unconditional ones. Without loss of generality, we assume that X contains only continuous variables and its support is $\mathcal{X} = [0,1]^{d_x}$.[7,8] We define a countable set of hyper-cubes in $\mathcal{X}$ as

$$\begin{aligned} \mathcal{G}_{\text{c-cube}} &= \left\{ g_\ell(\cdot) = 1(\cdot \in C_\ell) : \ell \equiv (x, r) \in \mathcal{L}_{\text{c-cube}} \right\}, \text{ where} \\ C_\ell &= \left(\times_{j=1}^{d_x} (x_j, x_j + r] \right) \text{ and} \\ \mathcal{L}_{\text{c-cube}} &= \left\{ (x, q^{-1}) : \ q \cdot x \in \{0, 1, 2, \cdots, q-1\}^{d_x}, \text{ and } q = 1, 2, \cdots \right\}. \end{aligned} \tag{3.2}$$

For notation simplicity, we let $C_1 = C_{(\mathbf{0},1)} = \mathcal{X}$ and $g_1 = g_{(\mathbf{0},1)} = 1$. One can also consider other instrument functions that satisfy Andrews and Shi (2013, Assumption CI). When there are discrete components in Z, we can apply our analysis to the subsample determined by the corresponding discrete component in z_0. If all components of Z are discrete, then we can apply Bugni, Canay, and Shi (2017)'s subvector inference procedure for constant coefficients to each subsamples. Therefore, we assume all the elements in Z are continuous variables without loss of generality and $f_z(\cdot)$ be its probability density function (pdf). Following the same argument in AS, the moment conditions in (2.1) are equivalent to

$$\begin{aligned} &\mu_{\ell,j}(\theta, z_0) \geq 0 \ \text{ for } j = 1, \ldots, p \text{ and} \\ &\mu_{\ell,j}(\theta, z_0) = 0 \ \text{ for } j = p+1, \ldots, k, \text{ for all } \ell \in \mathcal{L}. \end{aligned} \tag{3.3}$$

where $\mu_\ell(\theta, z_0) = E_P[g_\ell(X) \cdot m(W, \theta) | Z = z_0] \cdot f_z(z_0)$.

[7] Suppose $X = \{X_1, X_2\}$ in which X_1 is a binary variable taking values in $\{0,1\}$ and X_2 is a continuous variable. Then $E_P[m_j(W, \theta_0(z_1, z_2)) | X, Z = z] \geq (=)\ 0$ if and only if $E_P[m_j(W, \theta_0(Z)) \cdot 1(X_1 = 0) | X_2, Z = z] \geq (=)\ 0$ and $E_P[m_j(W, \theta_0(Z)) \cdot 1(X_1 = 1) | X_2, Z = z] \geq (=)\ 0$. In other words, by expanding the number of moment conditions, we can rewrite the model so that X_1 is not in the conditioning set and X_2 remains in it. Therefore, it is no loss of generality to assume that X contains only continuous variables.

[8] We can always normalize an observed x_{ij} to the unit interval by applying the transformation $\Phi\left(\frac{x_{ij} - \bar{x}_j}{\hat{\sigma}_{x,j}}\right)$, where Φ is the standard normal CDF, and $(\bar{x}_j, \hat{\sigma}_{x,j})$ are sample mean and standard deviation of observations $\{x_{1j}, x_{2j}, \cdots, x_{nj}\}$, respectively. Note that such normalization will not affect the asymptotics of our proposed test because the sample mean and standard deviation of observations converge at a faster rate than our proposed test statistics.

Let $K(\cdot)$ denote a kernel function with support on $[-1,1]^{d_z}$ and h_n is a bandwidth. For $j=1,\ldots,k$, define

$$\hat{\mu}_{\ell,n}(\theta,z_0)=\frac{1}{nh_n^{d_z}}\sum_{i=1}^{n}K\Big(\frac{Z_i-z_0}{h_n}\Big)g_\ell(X_i)\cdot m(W_i,\theta)$$

which, under the assumptions formally stated in the next section, is a consistent estimator for $\mu_\ell(\theta,z_0)$; with undersmoothing, $\sqrt{nh_n^{d_z}}(\hat{\mu}_{\ell,n}(\theta,z_0)-\mu_\ell(\theta,z_0))$ converges in distribution to a k-dimensional mean zero Gaussian process with covariance kernel $\rho_2\cdot Cov_P[g_{\ell^{(1)}}(X)\cdot m(W,\theta^{(1)}),g_{\ell^{(2)}}(X)\cdot m(W,\theta^{(2)})|Z=z_0]\cdot f_z(z_0)$, where the constant $\rho_2=\int_u K^2(u)du$. Let $\hat{\mu}_{1,n}(\theta,z_0)=n^{-1}h_n^{-d_z}\sum_{i=1}^n K\left(\frac{Z_i-z_0}{h_n}\right)m(W_i,\theta)$. We define

$$\begin{aligned}
\widehat{\Sigma}_n(\theta,1,z_0)&=\frac{1}{nh_n^{d_z}}\sum_{i=1}^{n}\Big(K\Big(\frac{Z_i-z_0}{h_n}\Big)(m(W_i,\theta)-\hat{\mu}_{1,n}(\theta,z_0)\Big)\Big(K\Big(\frac{Z_i-z_0}{h_n}\Big)m(W_i,\theta)-\hat{\mu}_{1,n}(\theta,z_0)\Big)',\\
\widehat{\Sigma}_n(\theta,\ell,z_0)&=\frac{1}{nh_n^{d_z}}\sum_{i=1}^{n}\Big(K\Big(\frac{Z_i-z_0}{h_n}\Big)g_\ell(X_i)m(W_i,\theta)-\hat{\mu}_{\ell,n}(\theta,z_0)\Big)\\
&\qquad\cdot\Big(K\Big(\frac{Z_i-z_0}{h_n}\Big)g_\ell(X_i)m(W_i,\theta)-\hat{\mu}_{\ell,n}(\theta,z_0))\Big)',\\
\widehat{\Sigma}_{\epsilon,n}(\theta,\ell,z_0)&=\widehat{\Sigma}_n(\theta,\ell,z_0)+\epsilon\cdot\text{diag}\left(\widehat{\Sigma}_n(\theta,1,z_0)\right).
\end{aligned}$$

Let $S(m,\Sigma)$ be a testing function, which can be chosen as one of the following two forms.

$$\begin{aligned}
S(m,\Sigma)&=\sum_{j=1}^{p}\left[\frac{m_j}{\sigma_j}\right]_-^2+\sum_{j=p+1}^{k}\left[\frac{m_j}{\sigma_j}\right]^2,\ \text{or}\\
S(m,\Sigma)&=\max\left\{\left[\frac{m_1}{\sigma_1}\right]_-^2,\ldots,\left[\frac{m_p}{\sigma_p}\right]_-^2,\left[\frac{m_{p+1}}{\sigma_{p+1}}\right]^2,\ldots,\left[\frac{m_k}{\sigma_k}\right]^2\right\}
\end{aligned}$$

where $[a]_-=\min\{0,a\}$ and $\sigma_j=\sqrt{\Sigma_{jj}}$. Then for a fixed value of θ_1, we can define the following Cramér-von-Mises-type (CvM) (profiled) test statistic as

$$\widehat{TS}_n(\theta_1,z_0)\equiv\inf_{\theta\in\Theta(\theta_1)}\widehat{T}_n(\theta,z_0),\tag{3.4}$$

where $\Theta(\theta_1)\equiv\{\tilde{\theta}\in\Theta:\ \tilde{\theta}_1=\theta_1\}$ is the possible value that the rest of parameters can take when the first parameter is fixed at θ_1, and

$$\widehat{T}_n(\theta, z_0) = \sum_{q=1}^{Q_n} \frac{1}{q^2+100} \sum_{\{\ell:\ r=q^{-1}\}} q^{-d_x} S(\sqrt{nh_n^{d_z}}\hat{\mu}_{\ell,n}(\theta, z_0), \widehat{\Sigma}_{\epsilon,\ell,n}(\theta, z_0)) \tag{3.5}$$

with $Q_n \to \infty$ as $n \to \infty$.[9]

Next, we approximate the distribution of $\widehat{TS}_n(\theta_1, z_0)$ to construct the critical value. We consider multiplier bootstrap. Let $\{U_i : i = 1, \dots, n\}$ be a sequence of pseudo-random variables with zero mean and unit variance that are independent of the sample path. The multiplier bootstrap process is

$$\Psi_n^u(\theta, \ell, z_0) = \frac{1}{\sqrt{nh_n^{d_z}}} \sum_{i=1}^{n} U_i \Big(K\Big(\frac{Z_i - z_0}{h_n}\Big) g_\ell(X_i) \cdot m(W_i, \theta) - \hat{\mu}_{\ell,n}(\theta, z_0) \Big).$$

Following Bugni, Canay, and Shi (2017), we define the slackness function as $\hat{\nu}_{\ell,n}(\theta, z_0) = \kappa_n^{-1}\sqrt{nh_n^{d_z}}\hat{\mu}_\ell(\theta, z_0)$, where $\kappa_n = \sqrt{\log(n)}$. The bootstrap version of simulated CvM test statistic for θ as

$$\widehat{T}_n^u(\theta, z_0) = \sum_{q=1}^{Q_n} \frac{1}{q^2+100} \sum_{\{\ell:\ r=q^{-1}\}} q^{-d_x} S\big(\Psi_n^u(\theta, \ell, z_0) + \hat{\nu}_{\ell,n}(\theta, z_0), \widehat{\Sigma}_{\epsilon,\ell,n}(\theta, z_0)\big).$$

And for a fixed value of θ_1, the bootstrap test statistic is[10]

$$\widehat{TS}_n^u(\theta_1, z_0) \equiv \min_{\theta \in \Theta(\theta_1)} \widehat{T}_n^u(\theta, z_0).$$

For a fixed positive number η, for example, 10^{-6}, define $\widehat{C}_{\eta,n}(\theta_1, \alpha)$ as the $(1 - \alpha + \eta)$-th

[9] Note that our test with non-standardized moment conditions would still work. That is, our test is still valid if we replace $\widehat{\Sigma}_{\epsilon,\ell,n}(\theta, z_0)$ with the identity matrix in (3.4). In the main text, we consider the standardized version. In Appendix D.3, we also report the CS with non-standardized moment conditions for our empirical application, and the results are similar qualitatively.

[10] The statistic $\widehat{TS}_n^u(\theta_1)$ defined here is analogous to the statistic $T_n^{PR}(\lambda_0)$ of (2.13) in Bugni, Canay, and Shi (2017). As we show later, critical value based on $\widehat{TS}_n^u(\theta_1)$ delivers valid inference. We might, in addition, consider an alternative bootstrap statistic $T_n^{DR}(\theta_1)$ analogous to their $T_n^{DR}(\lambda_0)$, and use $\min\{\widehat{TS}_n^{DR}(\theta_1), \widehat{TS}_n^u(\theta_1)\}$ for a potential power improvement. Please see discussions in Bugni, Canay, and Shi (2017, section 4.1) for a detailed discussion.

quantile of the conditional distribution of $\widehat{TS}_n^u(\theta_1)$ given data plus η, i.e.,

$$\widehat{C}_{\eta,n}(\theta_1,\alpha)=\sup\Big\{C\Big|P^u\big(\widehat{TS}_n^u(\theta_1,z_0)\le C\big)\le 1-\alpha+\eta\Big\}+\eta. \tag{3.6}$$

The confidence set for $\theta_{0,1}(z_0)$ is then given as

$$\widehat{CS}_n=\{\theta_1:\widehat{TS}_n(\theta_1,z_0)\le\widehat{C}_{\eta,n}(\theta_1,\alpha)\}. \tag{3.7}$$

3.1 Asymptotics of Confidence Sets

Let $\{W_i\}_{i=1}^n$ denote a random sample of size n generated from P. Let $\mathcal{P}$ denote the set of P that we consider. Let F_z, F_x, and F_{xz} denote the marginal distributions of Z, X, and (X,Z) under P. We now introduce the regularity conditions for establishing the asymptotic properties of the proposed confidence sets in (3.7).

Assumption 3.1 *$\{(X_i,Y_i,Z_i)\}_{i=1}^n$ is a random sample of i.i.d. observations.*

Assumption 3.2 *Θ is compact and convex.*

One special case of Assumption 3.2 is that Θ is a Cartesian product of d_θ closed intervals $\Theta=\Pi_{j=1}^{d_\theta}[\theta_{j\ell},\theta_{ju}]$, in which case $\Theta(\theta_1)\equiv\{\tilde{\theta}\in\Theta:\ \tilde{\theta}_1=\theta_1\}$ is independent with θ_1, and that $\Theta_{-1}\equiv\Pi_{j=2}^{d_\theta}[\theta_{j\ell},\theta_{ju}]$. We next impose conditions on the moment functions $\{m_j(W,\theta):\theta\in\Theta\}$ for $j=1,\dots,k$ to regulate their complexity.

Assumption 3.3 *Assume that for fixed $\delta>0$ and $0<Q<\infty$ not depending on P,*

i. $\max_{j=1,\dots,k}\left|m_j(w,\theta)\right|\le M(w)$ *for all $w\in\mathcal{W}$, for all $\theta\in\Theta$ for some envelope function $M(w)$;*

ii. $E_P[M(W)^4|Z=z]\le Q<\infty$ *on $\mathcal{N}_\delta(z_0)$ for all $P\in\mathcal{P}$;*

iii. the processes $\{m_j(W_{n,i},\theta):\theta\in\Theta,i\le n,1\le n\}$ for $j=1,\dots,k$ are manageable with respect to the envelope functions $\{M(W_{n,i}):i\le n,1\le n\}$ where $\{W_{n,i}:i\le n,1\le n\}$ is a row-wise i.i.d. triangular array with $W_{n,i}\sim P_n$ for any sequence $\{P_n\in\mathcal{P}\}$.

Assumption 3.3 implies that $\{n^{-1/2}h_n^{-d_z/2}K((Z_i - z_0)/h_n) \cdot g_\ell(X_i)m_j(W_{n,i},\theta) : \theta \in \Theta, \ell \in \mathcal{L}, i \leq n, 1 \leq n\}$ are manageable with respect to the envelope functions $\{n^{-1/2}h_n^{-d_z/2}K((Z_i - z_0)/h_n) \cdot M(W_{n,i}) : i \leq n, 1 \leq n\}$.

Assumption 3.4 *For the same δ and Q as in Assumption 3.3, assume that*

i. $f_z(z) \geq \delta > 0$ and is continuous on $\mathcal{N}_\delta(z_0) \subset \mathcal{Z}$;

ii. $f_z(z)$ is twice continuously differentiable on $\mathcal{N}_\delta(z_0)$;

iii. $|f_z(z)| \leq Q$, $|f_z'(z)| \leq Q$ and $|f_z''(z)| \leq Q$ on $\mathcal{N}_\delta(z_0)$.

where $\mathcal{N}_\delta(z_0) = \mathcal{N}_\delta(z_0) \equiv \{z : \|z - z_0\| \leq \delta\}$.

Assumption 3.4 imposes some regularity conditions on the distribution of Z and assumes z_0 is in the interior of the support. We next impose smoothness conditions on the conditional moment conditions.

Assumption 3.5 *Let $\mu_j(\theta, x, z) = E_P[m_j(W,\theta)|X = x, Z = z]$. For all $x \in \mathcal{X}$, $\mu_j(\theta, x, z)$ is twice continuously differentiable on $\Theta \times \mathcal{N}_\delta(z_0)$. Furthermore, for all $x \in \mathcal{X}$, for the same δ and Q as in Assumption 3.3 and for all $j = 1, \ldots, k$,*

i. $\|\partial\mu_j(\theta, x, z)/\partial\theta\| \leq Q$ and $\|\partial^2\mu_j(\theta, x, z)/\partial\theta\partial\theta'\| \leq Q$ on $\Theta \times \mathcal{N}_\delta(z_0)$;

ii. $|\mu_j(\theta, x, z)| \leq Q$, $|\partial\mu_j(\theta, x, z)/\partial z| \leq Q$ and $|\partial^2\mu_j(\theta, x, z)/\partial z\partial z| \leq Q$ on $\Theta \times \mathcal{N}_\delta(z_0)$.

Assumption 3.6 *Assume that*

i. The $K(\cdot)$ is a non-negative symmetric bounded kernel with a compact support in R (say $[-1, 1]$).

ii. $\int K(u)du = 1$ and $\int u_j K(u)du = 0$.

iii. $h_n \to 0$, $nh_n \to \infty$ and $nh_n^{d_z+4} \to 0$ as $n \to \infty$.

Assumption 3.6 imposes conditions on kernel function and bandwidth. Assumption 3.6 (i)-(ii) are satisfied for commonly used second-order kernels. All of our results can be extended to higher-order kernels straightforwardly. Assumption 3.6(iii) requires undersmoothing, so the bias term is asymptotically negligible. This is standard practice for nonparametric estimators being asymptotically normally distributed with mean zero and is also adopted in AS.

Assumption 3.7 *Assume that* $\kappa_n \to \infty$ *and* $\kappa_n^2 n^{-1} h_n^{-d_z} \to 0$.

Assumption 3.7 specifies the condition for the slackness tuning parameter κ_n, and it is satisfied if κ_n is proportional to $\log(n)$, or a power of $\log(n)$.

Assumption 3.8 *Assume that uniformly over* $P \in \mathcal{P}$ *given in Assumption 3.3, the following hold,*

$$\lim_{\delta \downarrow 0} \sup_{\|(\theta^{(1)} - \theta^{(2)})\| \leq \delta} \sup_{\ell \in \mathcal{L}} \max_{j=1\ldots,k} \left| Var\big(g_\ell(X) \cdot (m_j(W, \theta^{(1)}) - m_j(W, \theta^{(2)})) \big| Z = z_0\big) \right| \to 0.$$

Assumption 3.8 is imposed to ensure that when along a (sub)sequence of distributions such that $\widehat{\Psi}_n(\theta, \ell, z_0) = \sqrt{nh_n^{d_z}}(\hat{\mu}_{\ell,n}(\theta, z_0) - \mu_\ell(\theta, z_0))$ weakly converges to a tight Gaussian process, the limiting process will have a continuous path in θ uniformly over $\ell \in \mathcal{L}$. Define population-level quantities:

$$\Sigma((\theta^{(1)}, \ell^{(1)}), (\theta^{(2)}, \ell^{(2)})) = \rho_2 \cdot Cov_P(g_{\ell^{(1)}}(X) \cdot m(W, \theta^{(1)}), g_{\ell^{(2)}}(X) \cdot m(W, \theta^{(2)}) | Z = z_0) \cdot f_z(z_0)$$
$$\Sigma((\theta, \ell)) = \Sigma((\theta, \ell), (\theta, \ell)),$$
$$\Sigma((\theta, 1)) = \rho_2 \cdot Cov_P(m(W, \theta), m(W, \theta) | Z = z_0) \cdot f_z(z_0),$$
$$\Sigma_\epsilon((\theta, \ell)) = \Sigma((\theta, \ell)) + \epsilon \cdot \Sigma((\theta, 1)),$$

and the population counterpart of the test statistics,

$$T_P(\theta, z_0) = \sum_{q=1}^{\infty} \frac{1}{q^2 + 100} \sum_{\{\ell:\ r = q^{-1}\}} q^{-d_x} S(\mu_\ell(\theta, z_0), \Sigma_{\epsilon,\ell}((\theta))). \tag{3.8}$$

Assumption 3.9 *Let* $\Theta(\theta_1) \equiv \{\tilde{\theta} \in \Theta : \tilde{\theta}_1 = \theta_1\}$ *and* $\Theta_P(z_0)$ *as defined in Equation (2.2). Let* $\mathcal{P}_0$ *be the collection of* $P \in \mathcal{P}$ *such that* $\Theta_P(z_0)$ *is not empty. Then for all* $P \in \mathcal{P}_0$ *and for all* $\theta \in \Theta(\theta_1)$, $T_P(\theta, z_0) \geq c \min\{\delta, \inf_{\tilde{\theta} \in \Theta(\theta_1) \cap \Theta_P(z_0)} \|\theta - \tilde{\theta}\|^2\}$ *for some constants* $c > 0$ *and* $\delta > 0$ *that are independent of* θ_1 *and* z_0.

Assumption 3.9 is an identification strength assumption. It is a type of polynomial minorant condition introduced by Chernozhukov, Hong, and Tamer (2007). A similar condition is also assumed in Bugni, Canay, and Shi (2017, Assumption A.3) for subvector inference in unconditional moment inequality models. This assumption excludes weakly identified models. For instance, it requires the instrumental and endogenous variables to have a correlation bounded away from zero.

We define $\mathcal{H}_0$ as the collection of (θ_1, P) such that $P \in \mathcal{P}$ and there exists a $\theta_{-1} \in \Theta_{-1}$ such that $(\theta_1, \theta_{-1}) \in \Theta_P(z_0)$. That is,

$$\mathcal{H}_0 \equiv \{(\theta_1, P) : P \in \mathcal{P}, \text{ exist } \theta_{-1} \in \Theta_{-1} \text{ such that } (\theta_1, \theta_{-1}) \in \Theta_P(z_0)\}. \tag{3.9}$$

Theorem 3.1 *Let the confidence level be* $1 - \alpha$. *Suppose Assumptions 3.1-3.9 hold, then*

$$\liminf_{n \to \infty} \inf_{(\theta_1, P) \in \mathcal{H}_0} Pr(\theta_1 \in \widehat{CS}_n) \geq 1 - \alpha. \tag{3.10}$$

In addition, if there exists $(\theta_1^*, P^*) \in \mathcal{H}_0$ *such that the limiting distribution function under* P^* *of* $\widehat{TS}_n(\theta_1, z_0)$ *is continuous and strictly increasing at its* $(1 - \alpha)$*-th quantile, then*

$$\lim_{\eta \downarrow 0} \liminf_{n \to \infty} \inf_{(\theta_1, P) \in \mathcal{H}_0} Pr(\theta_1 \in \widehat{CS}_n) = 1 - \alpha. \tag{3.11}$$

3.2 Joint Confidence Set

The confidence set characterized in the Theorem 3.1 depends on z_0. In some applications, researchers may be interested in a joint inference on $\theta_{01}(\cdot)$ evaluated at multiple pre-specified values: $\mathcal{Z}^T = \{z_1, z_2, \cdots, z_T\}$.[11] The results of Theorem 3.1 can be

[11] Researchers may also be interested in the confidence band for the functional parameter $\theta_{01}(\cdot)$. This is beyond the scope of this paper, and we will leave it for future research.

readily extended to analyze this case. One way to proceed is to define $\widehat{TS}_n^u(\tilde{\theta}_1^T, \mathcal{Z}^T) = \max_{t=1,2,\cdots,T} \widehat{TS}_n^u(\theta_{1t}, z_t)$ and the critical value $\widehat{C}_{\eta,n}^{joint}(\tilde{\theta}_1^T, \alpha)$ as

$$\widehat{C}_{\eta,n}^{joint}(\tilde{\theta}_1^T, \alpha) = \sup\Big\{C \Big| P^u\big(\widehat{TS}_n^u(\tilde{\theta}_1^T, \mathcal{Z}^T) \leq C\big) \leq 1-\alpha+\eta\Big\} + \eta,$$

where $\tilde{\theta}_1^T \equiv (\theta_{11}, \theta_{12}, \cdots, \theta_{1t}, \cdots, \theta_{1T})$ is a generic $T \times 1$ vector. The joint confidence set for $\{\theta_{01}(z_t) : t = 1, \ldots, T\}$ is then given as

$$\widehat{CS}_n^{joint} = \Big\{\tilde{\theta}_1^T : \max_{t=1,\ldots,T} \widehat{TS}_n(\theta_{1t}, z_t) \leq \widehat{C}_{\eta,n}^{joint}(\tilde{\theta}_1^T, \alpha)\Big\}. \tag{3.12}$$

where is $\widehat{TS}_n(\theta_{1t}, z_t)$ defined in the same way as in Equation (3.4).

Computing the joint confidence set $\widehat{CS}_n^{joint}$ given in Equation (3.12) can be time-consuming because one needs to search in the T-dimensional space. To see this, suppose $T = 10$ and for each z_t, and consider 100 grid points for $\theta_{01}(z_t)$. In this case, there are 100^{10} grid points for the vector $\tilde{\theta}_1^T$, and, consequently, one needs to invert the corresponding test 100^{10} times. When the number of z's being considered gets larger, it is almost impossible to compute such a joint confidence set. Therefore, we utilize the fact that for a finite number of different values of z's, $z_1, \ldots, z_T$, the confidence sets for $\theta_{01}(z_1), \ldots, \theta_{01}(z_T)$ are asymptotically mutually independent because when the bandwidth h gets smaller with sample size, we will use subsamples that are mutually exclusive to compute each confidence set. Then, a valid joint confidence set with 1-α confidence level for $\{\theta_{01}(z_t) : t = 1, \ldots, T\}$ is then given as

$$\widetilde{CS}_n^{joint} = \times_{t=1,\ldots,T} \widehat{CS}_n(z_t, \alpha_T), \tag{3.13}$$

where for each t, $\widehat{CS}_n(z_t, \alpha_T)$ is a valid confidence set with confidence level $1 - \alpha_T$ for $\theta_{01}(z_t)$ as in Equation (3.7) and $(1 - \alpha_T)^T = 1 - \alpha$. It is much less time consuming to compute $\widetilde{CS}_n^{joint}$ than $\widehat{CS}_n^{joint}$. Again, suppose $T = 10$ and for each z_t, we consider 100 grid points for $\theta_{01}(z_t)$, then to obtain $\widetilde{CS}_n^{joint}$, we only need to invert the test $10 \times 100 = 1000$

times.[12] Therefore, even if the number of z's we consider gets larger, it is still feasible to compute $\widetilde{CS}_n^{joint}$. The drawback of $\widetilde{CS}_n^{joint}$ is that when the number of of z's is too large and those z's can be very dense in $\mathcal{Z}$. In general, we would require a larger sample size so that the confidence sets at different z's are mutually independent.

3.3 Specification Test

In many empirical settings, given the set $\mathcal{Z}^T = \{z_1, \cdots, z_T\}$ of interest, researchers may want to examine whether the model is correctly specified over this set. To be specific, consider the following null hypothesis:

$$\mathcal{P}_0 \equiv \{P \in \mathcal{P} : \text{There exists a } \theta(\cdot) \text{ such that (2.1) holds for all } z \in \mathcal{Z}^T\}. \tag{3.14}$$

Note that the condition stated in (3.14) is a necessary condition of the stronger statement in (2.1), which requires the existence of a function $\theta_0(\cdot)$ such that the moment inequalities to hold for all $z \in \mathcal{Z}$.[13] For this reason, a rejection of (3.14) implies the rejection of the original model in (2.1), but not vice versa. Still, empirical researchers can consider testing (3.14) as a practical way of checking the model specification and can pick a larger number of grid points (of z) to make the testing result more credible. Note also that the null DGP set $\mathcal{P}_0$ implicitly depends on the grid points $\mathcal{Z}^T$, which we omit the dependence for the ease of notation.

For testing the H_0 of $P \in \mathcal{P}_0$ against H_1 of $P \in \mathcal{P}/\mathcal{P}_0$, one can certainly construct the confidence set for $\theta_0(z)$ and verify if this confidence set is empty. However, as discussed in Bugni, Canay, and Shi (2015), checking the emptiness of the confidence set can be unnecessarily costly in computation, and the test statistics defined as the infimum (or supremum) of an appropriate sample objective function can achieve better power.

[12] When the dimension of the parameter vector is high, instead of considering a fixed grid points, one can use the EAM algorithm of Kaido, Molinari, and Stoye (2019) to select testing points to reduce computation cost. However, the computation simplification of the product-confidence set still applies, in addition to the savings brought by the EAM algorithm.

[13] In this sense, we are testing a collection of local specifications instead of the global specification.

Therefore, we consider the following test statistics,

$$\widehat{T}_n \equiv \max_{t=1,\cdots,T}\left[\min_{\theta\in\Theta}\widehat{T}_n(\theta,z_t)\right],$$

and its bootstrapped analog

$$\widehat{T}_n^u \equiv \max_{t=1,\cdots,T}\left[\min_{\theta\in\Theta}\widehat{T}_n^u(\theta,z_t)\right].$$

We set the critical value $C^u_{\eta,n}(\alpha)$ as the $(1-\alpha+\eta)$-th quantile of $\widehat{T}^u_n$ plus η, and define the test be $\phi_n = 1[\hat{T}_n > C^u_{\eta,n}(\alpha)]$. It is easy to see that the test statistic $\widehat{T}_n$ and $C^u_{\eta,n}(\alpha)$ utilize respectively $\widehat{T}_n(\theta,z_t)$ and $\widehat{T}^u_n(\theta,z_t)$, both of which are used earlier for constructing CSs of (3.1). The following theorem establishes the consistency of the proposed procedure above for testing the null of (3.14).

Theorem 3.2 *Suppose Assumptions 3.1-3.9 hold, then*

$$\limsup_{n\to\infty}\sup_{P\in\mathcal{P}_0} Pr(\phi_n = 1) \le \alpha. \tag{3.15}$$

In addition, if these exists $P^* \in \mathcal{P}$ such that the limiting distribution function under P^* of $\hat{T}_n$ is continuous and strictly increasing at its $(1-\alpha)$-th quantile, then

$$\lim_{\eta\downarrow 0}\limsup_{n\to\infty}\sup_{P\in\mathcal{P}_0} Pr(\phi_n = 1) = \alpha. \tag{3.16}$$

Remark 3.1 *In calculating the quantile of $\widehat{T}^u_n$, one can replace the minimization region Θ with $\widehat{\Theta}_P(z_t)$, a consistent estimator of the identified set $\Theta_P(z_t)$. This would allow us to use other GMS functions. Please see Bugni, Canay, and Shi (2015, footnote 8) for more discussions on the choice of the minimization region and slackness functions.*

Corollary 3.1 *Fix $\mathcal{Z}^T = \{z_1, z_2, \cdots, z_T\}$. Suppose the conditions for Theorem 3.2 are satisfied for all $z \in \mathcal{Z}^T$. Let $T_P(\theta, z_t)$ be as defined in Equation (3.8) with z_t in place of*

z_0 and P_n be a sequence of DGP such that

$$c_n = \max_{t=1,\cdots,T} \inf_{\theta\in\Theta} T_{P_n}(\theta, z_t) > 0.$$

Then for any chosen $\eta < +\infty$, we have $\liminf_{n\to\infty} Pr(\phi_n = 1) = 1$ if $c_n \to c > 0$. If $nh^{d_z} c_n \to c > 0$, and let $r(c) \equiv \liminf_{n\to\infty} Pr(\phi_n = 1)$, then we have $\lim_{c\to+\infty} r(c) = 1$.

The condition $\max_{t=1,\cdots,T} \inf_{\theta\in\Theta} T_{P_n}(\theta, z_t) = c_n > 0$ is a high level condition. $c_n \to c \in (0, \infty)$ can occur if a moment inequality is violated at a particular z_t. For example, if for some $j = 1, \cdots, p$, $E_{P_n}[m_j(W, \theta_0(Z))|X, z = z_t] < -\delta < 0$ over a subset of $\tilde{\mathcal{X}}_{z_t}$ with $Pr(X \in \tilde{\mathcal{X}}_{z_t}|Z = z_t) > 0$, then we can expect $c_n \to c > 0$. It can also occur when $|E_p[m_j(W, \theta_0(Z))|X, z = z_t]| > \delta > 0$ over a subset of $\tilde{\mathcal{X}}_{z_t}$ with $Pr(X \in \tilde{\mathcal{X}}_{z_t}|Z = z_t) > 0$ for some $j = p+1, \cdots k$.

Remark 3.2 *Our specification test can also test other restrictions on the $\theta_0(z)$. For example, one may be interested in if $\theta_0(z) \equiv \theta_0$ for all $z \in \mathcal{Z}^T$, where θ_0 is an unknown constant but with a known possible region of S. To test this hypothesis, we can modify the test statistics to*

$$\widehat{T}_n \;\; \equiv \;\; \min_{\theta\in S} \left[\max_{t=1,\cdots,T} \widehat{T}_n(\theta, z_t) \right].$$

Another possible scenario is that researchers may impose a parametric assumption on $\theta_0(z)$ such that $\theta_0(z) \equiv \varphi(z, \gamma_0)$, where φ is known up to a finite-dimensional parameter $\gamma_0 \in \Gamma$. Then, the test statistics can be defined as

$$\widehat{T}_n \;\; \equiv \;\; \min_{\gamma\in\Gamma} \left[\max_{t=1,\cdots,T} \widehat{T}_n(\varphi(z_t, \gamma), z_t) \right].$$

In the above two cases, if the test rejects, then we can interpret it as either the initial moment inequalities are misspecified, or the extra parametric assumption on $\theta_0(z)$ is misspecified, or both.

4 Simulations

This section provides some Monte Carlo simulations to illustrate our method and demonstrate its finite sample performance. In Section 4.1, we mainly focus on the property of the proposed confidence sets. In Section 4.2, we investigate the property of the proposed specification test. We consider four sample sizes $n \in \{500, 1000, 2000, 4000\}$, set the number of bootstrap samples $B = 1000$, and the number of replications $R = 1000$.

There are several tuning parameters we need to decide on when implementing our tests. Here, we summarize our recommendations.

1. $\kappa_n = \sqrt{\log n}$ which is recommended by Andrews and Soares (2010) and Bugni, Canay, and Shi (2017).

2. We use the Epanechnikov kernel. As in Andrews and Shi (2014), we consider the bandwidth $h = \tau \times 4.68\hat{\sigma}_z n^{-2/7}$ with $\tau = 0.5$, where $\hat{\sigma}_z$ is the estimated standard deviation of Z_i.[14]

3. $\eta = 10^{-6}$ which is recommended by Andrews and Shi (2013) and Andrews and Shi (2014).

4. $\epsilon = 1/20$ which is recommended by Andrews and Shi (2014).

5. Q_n is set such that the smallest cube contains, on average, no smaller than 15 sample points.[15]

4.1 Finite Sample Performances of the CS

We illustrate our method by linear regression with interval observed outcomes that we introduced in Example 2.2. Again, the latent variable regression is given by

$$Y = X_1\theta_{10}(Z) + X_2\theta_{20}(Z) + \varepsilon,$$

[14] If Z is a vector with a generic element Z_d, $d = 1, 2, \cdots, d_z$, then we will use a product kernel and set the bandwidth for the d-th dimension as $h_d = \tau \times 4.68\hat{\sigma}_{z_d} n^{-2/7}$, where $\hat{\sigma}_{z_d}$ is the standard deviation of $Z_{i,d}$. We also try other τ values between 0.1 and 1 in our main simulation; the results are similar.

[15] In our simulations, we set $Q_n = 10$, and the average sample size for the smallest cube is around 27 when $n = 2000$.

where Y is not observe but known to belong to $[Y_\ell, Y_u]$. $X_1 \sim N(0,1)$, $X_2 \sim N(0,1)$, $Z \sim U[2,6]$, $\epsilon \sim N(0,1)$ are all mutually independent. For some $\delta > 0$, let $Y_u = \delta(\text{Ceil}[Y/\delta])$ and $Y_\ell = \delta(\text{Ceil}[Y/\delta] - 1)$, where $\text{Ceil}[x]$ rounds x to integer toward $+\infty$. Under this construction, the bracket length $Y_u - Y_\ell$ is exactly δ. We consider the following varying coefficients:

$$\theta_{10}(z) = (1.6 + 0.6z)e^{-0.4(z-3)^2}, \quad \theta_{20}(z) = 2(1 + \cos(z))$$

This specification of $\theta_{10}(z)$ is taken from Cai, Fang, Lin, and Su (2019). We focus on $z_0 = 4$, which implies the true value of $\theta_{10}(z_0)$ equals to 2.68. In this model, the upper and lower bounds of the identified set for $\theta_{10}(z_0)$ is $[\theta_{1,\ell b}, \theta_{1,ub}]$, where

$$\theta_{1,\ell b} = \inf_{\theta \in \Theta} \theta_1 \quad \text{s.t.} \quad E_P[Y_\ell | X, Z = z_0] \leq x^\intercal \theta \leq E_P[Y_u | X, Z = z_0], \quad \text{a.s. } X,$$

$$\theta_{1,ub} = \sup_{\theta \in \Theta} \theta_1 \quad \text{s.t.} \quad E_P[Y_\ell | X, Z = z_0] \leq x^\intercal \theta \leq E_P[Y_u | X, Z = z_0], \quad \text{a.s. } X.$$

For this linear regression with interval-observed outcome variable designs, we consider interval lengths $\delta = 0.5$, which implies the identified set to be $[2.6, 2.73]$.[16] We calculate coverage frequencies at 95% nominal levels for different values of θ_1 deviating away from the upper boundary of the identified set, that is, $\theta_{1,ub} + c$ for $c \geq 0$. In Figure 1, we plot the coverage frequency against the distance to the upper boundary c. We can see that the coverage frequency is no smaller than the nominal level at the upper bound ($c = 0$), which shows that our CS is asymptotically valid. We can also see that the coverage probability declines quickly when moving away from the identified set for a given sample size and also decreases quickly as the sample size increases for each given c value. This shows that our CS has a good finite sample power. The pattern of the coverage frequency near the lower boundary is similar and, therefore, omitted.

Next, we examine the finite sample performance of the joint CS characterized in

[16] The "approximated identified sets" reported here are calculated by evaluating sample objective functions with a very large sample size ($n = 100,000$) and $Q_n = 10$. Therefore, these sets are essentially approximations of the approximated identified region of the set of unconditional moment inequalities corresponding to $Q_n = 10$, and they should be larger than the true identified sets of the conditional moment inequalities. We also consider interval length $\delta = 0.1$ and $\delta = 1.0$. The results are qualitatively similar and therefore omitted to save space.

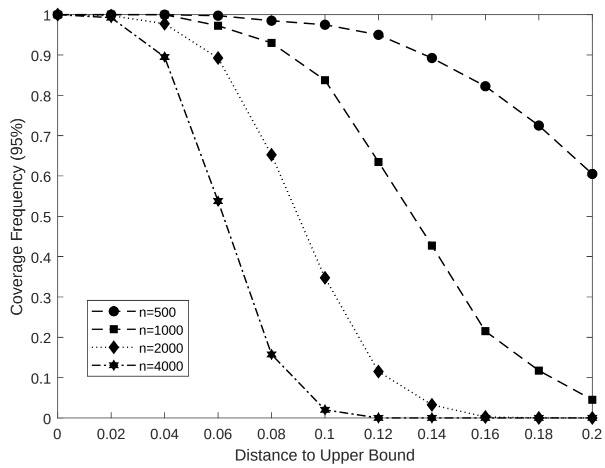


Figure 1: Coverage Frequency of $\widehat{CS}_n$ at a Single z

Equation (3.13). Instead of focusing on $z_0 = 4$, we consider a set of z values of $\mathcal{Z}^T = \{3.6, 4.0, 4.4, 4.8, 5.2\}$. Our goal is to construct a joint CS for the vector

$$\overrightarrow{\theta_{01}} \equiv (\theta_{01}(3.6), \theta_{01}(4.0), \cdots, \theta_{01}(5.2))' \in \mathbb{R}^5.$$

Note that the identified set for the vector $\overrightarrow{\theta_{01}}$ is a Cartesian product of the following form

$$[\theta_{1,\ell b}(3.6), \theta_{1,ub}(3.6)] \times [\theta_{1,\ell b}(4), \theta_{1,ub}(4)] \times \cdots \times [\theta_{1,\ell b}(5.2), \theta_{1,ub}(5.2)],$$

where $\theta_{1,\ell b}(z)$ and $\theta_{1,ub}(z)$ are the lower and upper bound of the identified set for $\theta_{01}(z)$. Because $\overrightarrow{\theta_{01}}$ is a multi-dimensional vector, it is difficult to draw the coverage probability for each $\overrightarrow{\theta_{01}}$. To have an intuitive comparison with the results in Figure 1, we report the coverage frequency of the joint CS for a deviation from the upper boundary of the identified set, namely $\overrightarrow{\theta_{1,ub}} + c\iota$, where $\overrightarrow{\theta_{1,ub}} \equiv (\theta_{1,ub}(3.6), \theta_{1,ub}(4), \cdots, \theta_{1,ub}(5.2))'$ is the upper boundary of the identified set (evaluated at $\mathcal{Z}$), ι is a vector of ones with same dimension as $\overrightarrow{\theta_{1,ub}}$, and $c \geq 0$ measures the size of the deviation. Increasing c again means that we are moving away from the identified set. Similar to the CS at a single z value, we can see from Figure 2 that the coverage frequencies decline as c increases for all the sample sizes, which shows that the joint CS also has a good finite sample power property.

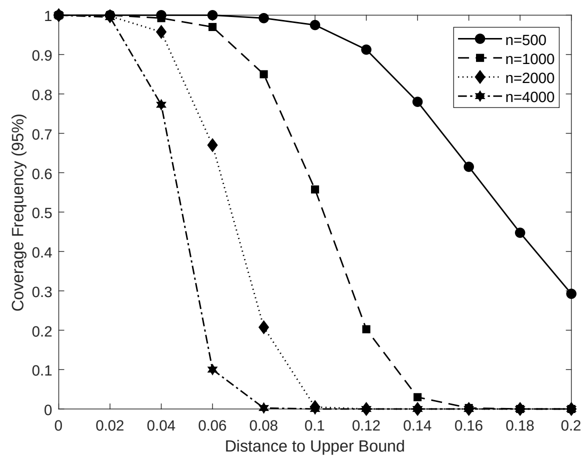


Figure 2: Coverage Frequency of the Joint CS

4.2 Specification Test

In this subsection, we examine the finite sample performance of our specification test, for which we also consider the interval data example but change the DGP so that the moment inequalities are misspecified. Specifically, the model is the same as the one in Section 4.1, except now we also consider cases in which $\delta < 0$. The model is mis-specified in such cases, and we should expect high rejection frequency. We conduct the test at the same five grid points $\mathcal{Z}^T = \{3.6, 4.0, 4.4, 4.8, 5.2\}$.

The following Table 1 reports the rejection frequencies under different significance levels α and δ. When the model is correctly specified and has a positive interval length ($\delta > 0$), the rejection frequency is very low and close to zero. This is not surprising because the true model lies in the "interior" of the null hypothesis. When the model is correctly specified but point-identified ($\delta = 0$), we are in the knife-edge case, and the rejection frequency is close to the nominal value when the sample size is large enough. Finally, when the model is misspecified ($\delta > 0$), our test can detect it and show good power—the rejection frequencies increase as the size of the misspecification increases.

Table 1: Rejection Frequency: Linear Regression with Interval Outcome

δ	n	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
$\delta = -1.0$	$n = 500$	1.000	0.997	0.952
	$n = 1000$	1.000	1.000	1.000
	$n = 2000$	1.000	1.000	1.000
$\delta = -0.5$	$n = 500$	0.992	0.967	0.715
	$n = 1000$	1.000	1.000	0.992
	$n = 2000$	1.000	1.000	1.000
$\delta = -0.2$	$n = 500$	0.650	0.420	0.112
	$n = 1000$	0.917	0.735	0.335
	$n = 2000$	0.997	0.975	0.665
$\delta = 0.0$	$n = 500$	0.117	0.057	0.012
	$n = 1000$	0.102	0.035	0.005
	$n = 2000$	0.145	0.067	0.005
$\delta = 0.2$	$n = 500$	0.015	0.002	0.0050
	$n = 1000$	0.005	0.002	0.000
	$n = 2000$	0.005	0.000	0.000
$\delta = 0.5$	$n = 500$	0.002	0.000	0.000
	$n = 1000$	0.000	0.000	0.000
	$n = 2000$	0.000	0.000	0.000
$\delta = 1.0$	$n = 500$	0.000	0.000	0.000
	$n = 1000$	0.000	0.000	0.000
	$n = 2000$	0.000	0.000	0.000

5 Empirical Illustration

In this section, we illustrate our method by estimating the return to education using a subset of China's 2005 "1% population census", which is also known as the "mini-census". It is well documented that the return to education in China is heterogeneous across regions with different development levels, and it is very crucial for policy-makers to account for such heterogeneity when designing new policies (see discussions in Heckman, 2005). On the other hand, because of the endogeneity of education, researchers have been employing the IV approach to identify the causal effect, where the parents' education is often used as the instrumental variable. For example, using the mother and father's education as one of the key IVs, Heckman and Li (2004) estimated that in China, a four-year college education increases wages by about 43%.[17] However, Liu, Mourifié, and Wan (2020, Table S1) found that one needs to be cautious about the validity of the parent's education as the IV for some provinces. In this empirical scenario, our method would be useful to partially identify the causal effect of a return to schooling under a weaker assumption while nonparametrically accounting for its dependence on the regional development level.

After matching children with their parents, our data set contains 176,458 individuals between 18 and 60 years of age. It covers all 31 provinces of China and 343 prefectures. For illustration, we retain the subsample for which the IV-validity was rejected in Liu, Mourifié, and Wan (2020), which results in 44,112 observations.[18] The core variables are the logarithm of the monthly wage (outcome variable Y) in 2005 Chinese Yuan, a prefecture-level average of the logarithm of monthly income (Z), an education level (X_1), and the mother's education level X_{IV}. Both education levels are classified into three categories: elementary school and below, middle school, high school and above. In this exercise, we use local (prefecture) level contemporaneous average income as the proxy for the regional development level. The descriptive statistics are reported in Table 2.

Table 2: Descriptive Statistics

Variables	Average	Std	Max	Min
Log-wage (Y)	6.26	0.89	10.5	2.30
Local Average Income (Z)	6.51	0.51	7.48	5.33
Education (X_1)	1.09	0.69	2	0
Mother's Education (X_{IV})	0.31	0.58	2	0

We consider the model that we discussed earlier in Equation (2.4),

$$E_P[X_{IV}(Y - X_1\theta_{01}(Z) - \theta_{02}(Z))|Z = z] \geq 0,$$
$$E_P[Y - X_1\theta_{01}(Z) - \theta_{02}(Z)|Z = z] = 0$$

We create a grid of $\mathcal{Z}^T = \{5.8, \cdots, 6.7, 6.8, 7.1, 7.2, 7.3\}$ and construct a 95% joint CS

[17] They used the data from the China Urban Household Income and Expenditure Survey(CUHIES) for 2000.

[18] These provinces are Shanghai, Hubei, Guangdong, Chongqing, Xizang, and Qinghai.

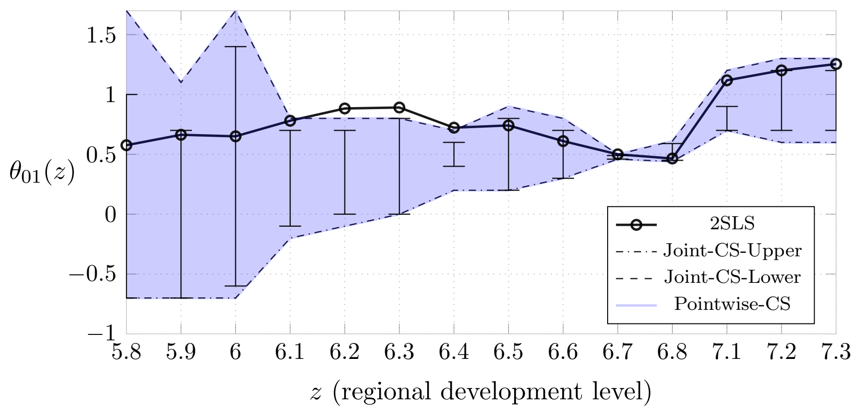


Figure 3: Confidence Intervals (95%) for Return to Education

for $\theta_{01}(z)$ with $z \in \mathcal{Z}^T$.[19] The choice of tuning parameters is the same as those in our simulation studies, except that we increase the number of bootstraps to 8000 to increase accuracy. We specify the parameters space as $[-1, 2] \times [4, 6]$. The results are reported in Figure 3. We connect the upper and lower points of the joint confidence sets and plot it as the (blue) shadowed area. As a comparison, for each $z \in \mathcal{Z}^T$, we also plot the pointwise confidence intervals as the black vertical lines. Finally, the black line with circle markers plots the two-stage least square estimator $\hat{\theta}_{2SLS}(z)$ using the observations whose $Z_i \in [z-0.1, z+0.1]$. We chose the half-window length as 0.1 to ensure the matrices in the 2SLS are not rank-deficient; we also tried larger numbers and obtained similar results. The 2SLS estimate using the entire sample (not binning on z) is 1.02, with a 95% confidence interval of $[0.990, 1.045]$. As a robustness analysis, we also calculate the confidence sets without standardizing the moments and using subsamples defined by gender and age. The results are similar and collected in Figures 6 and 7 in the Appendix D.3.

We have two observations from Figure 3. First, there is indeed substantial heterogeneity in the return to education. If one is willing to assume IV validity, then the 2SLS estimates suggest that the return to education varies substantially across different local

[19]In this example, the bandwidth $h \approx 0.056$, and there are few overlap observations when constructing confidence intervals at each z. The grid does not contain the two points 6.9 and 7.0 because there are no observations within their h-neighbourhood. For this sample of 44,112 observations, the total computation time of brute-forth grid search is already manageable (a few hours); however, it can be much improved by more sophisticated algorithms, e.g. the EAM algorithm of Kaido, Molinari, and Stoye (2019).

development levels: the estimates range from 0.5 to 1.2, which is far wider than the 95% CI $[0.990, 1.045]$ of the pooled 2SLS. On the other hand, based on the moment inequality model, we can see that the location of confidence intervals for $\theta_{01}(z)$ also varies substantially across different values of z. The width of the confidence intervals changes significantly, too. This suggests that after conditioning on different values of z, the data and model offer different levels of identification power for the parameter of interest $\theta_{01}(z)$. These features will not be observed if we do not allow $\theta_{01}(z)$ to vary across z. Regardless of the point or partial identification approach, the results show the empirical virtue of considering a model that allows for varying coefficients.

The second observation based on Figure 3 is that the 2SLS estimates are close to the upper boundaries of the pointwise or joint CS for nearly all z values. Therefore, even if one considers the heterogeneity in the return to education, it is still possible to make misleading policy recommendations based on 2SLS when the IV validity assumption is violated. For example, our results show that the return to education can be much lower (even negative) for the relatively under-developed areas than the 2SLS estimates, which may result from a frictional labor market or weak infrastructure. A policy implication is that the government needs to improve the labor market conditions or local infrastructures before investing in education. Our model thus offers additional information on top of the traditional varying coefficient models.

6 Conclusion

This paper provides an inference procedure for varying coefficients defined by moment inequalities and/or equalities. The proposed procedure is based on multiplier-bootstrap and, as shown, can be readily used to construct confidence sets for the parameters' subvector of interest. We show the resulting confidence sets are asymptotically valid uniformly over a broad family of DGPs and robust to partial identification. We also propose a specification test for a set of necessary implications of the varying coefficient models we considered. We illustrate the proposed method in simulation and empirical studies.

Appendix

A Notations

We introduce more notations. Let Ω be a specified closed set of $k \times k$ covariance matrices. Recall that

$$\Sigma_P((\theta^{(1)}, \ell^{(1)}), (\theta^{(2)}, \ell^{(2)})) = \rho_2 \cdot Cov_P(g_{\ell^{(1)}}(X) \cdot m(W, \theta^{(1)}), g_{\ell^{(2)}}(X) \cdot m(W, \theta^{(2)})|Z = z_0) \cdot f_z(z_0)$$
$$\Sigma_P((\theta, \ell)) = \Sigma_P((\theta, \ell), (\theta, \ell)),$$
$$\Sigma_P((\theta, 1)) = \rho_2 \cdot Cov_P(m(W, \theta), m(W, \theta)|Z = z_0) \cdot f_z(z_0),$$
$$\Sigma_{P,\epsilon}((\theta, \ell)) = \Sigma_P((\theta, \ell)) + \epsilon \cdot \Sigma_P((\theta, 1)),$$
$$\mu_\ell(\theta, z_0) = E_P(m(W, \theta)g_\ell(X)|Z = z_0) \cdot f_z(z_0).$$

For a given pair of $(\ell^{(1)}, \ell^{(2)})$, let $\mathcal{C}(\Theta^2)$ denote the space of continuous functions $\Sigma_P((\cdot, \ell^{(1)}), (\cdot, \ell^{(2)})) : \Theta^2 \to \Omega$. For notation simplicity, we write Σ_P to denote $\Sigma_P((\theta^{(1)}, \ell^{(1)}), (\theta^{(2)}, \ell^{(2)}))$ when it causes no confusion.

For a given θ_1, define

$$\Lambda_{n,P}(\theta_1) = \{(\theta, \xi) \in \Theta(\theta_1) \times \{R^k_{\pm\infty}\}_{\ell \in \mathcal{L}} : \xi_\ell = \sqrt{nh_n^{d_z}}\mu_\ell(\theta, z_0)\},$$
$$\Lambda^*_{n,P}(\theta_1) = \{(\theta, \xi) \in \Theta(\theta_1) \times \{R^k_{\pm\infty}\}_{\ell \in \mathcal{L}} : \xi_\ell = \kappa_n^{-1}\sqrt{nh_n^{d_z}}\mu_\ell(\theta, z_0)\},$$
$$\widehat{\Lambda}^*_{n,P}(\theta_1) = \{(\theta, \xi) \in \Theta(\theta_1) \times \{R^k_{\pm\infty}\}_{\ell \in \mathcal{L}} : \xi_\ell = \kappa_n^{-1}\sqrt{nh_n^{d_z}}\hat{\mu}_\ell(\theta, z_0)\}.$$

where $\mu_\ell(\theta, z_0) = E_P[g_\ell(X) \cdot m(W, \theta)|Z = z_0] \cdot f_z(z_0)$.

For any two points (θ, ξ) and (θ', ξ') in $\Theta \times \{R^k_{\pm\infty}\}_{\ell \in \mathcal{L}}$, define the metric as

$$d((\theta, \xi), (\theta', \xi')) = \Big[\sum_{j=1}^{d_\theta}(\Phi(\theta_j) - \Phi(\theta'_j))^2 + \sum_{q=1}^{\infty}\frac{1}{q^2 + 100}\sum_{\{\ell:\ r=q^{-1}\}} q^{-d_x}\sum_{j=1}^{k}(\Phi(\xi_{j,\ell}) - \Phi(\xi'_{j,\ell}))^2\Big]^{1/2},$$

where $\Phi(\cdot)$ is the CDF of the standard normal. Then it is true that the space $(\Theta \times \{R^k_{\pm\infty}\}_{\ell\in\mathcal{L}}, d)$ constitutes a compact metric space because that $R_{\pm\infty}$ is a compact space under metric d_R with $d_R(r,r') = |\Phi(r) - \Phi(r)|$, $r, r' \in R_{\pm\infty}$. Let $\mathcal{S}(\Theta \times \{R^k_{\pm\infty}\}_{\ell\in\mathcal{L}})$ denote the collection of compact subsets of the metric space $(\Theta \times \{R^k_{\pm\infty}\}_{\ell\in\mathcal{L}}, d)$. Note that this is true only when the dimension of $\Theta \times \{R^k_{\pm\infty}\}_{\ell\in\mathcal{L}}$ is countably many infinite and this is the main reason that we have to use instrument functions $\mathcal{G}_{\text{c-cube}}$ that is countably many. Let d_H denote the Hausdorff metric associated to the metric d, i.e., for any sets $A, B \subseteq \Theta \times \{R^k_{\pm\infty}\}_{\ell\in\mathcal{L}}$,

$$d_H(A,B) = \max\Big\{ \sup_{(\theta,\xi)\in A} \inf_{(\theta',\xi')\in B} d((\theta,\xi),(\theta',\xi')), \sup_{(\theta',\xi')\in B} \inf_{(\theta,\xi)\in A} d((\theta,\xi),(\theta',\xi')) \Big\}.$$

At last, define the metric space $(\Theta(\theta_1) \times \{R^k_{\pm\infty}\}_{\ell\in\mathcal{L}}, d)$ and the collection of its compact subsets $\mathcal{S}(\Theta(\theta_1) \times \{R^k_{\pm\infty}\}_{\ell\in\mathcal{L}})$ analogously.

B Lemmas

In this section, we abbreviate $\widehat{TS}_n(\theta_1, z_0)$ as $\widehat{TS}_n(\theta_1)$ when it causes no confusion; but it is understood that the test statistic depends on the pre-chosen z_0 value.

Lemma B.1 *Suppose Assumptions 3.1-3.9 hold. Let $\{(\lambda_{u_n}, P_{u_n} \in \mathcal{H}_0)\}_{n\geq 1}$ be a (sub)sequence of parameters and distributions such that for some $(\Sigma, \Lambda_{\mathcal{L}}) \in \{\mathcal{C}(\theta^2)\}_{(\ell_1,\ell_2)\in\mathcal{L}^2} \times \mathcal{S}(\Theta \times \{R^k_{\pm\infty}\}_{\ell\in\mathcal{L}})$, (i) $\Sigma_{P_{u_n}} \to \Sigma$ uniformly and (ii) $\Lambda_{u_n,P_{u_n}}(\theta_{u_n}) \overset{H}{\to} \Lambda_{\mathcal{L}}$. Then, along the (sub)sequence,*

$$\widehat{TS}_{u_n}(\theta_{1,u_n}) \overset{d}{\to} \inf_{(\theta,\lambda_{\mathcal{L}})\in\Lambda_{\mathcal{L}}} \sum_{q=1}^{\infty} \frac{1}{q^2+100} \sum_{\{\ell:\ r=q^{-1}\}} q^{-d_x} S(\Psi_\Sigma(\theta,\ell) + \lambda_\ell, \Sigma_\epsilon(\theta,\ell)), \tag{B.1}$$

where $\Psi_\Sigma : \Theta \times \mathcal{L} \to R^k$ is a R^k-valued tight Guassian process with covariance kernel $\Sigma \in \mathcal{C}(\theta^2)$, and $\Sigma_\epsilon = \Sigma(\theta,\ell) + \epsilon\Sigma(\theta,1)$.

Proof. Without loss of generality, we let $u_n = n$. Recall that

$$\widehat{TS}_n(\theta_1) \equiv \inf_{\theta \in \Theta(\theta_1)} \widehat{T}_n(\theta, z_0),$$

where $\Theta(\theta_1) \equiv \{\tilde{\theta} \in \Theta: \ \tilde{\theta}_1 = \theta_1\}$ and

$$\widehat{T}_n(\theta, z_0) = \sum_{q=1}^{\infty} \frac{1}{q^2+100} \sum_{\{\ell: \ r=q^{-1}\}} q^{-d_x} S(\sqrt{nh_n^{d_z}}\hat{\mu}_n(\theta, \ell, z_0), \widehat{\Sigma}_{\epsilon,n}(\theta, \ell, z_0)).$$

Let $\widehat{\Psi}_n(\theta, \ell, z_0) = \sqrt{nh_n^{d_z}}(\hat{\mu}_{\ell,n}(\theta, z_0) - \mu_\ell(\theta, z_0))$. We have

$$\begin{aligned}\widehat{TS}_n(\theta_1) &= \inf_{\theta \in \Theta(\theta_1)} \sum_{q=1}^{\infty} \frac{1}{q^2+100} \sum_{\{\ell: \ r=q^{-1}\}} q^{-d_x} S(\sqrt{nh_n^{d_z}}\hat{\mu}_{\ell,n}(\theta, z_0), \widehat{\Sigma}_{\epsilon,n}(\theta, \ell, z_0)) \\ &= \inf_{(\theta,\xi) \in \Lambda_{n,P}(\theta_1)} \sum_{q=1}^{\infty} \frac{1}{q^2+100} \sum_{\{\ell: \ r=q^{-1}\}} q^{-d_x} S(\widehat{\Psi}_n(\theta, \ell, z_0) + \xi_\ell, \widehat{\Sigma}_{\epsilon,n}(\theta, \ell, z_0)).\end{aligned}$$

For a generic uniform continuous function $\gamma : \Theta \times \mathcal{L} \to \mathbb{R}^K$, define

$$\begin{aligned}g_n(\gamma(\cdot), \Sigma(\cdot)) &\equiv \inf_{(\theta,\xi) \in \Lambda_{n,P}(\theta_1)} \sum_{q=1}^{\infty} \frac{1}{q^2+100} \sum_{\{\ell: \ r=q^{-1}\}} q^{-d_x} S(\gamma(\theta, \ell) + \xi_\ell, \Sigma_\epsilon(\theta, \ell)), \text{ and} \\ g(\gamma(\cdot), \Sigma(\cdot)) &\equiv \inf_{(\theta,\xi) \in \Lambda_{\mathcal{L}}(\theta_1)} \sum_{q=1}^{\infty} \frac{1}{q^2+100} \sum_{\{\ell: \ r=q^{-1}\}} q^{-d_x} S(\gamma(\theta, \ell) + \xi_\ell, \Sigma_\epsilon(\theta, \ell)).\end{aligned}$$

Let $\{\gamma_n(\cdot), \Sigma_n(\cdot)\}_{n \geq 1}$ be a sequence of functions such that

$$\lim_{n \to \infty} \sup_{\theta \in \Theta(\theta_1)} \sum_{q=1}^{\infty} \frac{1}{q^2+100} \sum_{\{\ell: \ r=q^{-1}\}} q^{-d_x} \|(\gamma_n(\theta, \ell), \Sigma_n(\theta, \ell)) - (\gamma_n(\theta, \ell), \Sigma(\theta, \ell))\| = 0$$

where $\| \cdot \|$ denotes the Euclidean norm, then by the same argument of Theorem 3.1 of Bugni, Canay, and Shi (2015), we can show that

$$\lim_{n \to \infty} g_n(\gamma_n(\cdot), \Sigma_n(\cdot)) = g(\gamma(\cdot), \Sigma(\cdot)).$$

Therefore, Lemma B.1 holds following the extended continuous mapping theorem (Van

Der Vaart and Wellner, 1996, Theorem 1.11.1) and by observing $\Psi_n \xrightarrow{d} \Psi_\Sigma$. $\square$

Lemma B.2 *Suppose Assumptions 3.1-3.9 hold. Let* $\{(\lambda_{u_n}, P_{u_n} \in \mathcal{H}_0)\}_{n\geq 1}$ *be a (sub)sequence of parameters and distributions such that for some* $(\Sigma, \Lambda^*_{\mathcal{L}}) \in \{\mathcal{C}(\theta^2)\}_{(\ell_1,\ell_2)\in\mathcal{L}^2} \times \mathcal{S}^2(\Theta \times \{R^k_{\pm\infty}\}_{\ell\in\mathcal{L}})$*, (i)* $\Sigma_{P_{u_n}} \to \Sigma$ *uniformly and (ii)* $\Lambda^*_{u_n,P_{u_n},\mathcal{L}}(\theta_{u_n}) \xrightarrow{H} \Lambda^*_{\mathcal{L}}$*. Then, there exists a further subsequence* $\{k_n\}_{n\geq 1}$ *of* $\{u_n\}_{n\geq 1}$*,*

$$\widehat{TS}^u_{k_n}(\theta_{k_n}) \xrightarrow{d} \inf_{(\theta,\lambda_{\mathcal{L}})\in\Lambda^*_{\mathcal{L}}} \sum_{q=1}^{\infty} \frac{1}{q^2+100} \sum_{\{\ell:\ r=q^{-1}\}} q^{-d_x} S(\nu_\Sigma(\theta,\ell)+\lambda_\ell, \Sigma_\epsilon(\theta,\ell)), \tag{B.2}$$

conditional on the sample path almost surely.

Proof. First, by (ii) of Lemma B.5, we have

$$\sup_{(\theta,\ell))\in(\Theta(\theta_1),\mathcal{L})} \|\widehat{\Sigma}_n((\theta,\ell)) - \Sigma_P((\theta,\ell))\| \xrightarrow{p} 0,$$

and this is sufficient to show that

$$\sup_{(\theta,\ell))\in(\Theta(\theta_1),\mathcal{L})} \|\widehat{\Sigma}_{\epsilon,n}((\theta,\ell)) - \Sigma_{\epsilon,P}((\theta,\ell))\| \xrightarrow{p} 0.$$

Second, note that

$$\kappa_n^{-1}\sqrt{nh_n^{d_z}}\hat{\mu}_\ell(\theta,z_0) = \kappa_n^{-1}\widehat{\Psi}_n(\theta,\ell,z_0) + \kappa_n^{-1}\sqrt{nh_n^{d_z}}\mu_\ell(\theta,z_0)$$

and by (i) of Lemma B.5 and the fact that $\kappa_n^{-1} \to 0$, we have $d_H(\Lambda^*_{n,P}(\theta_1), \widehat{\Lambda}^*_{n,P}(\theta_1)) \xrightarrow{p} 0$. Then given that $d_H(\Lambda^*_{n,P}(\theta_1), \Lambda^*_{\mathcal{L}}) \to 0$, we have $d_H(\Lambda^*_{n,P}(\theta_1), \Lambda^*_{\mathcal{L}}) \xrightarrow{p} 0$.

Therefore, there exists a subsequence $\{k_n\}_{n\geq 1}$ of $\{n\}_{n\geq 1}$ such that (a) $\widehat{\Psi}_{k_n}(\cdot) \Rightarrow \Psi_\Sigma$ conditional on sample path almost surely, (b) $\sup_{(\theta,\ell))\in(\Theta(\theta_1),\mathcal{L})} \|\widehat{\Sigma}_n((\theta,\ell)) - \Sigma_P((\theta,\ell))\| \xrightarrow{a.s.} 0$ and (c) $d_H(\Lambda^*_{n,P}(\theta_1), \Lambda^*_{\mathcal{L}}) \xrightarrow{a.s.} 0$. Then by the same proof of Lemma B.1 and by conditional on the sample path, we have

$$\widehat{TS}^u_{k_n}(\theta_{k_n}) \xrightarrow{d} \inf_{(\theta,\lambda_{\mathcal{L}})\in\Lambda^*_{\mathcal{L}}} \sum_{q=1}^{\infty} \frac{1}{q^2+100} \sum_{\{\ell:\ r=q^{-1}\}} q^{-d_x} S(\nu_\Sigma(\theta,\ell)+\lambda_\ell, \Sigma_\epsilon(\theta,\ell)),$$

conditional on the sample path almost surely.□

Lemma B.3 *Let* $\{(\theta_{1,u_n}, P_{u_n} \in \mathcal{H}_0)\}_{n\geq 1}$ *be a (sub)sequence of parameters and distributions such that for some* $(\Sigma, \Lambda_{\mathcal{L}}, \Lambda^*_{\mathcal{L}}) \in \{\mathcal{C}(\theta^2)\}_{(\ell_1,\ell_2)\in\mathcal{L}^2} \times \mathcal{S}^2(\Theta \times \{R^k_{\pm\infty}\}_{\ell\in\mathcal{L}})$*, (i)* $\Sigma_{P_{u_n}} \to \Sigma$ *uniformly, (ii)* $\Lambda_{u_n,P_{u_n},\mathcal{L}}(\theta_{1,u_n}) \xrightarrow{H} \Lambda_{\mathcal{L}}$ *and (iii)* $\Lambda^*_{u_n,P_{u_n},\mathcal{L}}(\theta_{1,u_n}) \xrightarrow{H} \Lambda^*_{\mathcal{L}}$*. Suppose Assumptions 3.1-3.9 hold. Then we have that for all* $(\theta, \xi^*) \in \Lambda^*_{\mathcal{L}}$ *such that* $\xi^*(\ell) \in R^p_{+\infty}(-\infty, \infty] \times R^{k-p}$ *for all* $\ell \in \mathcal{L}$ *with*

$$\sum_{q=1}^{\infty} \frac{1}{q^2+100} \sum_{\{\ell:\ r=q^{-1}\}} q^{-d_x} S(\xi^*(\ell), \Sigma_\epsilon(\theta,\ell)) < \infty,$$

there exists ξ *such that* $(\theta, \xi) \in \Lambda_{\mathcal{L}}$ *and* $\xi_j(\ell) \geq \xi^*_j(\ell)$ *for* $j \leq p$ *and* $\xi_j(\ell) = \xi^*_j(\ell)$ *for* $p < j \leq k$ *for all* $\ell \in \mathcal{L}$*.*

Proof. We apply the proof of Lemma S.3.8 of Bugni, Canay, and Shi (2017) to show our case. Without loss of generality, let $u_n = n$. If $(\theta, \xi^*) \in \Lambda^*_{\mathcal{L}}$, there exists a sequence $\{\theta_n\}$ such that $\theta_n \in \Theta(\theta_{1,n})$ with $\theta_n \to \theta$, and $\kappa_n^{-1}\sqrt{nh_n^{d_z}}\mu_\ell(\theta_n, z_0) \to \xi^*(\ell)$ for all $\ell \in \mathcal{L}$. Similar to (S.16) of Bugni, Canay, and Shi (2017), there exists a sequence of $\tilde{\theta}_n \in \Theta_{P_n}(\theta_{1,n}, z_0)$ such that $\|\theta_n - \tilde{\theta}_n\| \leq O(\kappa_n/\sqrt{nh_n^{d_x}})$. To see this, note that

$$\begin{aligned}\kappa_n^{-2} n h_n^{d_z} T_{P_n}(\theta_n, z_0) &= \sum_{q=1}^{\infty} \frac{1}{q^2+100} \sum_{\{\ell:\ r=q^{-1}\}} q^{-d_x} S(\kappa_n^{-1}\sqrt{nh_n^{d_z}}\mu_\ell(\theta_n, z_0), \Sigma_\epsilon(\theta,\ell)) \\ &\to \sum_{q=1}^{\infty} \frac{1}{q^2+100} \sum_{\{\ell:\ r=q^{-1}\}} q^{-d_x} S(\xi^*(\ell), \Sigma_\epsilon(\theta,\ell)) < \infty.\end{aligned}$$

Therefore, by Assumption 3.9,

$$O(\kappa_n^2 n^{-1} h_n^{-d_z}) = c^{-1} T_{P_n}(\theta_n, z_0) \geq \min\{\delta, \inf_{\tilde{\theta}\in\Theta(\theta_1)\cap\Theta_P(z_0)} \|\theta - \tilde{\theta}\|^2\},$$

and this further implies that there exists a sequence of $\tilde{\theta}_n \in \Theta_{P_n}(\theta_{1,n}, z_0)$ such that $\|\theta_n - \tilde{\theta}_n\| \leq O(\kappa_n/\sqrt{nh_n^{d_x}})$.

Define $\hat{\theta}_n = (1-\kappa_n^{-1})\tilde{\theta}_n + \kappa_n^{-1}\theta_n$. By the same arguments of (S.17) and (S.18), we have

$$\sqrt{nh_n^{d_z}}\mu_\ell(\hat{\theta}_n, z_0) = \kappa_n^{-1}\sqrt{nh_n^{d_z}}\mu_\ell(\theta_n, z_0) + \epsilon_{1,n}(\ell) + \epsilon_{2,n}(\ell)$$

where $\epsilon_{1,n}(\ell) = (\nabla_\theta\mu_\ell(\theta_n^{**}, z_0) - \nabla_\theta\mu_\ell(\theta_n^*, z_0))\sqrt{nh_n^{d_z}}(\hat{\theta}_n - \theta_n)$ with θ_n^* and θ_n^{**} both being between $\hat{\theta}_n$ and θ_n, and $\epsilon_{2,n}(\ell) = (1-\kappa_n^{-1})\sqrt{nh_n^{d_z}}\mu_\ell(\tilde{\theta}_n, z_0)$. Note that $\tilde{\theta}_n \in \Theta_{P_n}(\theta_{1,n}, z_0)$ and $\kappa_n^{-1} \to 0$, so it follows that $\epsilon_{2,n,j}(\ell) \geq 0$ for $j \leq p$ and $\epsilon_{2,n,j}(\ell) = 0$ for $j > p$ for all ℓ. Note that $\nabla_\theta\mu_\ell(\theta, z_0) = E[g_\ell(X)\mu_\ell(\theta, X, Z)|Z = z_0]$ and by Assumption 3.5 1., it is true that $\|\nabla_\theta\mu_\ell(\theta_n^{**}, z_0) - \nabla_\theta\mu_\ell(\theta_n^*, z_0)\| \leq cQ\|\theta_n^{**} - \theta_n^*\|$ for some positive constant c not depending on ℓ. Therefore, we have $\|\nabla_\theta\mu_\ell(\theta_n^{**}, z_0) - \nabla_\theta\mu_\ell(\theta_n^*, z_0)\| = o(1)$ uniformly over ℓ. By the fact that $\sqrt{nh_n^{d_x}}\|\hat{\theta}_n - \tilde{\theta}_n\| = O(1)$, we have uniformly over ℓ,

$$\|\epsilon_{1,n}(\ell)\| \leq \|(\nabla_\theta\mu_\ell(\theta_n^{**}, z_0) - \nabla_\theta\mu_\ell(\theta_n^*, z_0))\|\sqrt{nh_n^{d_x}}\|\hat{\theta}_n - \tilde{\theta}_n\| = o(1).$$

Given that the space $(\Theta \times \{R_{\pm\infty}^k\}_{\ell\in\mathcal{L}}, d)$ constitutes a compact metric space, it is true that there exists a subsequence $\{u_n\}$ of $\{n\}$ such that $\sqrt{u_n h_{u_n}^{d_z}}\mu_\ell(\hat{\theta}_{u_n}, z_0)$ and $\kappa_{u_n}^{-1}\sqrt{u_n h_{u_n}^{d_z}}\mu_\ell(\theta_{u_n}, z_0)$ converge for all ℓ. To be specific, $\{R_{\pm\infty}^k, d_k\}$ where for any two points $\delta_1, \delta_2 \in R_{\pm\infty}^k$, $d_k(\theta_1, \theta_2) = (\sum_{j=1}^k (\Phi(\theta_{1,j}) - \Phi(\theta_{2,j})^2)^{1/2}$ is a compact set. Note that because $\mathcal{L}$ is countable, we can order $\ell = 1, 2, \ldots$ with those ℓ's with smaller q being ordered first. For $\ell = 1$, then there exists a subsequence $\{a_{1,n}\}$ of $\{n\}$ so that

$$\xi_j(1) = \lim_{n\to\infty}\sqrt{a_{1,n}h_{a_{1,n}}^{d_z}}\mu_\ell(\hat{\theta}_{a_{1,n}}, z_0) \geq \lim_{n\to\infty}\kappa_{a_{1,n}}^{-1}\sqrt{a_{1,n}h_{a_{1,n}}^{d_z}}\mu_\ell(\theta_{a_{1,n}}, z_0) = \xi_j^*(1) \text{ for } j \leq p,$$
$$\xi_j(1) = \lim_{n\to\infty}\sqrt{a_{1,n}h_{a_{1,n}}^{d_z}}\mu_\ell(\hat{\theta}_{a_{1,n}}, z_0) = \lim_{n\to\infty}\kappa_{a_{1,n}}^{-1}\sqrt{a_{1,n}h_{a_{1,n}}^{d_z}}\mu_\ell(\theta_{a_{1,n}}, z_0) = \xi_j^*(1) \text{ for } j \leq p.$$

Similarly, for $\ell = 2$, there exists a subsequence $\{a_{2,n}\}$ of $\{a_{1,n}\}$ so that

$$\xi_j(2) = \lim_{n\to\infty}\sqrt{a_{2,n}h_{a_{2,n}}^{d_z}}\mu_\ell(\hat{\theta}_{a_{2,n}}, z_0) \geq \lim_{n\to\infty}\kappa_{a_{2,n}}^{-1}\sqrt{a_{2,n}h_{a_{2,n}}^{d_z}}\mu_\ell(\theta_{a_{2,n}}, z_0) = \xi_j^*(2) \text{ for } j \leq p,$$
$$\xi_j(2) = \lim_{n\to\infty}\sqrt{a_{2,n}h_{a_{2,n}}^{d_z}}\mu_\ell(\hat{\theta}_{a_{2,n}}, z_0) = \lim_{n\to\infty}\kappa_{a_{2,n}}^{-1}\sqrt{a_{2,n}h_{a_{2,n}}^{d_z}}\mu_\ell(\theta_{a_{2,n}}, z_0) = \xi_j^*(2) \text{ for } j \leq p.$$

Then we keep doing this for $\ell = 3, 4, \ldots$ and set $\{u_n\} = \{a_{n,n}\}$. This completes the proof.□

Lemma B.4 *Suppose Assumptions 3.1-3.9 hold. For any (sub)sequence* $\{(\theta_{u_n}, P_{u_n} \in \mathcal{H}_0)\}_{n\geq 1}$*, there exists a further subsequence* $\{k_n\}_{n\geq 1}$ *of* $\{u_n\}_{n\geq 1}$ *such that (i)* $\Sigma_{P_{k_n}} \to \Sigma$ *uniformly, (ii)* $\Lambda_{k_n,P_{k_n},\mathcal{L}}(\theta_{k_n}) \xrightarrow{H} \Lambda_{\mathcal{L}}$ *and (iii)* $\Lambda^*_{k_n,P_{k_n},\mathcal{L}}(\theta_{k_n}) \xrightarrow{H} \Lambda^*_{\mathcal{L}}$ *for some* $(\Sigma, \Lambda_{\mathcal{L}}, \Lambda^*_{\mathcal{L}}) \in \{\mathcal{C}(\theta^2)\}_{(\ell_1,\ell_2)\in\mathcal{L}^2} \times \mathcal{S}^2(\Theta \times \{R^k_{\pm\infty}\}_{\ell\in\mathcal{L}})$*.*

Proof. We apply the proof of Lemma D.7 of Bugni, Canay, and Shi (2015) to show our case. For $\ell = 1$, by the same arguments of Lemma D.7 of Bugni, Canay, and Shi (2015), we can show that there exists a subsequence $\{a_{1,n}\}$ of $\{n\}$ such that

$$\Sigma_{P_{a_{1,n}}}((\cdot,\ell_1),(\cdot,\ell_2)) \to \Sigma((\cdot,\ell_1),(\cdot,\ell_2)) \text{ uniformly for } \ell_1,\ell_2 \in \{1\},$$
$$\Lambda_{a_{1,n},P_{a_{1,n}},\ell}(\theta_{a_{1,n}}) \xrightarrow{H} \Lambda_\ell,$$
$$\Lambda^*_{a_{1,n},P_{a_{1,n}},\ell}(\theta_{a_{1,n}}) \xrightarrow{H} \Lambda^*_\ell,$$

for some $(\Sigma, \Lambda_{\mathcal{L}}, \Lambda^*_{\mathcal{L}}) \in \{\mathcal{C}(\theta^2)\}_{(\ell_1,\ell_2)\in\mathcal{L}^2} \times \mathcal{S}^2(\Theta \times \{R^k_{\pm\infty}\}_{\ell\in\mathcal{L}})$. For $\ell = 2$, we can show that there exists a subsequence $\{a_{2,n}\}$ of $\{a_{1,n}\}$ such that

$$\Sigma_{P_{a_{1,n}}}((\cdot,\ell_1),(\cdot,\ell_2)) \to \Sigma((\cdot,\ell_1),(\cdot,\ell_2)) \text{ uniformly for } \ell_1,\ell_2 \in \{1,2\},$$
$$\Lambda_{a_{2,n},P_{a_{2,n}},\ell}(\theta_{a_{2,n}}) \xrightarrow{H} \Lambda_\ell,$$
$$\Lambda^*_{a_{2,n},P_{a_{2,n}},\ell}(\theta_{a_{2,n}}) \xrightarrow{H} \Lambda^*_\ell.$$

Then we keep doing this for $\ell = 3, 4, \ldots$ and set $\{k_n\} = \{a_{n,n}\}$. This completes the proof.
□

Lemma B.5 *Suppose Assumptions 3.1-3.9 hold. Let* $\{P_{u_n} \in \mathcal{P}\}_{n\geq 1}$ *be a (sub)sequence of distributions such that for some* $\Sigma \in \{\mathcal{C}(\Theta^2)\}_{(\ell_1,\ell_2)\in\mathcal{L}^2}$, $\Sigma_{P_{u_n}} \to \Sigma$ *uniformly. Then, the following statements hold:*

(i) $\widehat{\Psi}_{u_n}(\cdot) \Rightarrow \Psi_\Sigma$*, where* Ψ_Σ *is a tight zero-mean Guassian process with covariance kernel*

Σ. *In addition, for any fixed $\epsilon > 0$, there exists a $\delta > 0$ such that*

$$Pr\Big(\sup_{\|\theta^{(1)}-\theta^{(2)}\|\leq\delta}\sup_{\ell\in\mathcal{L}}\|\Psi_\Sigma(\theta^{(1)},\ell)-\Psi_\Sigma(\theta^{(2)},\ell)\|\leq\epsilon\Big)=1.$$

(ii) We have

$$\sup_{(\theta^{(1)},\ell^{(1)}),(\theta^{(2)},\ell^{(2)})\in(\Theta(\theta_1),\mathcal{L})}\|\widehat{\Sigma}_n((\theta^{(1)},\ell^{(1)}),(\theta^{(2)},\ell^{(2)}))-\Sigma_P((\theta^{(1)},\ell^{(1)}),(\theta^{(2)},\ell^{(2)}))\|\xrightarrow{p}0,\ \textit{where}$$

$$\widehat{\Sigma}_n((\theta^{(1)},\ell^{(1)}),(\theta^{(2)},\ell^{(2)}))=\frac{1}{nh_n^{d_z}}\sum_{i=1}^{n}\Big(K\Big(\frac{Z_i-z_0}{h_n}\Big)g_{\ell^{(1)}}(X_i)m(W_i,\theta^{(1)})-\hat{\mu}_{\ell^{(1)},n}(\theta^{(1)},z_0)\Big)$$
$$\cdot\Big(K\Big(\frac{Z_i-z_0}{h_n}\Big)g_{\ell^{(2)}}(X_i)m(W_i,\theta^{(2)})-\hat{\mu}_{\ell^{(2)},n}(\theta^{(2)},z_0)\Big)'.$$

(iii) We have $\Psi_n^u(\cdot)\Rightarrow\Psi_\Sigma$ conditional on sample path with probability 1.

Proof. Parts (i) and (ii) are the same as those of Lemma AN3 of Andrews and Shi (2014). Given part (ii), the proof of part (iii) follows from the same argument of Theorem 4.1 of Hsu (2016). □

C Proof of Theorems

Proof of Theorem 3.1. Given Lemma B.1-Lemma B.5 above, the proof to Theorem 3.1 follows the same arguments of Equation (C.5) of Bugni, Canay, and Shi (2017), and we omit the details for brevity.

Proof of Theorem 3.2. The proof of Theorem 3.2 follows analogously from those in Theorem 3.1. In particular, the limiting distribution of $\min_{\theta\in\Theta}\widehat{T}_n(\theta,z_t)$ can be obtained in a similar way as in Lemma B.1. For a set of pre-chosen grid points $\{z_1,\ldots,z_T\}$, $\min_{\theta\in\Theta}\widehat{T}_n(\theta,z_t)$ are mutually asymptotically independent, so their asymptotic joint distribution is the product of their asymptotic marginal distributions. Finally, the max operator is a continuous function, so the limiting distribution of $\widehat{T}_n$ follows by continuous mapping theorem. The validity of multiplier bootstrap holds as shown in Lemma B.5.

The results in Corollary 3.1 hold because (i) the critical value $C_n^u(\alpha)$ is stochastically bounded, and (ii) $\frac{\widehat{T}_n}{nh^{d_z}} - c_n \xrightarrow{p} 0$.

D Additional Empirical and Simulation Results

In this appendix section, we report some additional simulations and empirical results.

D.1 Magnitude of $f_z(z_0)$.

Our Assumption 3.4 requires that $f_z(z_0) \geq \delta > 0$ in a neighboured of z_0. For a given instrument function g_ℓ, our test statistics involves estimating the conditional moment $\mu_\ell(\theta, z_0) = E[g_\ell(X_i)m(W_i, \theta)|Z = z_0]$. When $f_z(z_0)$ is small, there are fewer observations in the neighborhood of z_0. Given everything else equal, we expect that the confidence set for $\theta_{01}(z_0)$ will perform worse when $f_z(z_0)$ is small.

To verify this conjecture, we run a simulation that has the same design as Figure 1, except that we focus on $n = 2000$ and vary the underlying DGPs such that $f_z(z_0)$ varies. To be specific, we take Z to be a mixture of two independent uniform distributions Z_A and Z_B, where Z_A has a support of $[2, 3.5] \cup [4.5, 6]$ and Z_B has a support of $[3.5, 4.5]$, respectively. The mixing weight for Z_B, denoted by τ, takes values from the set $\{0.05, 0.1, 0.15, 0.2, 0.25\}$. Note that when $\tau = 0.25$, Z is a uniform distribution over $[2, 6]$, which is the same as the DGP considered in Figure 1. When τ is smaller, the density value $f_z(4)$ is lower.

Figure 4 plots the coverage probability for $\theta_1 \in [2.7, 2.9]$. Note that the upper boundary of the identified interval for $\theta_1(4)$ is approximately 2.73. We expect the coverage frequencies to decrease as θ_1 value moves away from the upper boundary. It is indeed true for all values of τ. However, when τ is small, the curve decreases slower, indicating that our confidence set has less power.

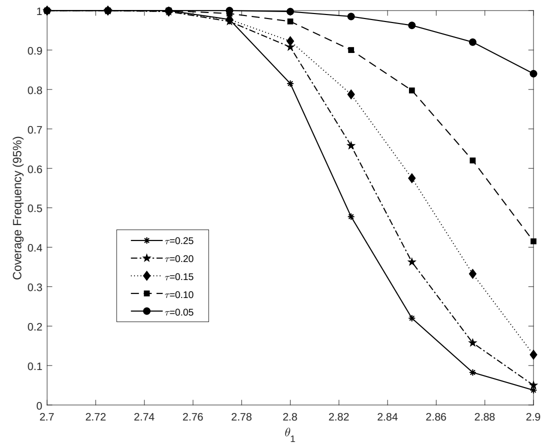


Figure 4: Coverage Frequency: Varying $f_z(z_0)$

D.2 Entry Game with Complete Information

In this section, we apply our method to a simple discrete choice game of complete information. Suppose two firms are making simultaneous binary choices:

$$Y_1 = 1\left\{\theta_{1,0}(Z)Y_2 - \varepsilon_1 \geq 0\right\};$$
$$Y_2 = 1\left\{\theta_{2,0}(Z)Y_1 - \varepsilon_2 \geq 0\right\},$$

where the coefficient $\theta_{1,0}(z) = -\frac{e^z-1}{e-1}$, $\theta_{2,0}(z) = -\frac{e^{1-z}-1}{e-1}$, $Z \sim U[0,1]$, and $(\varepsilon_1, \varepsilon_2) \sim \mathcal{N}\left(\left(\begin{smallmatrix}0\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}1 & \rho\\ \rho & 1\end{smallmatrix}\right)\right)$. In this model, the strength of the strategic interaction depends on the observed variable Z. We assume that players play a pure strategy Nash equilibrium, and when there are multiple equilibria, the nature tosses a fair coin to select. Researchers observe Y_1, Y_2 and Z, but do not know the functional form of $\theta_{j,0}(z)$, $j = 1, 2$. Researchers are also agnostic about the equilibrium selection mechanism.

Let $\Phi_\rho(t_1, t_2)$ be the probability of the event $\{\varepsilon_1 \leq t_1 \quad \& \quad \varepsilon_2 \leq t_2\}$. The necessary condition of Nash equilibrium implies the following conditional moment restrictions:

$$
\begin{aligned}
E_P\left[0.5-\Phi_\rho(\theta_{1,0}(Z),0)-(1-Y_1)Y_2 \mid Z=z\right] &\geq 0,\\
E_P\left[0.5-\Phi_\rho(0,\theta_{2,0}(Z))-Y_1(1-Y_2) \mid Z=z\right] &\geq 0,\\
E_P\left[\Phi_\rho(\theta_{1,0}(Z),\theta_{2,0}(Z))-Y_1Y_2 \mid Z=z\right] &= 0,\\
E_P\left[\Phi_\rho(0,0)-(1-Y_1)(1-Y_2) \mid Z=z\right] &= 0.
\end{aligned}
$$

In this model, the unknown parameters are $(\theta_{1,0}(\cdot),\theta_{2,0}(\cdot),\rho)$. However, ρ is identified from the fourth moment equality. Therefore, we solve ρ from the fourth equation and focus on the first three conditional moment restrictions:

$$\Phi_\rho(\theta_1,0) \leq 0.5-p(0,1|z), \tag{D.1}$$

$$\Phi_\rho(0,\theta_2) \leq 0.5-p(1,0|z), \tag{D.2}$$

$$\Phi_\rho(\theta_1,\theta_2) = p(1,1|z), \tag{D.3}$$

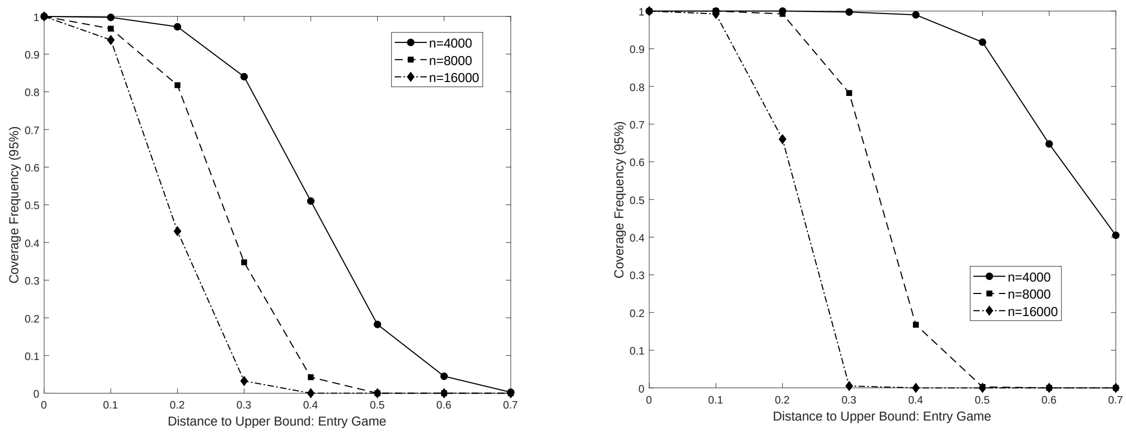
where $p(\ell,k|z) \equiv \Pr(Y_1=\ell,Y_2=k|Z=z)$. Note that given the joint normal distribution of epsilons, the upper and lower bound of the identified set for $\theta_{01}(z_0)$ can be analytically calculated from Equations (D.1) to (D.3). In particular, Equation (D.3) says that the joint identified set is a curve in the two-dimensional space. Equation (D.1) and Equation (D.2) provide the coordinates of the two endpoints of the curve.

D.2.1 Confidence sets

In this subsection, the first goal is to examine the performance of the confidence interval for $\theta_{1,0}(z_0)$ at $z_0=0.5$. Based on our calculation, when $\rho=0.5$, then true value is $\theta_{01}(0.5)=-0.3775$ and the identified set for $\theta_{01}(z_0)$ is $[-0.47,-0.29]$.

Figure 5a reports the coverage frequencies at 95% level under different sample sizes for $\theta_{1,ub}+c$ values when $\rho=0.5$, where $c\geq 0$ measures the distance of the testing value to the upper boundary of the identified set. We also considered other values of ρ and other significance levels but omitted the results due to the qualitative similarity. When c gets larger, the coverage frequencies decline dramatically and decline faster for larger sample sizes.

Next, we investigate the performance of the confidence set for $\theta_{01}(z)$, where $z \in \{0.2, 0.35, 0.5, 0.65, 0.8\}$. Similar to Section 4, we report the coverage frequency of the joint CS for $\overrightarrow{\theta_{1,ub}} + c\iota$, where $\overrightarrow{\theta_{1,ub}} \equiv (\theta_{1,ub}(0.2), \theta_{1,ub}(0.35), \cdots, \theta_{1,ub}(0.8))'$. The results are shown in Figure 5b. The patterns are similar to those reported for the single CS in that when we move away from the identified set, the joint coverage frequency also declines fast.



(a) Single z

(b) Multiple z

Figure 5: Coverage Frequency: Entry Game

D.2.2 Specification Test

To examine the performance of the specification test, We consider the same game and use the same set of inequalities, except that change the error terms $(\varepsilon_1, \varepsilon_2) \sim \mathcal{N}\left(\left(\begin{smallmatrix}-\delta \\ -\delta\end{smallmatrix}\right), \left(\begin{smallmatrix}1 & 0 \\ 0 & 1\end{smallmatrix}\right)\right)$. We assume that the researcher incorrectly parametrizes the joint distribution as to be $(\varepsilon_1, \varepsilon_2) \sim \mathcal{N}\left(\left(\begin{smallmatrix}0 \\ 0\end{smallmatrix}\right), \left(\begin{smallmatrix}1 & 0 \\ 0 & 1\end{smallmatrix}\right)\right)$. In this design, the size of δ measures the magnitude of the misspecification. For example, as $\delta \to +\infty$, the probability of the outcome $(0, 0)$ will converge to zero, but if under the misspecified model, for any given value of ρ, Our test rejects the model with large frequencies, and the rejection rate increases with both sample size n and misspecification magnitude δ. Table 3 reports the rejection frequencies when $\rho = 0$. When $\delta = 0$, the model is correctly specified, and the rejection frequencies are below nominal values across the broad. When $\delta > 0$, the model is mis-specified. Our test

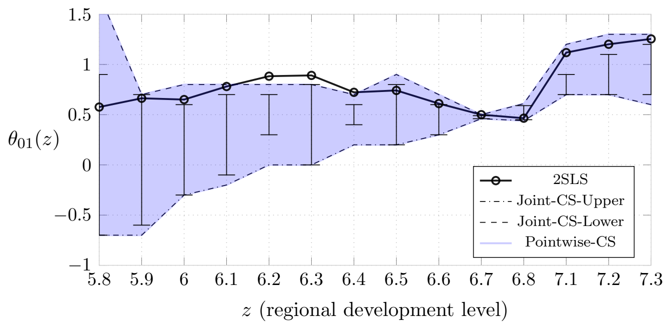


Figure 6: Confidence Intervals (95%) Without standardization

rejects the model with large frequencies and the rejection rate increases with sample size n and misspecification magnitude δ.

Table 3: Rejection Frequency: Entry Game

δ	n	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
$\delta = 0.0$	$n = 2000$	0.002	0.000	0.000
	$n = 4000$	0.005	0.000	0.000
	$n = 8000$	0.020	0.005	0.000
$\delta = 0.2$	$n = 2000$	0.597	0.145	0.000
	$n = 4000$	0.980	0.682	0.012
	$n = 8000$	1.000	1.000	0.407
$\delta = 0.4$	$n = 2000$	1.000	0.967	0.002
	$n = 4000$	1.000	1.000	0.402
	$n = 8000$	1.000	1.000	1.000

D.3 Additional Empirical Results

Figure 6 reports the joint and pointwise confidence set for the return to schooling without standardization. As we can see, the results are quite similar. Figure 7 reports the inference results with subsamples defined by gender and age.

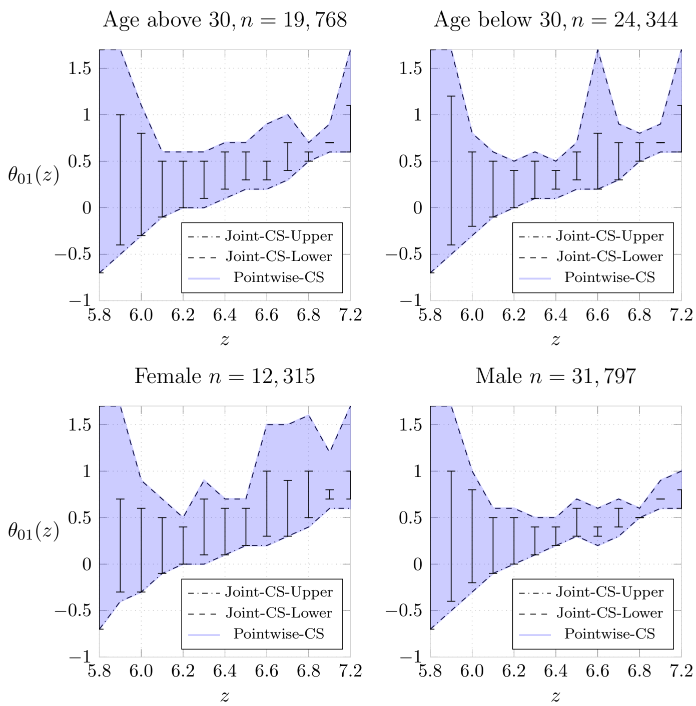


Figure 7: Confidence Intervals (95%) for Return to Education

E Additional Motivating Examples

This section lists some additional examples in which our method is potentially applicable.

Example E.1 (Quantile Regression with Interval-Outcome). *Consider a similar regression as Example 2.2 but under conditional quantile independence assumption:*

$$Y = X'\theta_0(Z) + \epsilon, \quad q_{\epsilon|X,Z}(\tau|X,Z) = 0, \quad a.s. - (X,Z) \tag{E.1}$$

where Y is a latent dependent variable and $q_{\epsilon|X,Z}(\tau|X,Z)$ denotes the τth conditional quantile of ϵ on X, Z. If Y were observed by researchers, it is the quantile varying coefficient model analyzed by Honda (2004). If Y is not directly observed but known to lie in the observed interval $[Y_\ell, Y_u]$, then the following moment inequalities hold for any $z \in \mathcal{Z}$:

$$E_P[\tau - 1\{Y_u \leq X'\theta_0(Z)\}|X, Z = z] \geq 0 \quad a.s.\ X \text{ and}$$
$$E_P[1\{Y_\ell \leq X'\theta_0(Z)\} - \tau|X, Z = z] \geq 0 \quad a.s.\ X.$$

Example E.2 (Quantile Regression with Censoring). *Consider again the quantile varying coefficient model in Equation (E.1). Suppose now Y is subject to censoring according to an observed binary variable $D \in \{0, 1\}$: Y is observed only when $D = 1$. Then, the following moment inequalities hold for any $z \in \mathcal{Z}$:*

$$E_P[\tau - 1\{Y \leq X'\theta_0(Z), D = 1\}|X, Z = z] \geq 0 \quad a.s.\ X \text{ and}$$
$$E_P[1\{Y \leq X'\theta_0(Z), D = 1\} + 1\{D = 0\} - \tau|X, Z = z] \geq 0 \quad a.s.\ X.$$

Example E.3 (Testing LATE Assumptions). *Consider a potential outcome model with binary treatment $D \in \{0,1\}$ and binary instrument $T \in \{0,1\}$. Let X_1 and X_0 be two potential outcomes, and D_0 and D_1 be two potential treatments. Let Z be a vector of covariates (here we name variables differently from the conventional treatment effect literature to match our notation). Suppose for any $z \in \mathcal{Z}$, we have (i) $(X_1, X_0, D_0, D_1) \perp T|Z = z$, (ii) $Pr(D = 1|T = 1, Z = z) \neq Pr(D = 1|T = 0, Z = z)$, and (iii) $D_1 \geq D_0$*

or $D_0 \geq D_1$ *a.s., then the conditional local average treatment effect* $E_P[X_1 - X_0|Z = z]$ *is identified by the Wald estimand. Mourifié and Wan (2017, Corollary 1) formulated the testable implication of LATE identifying assumptions (i)–(iii) as a set of conditional moment inequalities:*

$$
\begin{aligned}
&E_P[c_1(Z)D(1-T) - c_0(Z)DT|Z = z, X] \leq 0, \quad a.s.\ X \\
&E_P[c_0(Z)(1-D)T - c_1(Z)(1-D)(1-T)|Z = z, X] \leq 0, \quad a.s.\ X \\
&E_P[c_1(Z) - T|Z = z] = 0; \\
&E_P[c_0(Z) - (1-T)|Z = z] = 0.
\end{aligned}
$$

It fits the Model (2.1) with $\theta_0(Z) = (c_1(Z), c_0(Z))$ *be the varying coefficient, and* $W = (T, Z, D, X, Z')'$*. In this case, the random coefficients* $c_1(z)$ *and* $c_0(z)$ *are point-identified as the conditional probability* $Pr(T = 1|Z = z)$ *and* $Pr(T = 0|Z = z)$*, respectively. Researchers are interested in testing the model specification instead of estimation. Unlike Mourifié and Wan (2017)'s algorithm, we allow* Z *be either discrete or continuous.*[20]

[20] Mourifié and Wan (2017)'s implementation procedure is built upon the Stata package of Chernozhukov, Kim, Lee, and Rosen (2015) and accommodates only a single continuous conditioning variable. So a continuous Z needs to be discretized. Our method, on the other hand, allows for both discrete and continuous Z.

References

Ahmad, I., S. Leelahanon, and Q. Li (2005): "Efficient estimation of a semiparametric partially linear varying coefficient model," *The Annals of Statistics*, 33(1), 258–283.

Almunia, M., P. Antràs, D. Lopez-Rodriguez, and E. Morales (2021): "Venting out: Exports during a domestic slump," *American Economic Review*, 111(11), 3611–3662.

Andrews, D. W., S. Berry, and P. Jia (2004): "Confidence regions for parameters in discrete games with multiple equilibria, with an application to discount chain store location," *manuscript, Yale University*.

Andrews, D. W., and P. Guggenberger (2009): "Validity of subsampling and "plug-in asymptotic" inference for parameters defined by moment inequalities," *Econometric Theory*, 25(3), 669–709.

Andrews, D. W., and S. Kwon (2019): "Inference in moment inequality models that is robust to spurious precision under model misspecification," *Cowles Foundation Discussion Paper*.

Andrews, D. W., and X. Shi (2013): "Inference based on conditional moment inequalities," *Econometrica*, 81(2), 609–666.

——— (2014): "Nonparametric inference based on conditional moment inequalities," *Journal of Econometrics*, 179(1), 31–45.

——— (2017): "Inference based on many conditional moment inequalities," *Journal of econometrics*, 196(2), 275–287.

Andrews, D. W., and G. Soares (2010): "Inference for parameters defined by moment inequalities using generalized moment selection," *Econometrica*, 78(1), 119–157.

Ang, A., and J. Liu (2004): "How to discount cashflows with time-varying expected returns," *The Journal of Finance*, 59(6), 2745–2783.

Aradillas-López, A., and A. Gandhi (2016): "Estimation of games with ordered actions: An application to chain-store entry," *Quantitative Economics*, 7(3), 727–780.

Armstrong, T. B. (2014): "Weighted KS statistics for inference on conditional moment inequalities," *Journal of Econometrics*, 181(2), 92–116.

——— (2015): "Asymptotically exact inference in conditional moment inequality models," *Journal of Econometrics*, 186(1), 51–65.

——— (2018): "On the choice of test statistic for conditional moment inequalities," *Journal of Econometrics*, 203(2), 241–255.

Belloni, A., F. A. Bugni, and V. Chernozhukov (2019): "Subvector inference in PI models with many moment inequalities," Discussion paper.

Bontemps, C., and T. Magnac (2017): "Set identification, moment restrictions, and inference," *Annual Review of Economics*, 9, 103–129.

Bugni, F. A., I. A. Canay, and X. Shi (2015): "Specification tests for partially identified models defined by moment inequalities," *Journal of Econometrics*, 185(1), 259–282.

——— (2017): "Inference for subvectors and other functions of partially identified parameters in moment inequality models," *Quantitative Economics*, 8(1), 1–38.

Cai, Z. (2010): "Functional Coefficient Models for Economic and Financial Data," *Handbook of Functional Data Analysis, eds. F. Ferraty and Y. Romain*, pp. 166–186, Oxford University Press.

Cai, Z., L. Chen, and Y. Fang (2018): "A semiparametric quantile panel data model with an application to estimating the growth effect of FDI," *Journal of Econometrics*, 206(2), 531–553.

Cai, Z., M. Das, H. Xiong, and X. Wu (2006): "Functional coefficient instrumental variables models," *Journal of Econometrics*, 133(1), 207–241.

Cai, Z., J. Fan, and R. Li (2000): "Efficient estimation and inferences for varying-coefficient models," *Journal of the American Statistical Association*, 95(451), 888–902.

Cai, Z., Y. Fang, M. Lin, and J. Su (2019): "Inferences for a partially varying coefficient model with endogenous regressors," *Journal of Business & Economic Statistics*, 37(1), 158–170.

Cai, Z., and Y. Hong (2009): "Some recent developments in nonparametric finance," in *Nonparametric Econometric Methods*, pp. 379–432. Emerald Group Publishing Limited.

Cai, Z., Y. Ren, and B. Yang (2015): "A semiparametric conditional capital asset pricing model," *Journal of Banking & Finance*, 61, 117–126.

Cai, Z., and X. Xu (2008): "Nonparametric quantile estimations for dynamic smooth coefficient models," *Journal of the American Statistical Association*, 103(484), 1595–1608.

Canay, I. A., and A. M. Shaikh (2017): "Practical and theoretical advances in inference for partially identified models," *Advances in Economics and Econometrics*, 2, 271–306.

Card, D. (2001): "Estimating the return to schooling: Progress on some persistent econometric problems," *Econometrica*, 69(5), 1127–1160.

Chaney, T. (2018): "The gravity equation in international trade: An explanation," *Journal of Political Economy*, 126(1), 150–177.

Chen, R., and R. S. Tsay (1993): "Functional-coefficient autoregressive models," *Journal of the American Statistical Association*, 88(421), 298–308.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and confidence regions for parameter sets in econometric models 1," *Econometrica*, 75(5), 1243–1284.

Chernozhukov, V., W. Kim, S. Lee, and A. M. Rosen (2015): "Implementing intersection bounds in Stata," *The Stata Journal*, 15(1), 21–44.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection bounds: Estimation and inference," *Econometrica*, 81(2), 667–737.

Chernozhukov, V., W. K. Newey, and A. Santos (2023): "Constrained conditional moment restriction models," *Econometrica*, 91(2), 709–736.

Ciliberto, F., and E. Tamer (2009): "Market structure and multiple equilibria in airline markets," *Econometrica*, 77(6), 1791–1828.

Dingel, J. I. (2017): "The Determinants of Quality Specialization," *The Review of Economic Studies*, 84(4), 1551–1582.

Fan, J., and R. Li (2004): "New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis," *Journal of the American Statistical Association*, 99(467), 710–723.

Fan, J., and J.-T. Zhang (2000): "Two-step estimation of functional linear models with applications to longitudinal data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2), 303–322.

Fan, J., and W. Zhang (1999): "Statistical estimation in varying coefficient models," *The annals of Statistics*, 27(5), 1491–1518.

Hastie, T., and R. Tibshirani (1993): "Varying-coefficient models," *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4), 757–779.

Heckman, J. J. (2005): "China's human capital investment," *China Economic Review*, 16(1), 50–70.

Heckman, J. J., and X. Li (2004): "Selection bias, comparative advantage and heterogeneous returns to education: evidence from China in 2000," *Pacific Economic Review*, 9(3), 155–171.

Honda, T. (2004): "Quantile regression in varying coefficient models," *Journal of Statistical Planning and Inference*, 121(1), 113–125.

Hong, S. (2017): "Inference in semiparametric conditional moment models with partial identification," *Journal of econometrics*, 196(1), 156–179.

Hsu, Y.-C. (2016): "Multiplier bootstrap for empirical processes," Discussion paper, Institute of Economics, Academia Sinica, Taipei, Taiwan.

Hsu, Y.-C., and X. Shi (2017): "Model-selection tests for conditional moment restriction models," *The Econometrics Journal*, 20(1), 52–85.

Huber, M., and G. Mellace (2015): "Testing instrument validity for LATE identification based on inequality moment constraints," *Review of Economics and Statistics*, 97(2), 398–411.

Imbens, G. W., and T. Lemieux (2008): "Regression discontinuity designs: A guide to practice," *Journal of econometrics*, 142(2), 615–635.

Imbens, G. W., and C. F. Manski (2004): "Confidence intervals for partially identified parameters," *Econometrica*, 72(6), 1845–1857.

Kaido, H. (2017): "Asymptotically efficient estimation of weighted average derivatives with an interval censored variable," *Econometric Theory*, 33(5), 1218–1241.

Kaido, H., F. Molinari, and J. Stoye (2019): "Confidence intervals for projections of partially identified parameters," *Econometrica*, 87(4), 1397–1432.

Kédagni, D., and I. Mourifié (2020): "Generalized instrumental inequalities: testing the instrumental variable independence assumption," *Biometrika*, 107(3), 661–675.

Kim, K. i. (2008): "Set estimation and inference with models characterized by conditional moment inequalities," *Working Paper*.

Kitagawa, T. (2015): "A Test for Instrument Validity," *Econometrica*, 83(5), 2043–2063.

Lee, S., K. Song, and Y.-J. Whang (2013): "Testing functional inequalities," *Journal of Econometrics*, 172(1), 14–32.

Li, Q., C. J. Huang, D. Li, and T.-T. Fu (2002): "Semiparametric smooth coefficient models," *Journal of Business & Economic Statistics*, 20(3), 412–422.

Liu, S., I. Mourifié, and Y. Wan (2020): "Two-way exclusion restrictions in models with heterogeneous treatment effects," *The Econometrics Journal*, 23(3), 345–362.

Manski, C. F., and E. Tamer (2002): "Inference on regressions with interval data on a regressor or outcome," *Econometrica*, 70(2), 519–546.

Marcoux, M., T. M. Russell, and Y. Wan (2024): "A simple specification test for models with many conditional moment inequalities," *Journal of Econometrics*, 242(1), 105788.

Mayer, T., and S. Zignago (2011): "Notes on CEPII's distances measures: The GeoDist database," Working Papers 2011-25, CEPII research center.

Menzel, K. (2014): "Consistent estimation with many moment inequalities," *Journal of Econometrics*, 182(2), 329–350.

Morales, E., G. Sheu, and A. Zahler (2019): "Extended gravity," *The Review of economic studies*, 86(6), 2668–2712.

Mourifié, I., and Y. Wan (2017): "Testing local average treatment effect assumptions," *Review of Economics and Statistics*, 99(2), 305–313.

Nevo, A., and A. M. Rosen (2012): "Identification with imperfect instruments," *Review of Economics and Statistics*, 94(3), 659–671.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2015): "Moment inequalities and their application," *Econometrica*, 83(1), 315–334.

Romano, J. P., and A. M. Shaikh (2008): "Inference for identifiable parameters in partially identified econometric models," *Journal of Statistical Planning and Inference*, 138(9), 2786–2807.

——— (2010): "Inference for the identified set in partially identified econometric models," *Econometrica*, 78(1), 169–211.

Santos, A. (2012): "Inference in nonparametric instrumental variables with partial identification," *Econometrica*, 80(1), 213–275.

Schultz, T. P. (2003): "Human capital, schooling and health," *Economics & Human Biology*, 1(2), 207–221.

Su, L., I. Murtazashvili, and A. Ullah (2013): "Local linear GMM estimation of functional coefficient IV models with an application to estimating the rate of return to schooling," *Journal of Business & Economic Statistics*, 31(2), 184–207.

Sun, Z. (2023): "Instrument validity for heterogeneous causal effects," *Journal of Econometrics*, 237(2), 105523.

Tao, J. (2015): "On the Asymptotic Theory for Semiparametric GMM Models with Partial Identification," Discussion paper.

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

Wan, Y. (2013): "An integration-based approach to moment inequality models," *Manuscript. University of Toronto.*