Inference Based on Many Conditional Moment Inequalities

Donald W. K. Andrews
Cowles Foundation
Yale University

Xiaoxia Shi
Department of Economics
University of Wisconsin-Madison

June 2010; revised: April 26, 2016*

Abstract

In this paper, we construct confidence sets for models defined by many conditional moment inequalities/equalities. The conditional moment restrictions in the models can be finite, countably infinite, or uncountably infinite. To deal with the complication brought about by the vast number of moment restrictions, we exploit the manageability (Pollard (1990)) of the class of moment functions. We verify the manageability condition in five examples from the recent partial identification literature.

The proposed confidence sets are shown to have correct asymptotic size in a uniform sense and to exclude parameter values outside the identified set with probability approaching one. Monte Carlo experiments for a conditional stochastic dominance example and a random-coefficients binary-outcome example support the theoretical results.

*Andrews gratefully acknowledges the research support of the National Science Foundation via grant numbers SES-1058376 and SES-1355504. Shi gratefully acknowledges the research support of the University of Wisconsin-Madison Graduate School with funding from the Wisconsin Alumni Research Foundation. The first version of the results of this paper were given in Section 9 of Andrews and Shi (2009).

1 Introduction

In this paper, we extend the results in Andrews and Shi (2013a, b) (AS1, AS2) to cover models defined by many conditional moment inequalities and/or equalities ("MCMI" in short). The number of conditional moment inequalities/equalities can be countable or uncountable. Examples of models covered by the results include (1) conditional stochastic dominance, (2) random-coefficients binary-outcome models with instrumental variables, see Chesher and Rosen (2014), (3) convex moment prediction models, see Beresteanu, Molchanov, and Molinari (2010), (4) ordered-choice models with endogeneity and instruments, see Chesher and Smolinski (2012), and (5) discrete games identified by revealed preference, see Pakes, Porter, Ho, and Ishii (2015).

The main feature of an MCMI model is that the number of moment restrictions implied by the model is doubly "many." First, there are many (countable or uncountable) conditional moment restrictions, and second each conditional moment restriction implies infinitely many moment conditions. As in AS1 and AS2, we transform each conditional moment restriction into infinitely many unconditional ones using instrumental functions. After the transformation, the unconditional moment functions of the model form a class that is indexed by both the instrumental functions and the indices of the conditional moment restrictions. We exploit a manageability assumption on the class of conditional moment functions. With this assumption, we show that the class of transformed unconditional moment inequalities/equalities is also manageable and, in consequence, can be treated similarly to those to AS1 and AS2.

Thus, the manageability assumption on the class of conditional moment functions is crucial for our theoretical framework. This assumption is verified in the examples by deriving upper bounds on the covering numbers of the functional classes that arise. The upper bounds in the first two examples are derived by bounding the pseudo-dimensions of the functional classes. In the third example, they are derived using the Lipschitz continuity of the moment functions with respect to the index. These three examples are representative of cases where there are a continuum of conditional moment inequality/equalities. In the fourth and the fifth examples, the numbers of conditional moment inequalities/equalities are countable. For countable functional classes, we treat their elements as sequences and impose decreasing weights on them. The weights guarantee an appropriate bound for the covering numbers.

We note that the approach in this paper also is applicable to models defined by many unconditional moment inequalities/equalities. For such models, one simply omits the step that transforms the conditional moments restrictions into unconditional ones using instrumental functions.

This paper belongs to the moment inequality literature, which is now quite large. The most closely related paper is Chernozhukov, Chetverikov, and Kato (2014a), which studies models defined by many moment inequalities. They construct two types of tests, one based on a fixed critical value derived using a moderate deviation inequality, and the other based on a bootstrap critical value derived using distributional approximation theory for suprema of empirical processes developed in Chernozhukov, Chetverikov, and Kato (2013, 2014b). Both are based on a supremum-type test statistic, which is similar to, but different from, the KS statistic considered here. In one Monte Carlo example considered here, the one-step and two-step versions of their tests do not perform as well as the MCMI tests proposed in this paper. In the other Monte Carlo example considered here, their two-step bootstrap-based method performs better than the MCMI methods proposed in this paper at a large sample size, but not as well at smaller sample sizes.

Like this paper, Delgado and Escanciano (2013) consider tests for conditional stochastic dominance. They take a different approach from the approach in this paper.

Papers in the literature that consider conditional moment inequalities, but not MCMI, include Khan and Tamer (2009), Chetverikov (2012), AS1, Armstrong and Chan (2013), Chernozhukov, Lee, and Rosen (2013), Gandhi, Lu, and Shi (2013), Lee, Song, and Whang (2013), Andrews and Shi (2014), and Armstrong (2014a,b, 2015). Galichon and Henry (2009) provides related results. Papers in the literature that test a continuum of unconditional moment inequalities include papers on testing stochastic dominance and stochastic monotonicity, see Linton, Song, and Whang (2010) and references therein. Papers in the literature that test a continuum of inequalities that are not moment inequalities and, hence, to which the tests in this paper do not apply, include tests of Lorenz dominance, see Dardanoni and Forcina (1999) and Barrett, Donald, and Bhattacharya (2014), and tests of likelihood ratio (or density) ordering, see Beare and Moon (2015), Beare and Shi (2015), and references therein.

The rest of the paper is organized as follows. Section 2 specifies the model and describes the examples. Section 3 introduces the MCMI test statistics and confidence sets. Section 4 defines the critical values and gives a step-by-step guide for implementation. Section 5 shows the uniform asymptotic size of the proposed tests and confidence sets in the general setup. Section 6 gives the power results. Sections 7-9 verify the conditions imposed in Sections 5 and 6 for each of the examples. Sections 7 and 8 also provide finite-sample Monte Carlo results for the problem of testing conditional stochastic dominance and for the random-coefficients binary-outcome model with instruments. Section 10 concludes. An Appendix available on-line provides proofs and some additional simulation results.

For notational simplicity, throughout the paper, we let $(a_i)_{i=1}^n$ denote the n-vector $(a_1, ..., a_n)'$ for $a_i \in R$. We let $A := B$ denote that A equals B by definition or assumption.

2 Many Conditional Moment Inequalities/Equalities

2.1 Models

The models considered in this paper are of the following general form:

$$\begin{aligned}&E_{F_0}[m_j(W_i, \theta_0, \tau)|X_i] \geq 0 \text{ a.s. for } j = 1, ..., p \text{ and}\\&E_{F_0}[m_j(W_i, \theta_0, \tau)|X_i] = 0 \text{ a.s. for } j = p+1, ..., p+v, \ \forall \tau \in \mathcal{T},\end{aligned} \quad (2.1)$$

where $\mathcal{T}$ is a set of indices that may contain an infinite number of elements, θ_0 is the unknown true parameter value that belongs to a parameter space $\Theta \subset R^{d_\theta}$, the observations $\{W_i : i \leq n\}$ are i.i.d., F_0 is the unknown true distribution of W_i, X_i is a sub-vector of W_i, and $m(w, \theta, \tau) := (m_1(w, \theta, \tau), ..., m_{p+v}(w, \theta, \tau))'$ is a vector of known moment functions.[1]

In contrast, the parameter $\tau \in \mathcal{T}$ does not appear in the moment inequality/equality models considered in AS1 and AS2.

The object of interest is θ_0, which is not assumed to be point identified. The model restricts θ_0 to the *identified set* (which could be a singleton), which is defined by

$$\Theta_{F_0} := \{\theta \in \Theta : (2.1) \text{ holds with } \theta \text{ in place of } \theta_0\}. \quad (2.2)$$

We are interested in confidence sets (CS's) that cover the true value θ_0 with probability greater than or equal to $1 - \alpha$ for $\alpha \in (0, 1)$. We construct such CS's by inverting tests of

[1] The requirement that X_i is a sub-vector of W_i does not preclude X_i from containing excluded instruments because $m(W_i, \theta_0, \tau)$ is not required to vary with every element of W_i.

the null hypothesis that θ is the true value for each $\theta \in \Theta$. Let $T_n(\theta)$ be a test statistic and $c_{n,1-\alpha}(\theta)$ be a corresponding critical value for a test with nominal significance level α. Then, a nominal level $1-\alpha$ CS for the true value θ_0 is

$$CS_n := \{\theta \in \Theta : T_n(\theta) \leq c_{n,1-\alpha}(\theta)\}. \tag{2.3}$$

At each $\theta \in \Theta$, we test the validity of the moment conditions in (2.1) with θ_0 replaced with θ. The tests are of interest in their own right when (i) there is no parameter to estimate in the moment conditions, as in Example 1 below, or (ii) the validity of the moment conditions at a given θ has policy implications.

2.2 Examples

Models of the form described in (2.1) arise in many empirically relevant situations. Below are some examples.

Example 1 (Conditional Stochastic Dominance). Let $W := (Y_1, Y_2, X)$. Some economic theories imply that the distribution of Y_1 stochastically dominates that of Y_2 conditional on X. For an integer $s \geq 1$, the sth-order conditional stochastic dominance of Y_1 over Y_2 can be written as conditional moment inequalities:

$$\begin{aligned} &E_{F_0}[G_s(Y_2,\tau) - G_s(Y_1,\tau)|X] \geq 0 \text{ a.s. } \forall \tau \in \mathcal{T}, \text{ where} \\ &G_s(y,\tau) := (\tau - y)^{s-1} 1\{y \leq \tau\} \end{aligned} \tag{2.4}$$

and $\mathcal{T}$ contains the supports of Y_1 and Y_2. The tests developed below are directly applicable in this example without being inverted into a confidence set.

Stochastic dominance relationships have been used in income and welfare analysis, for example, in Anderson (1996, 2004), Davidson and Duclos (2000), and Bishop, Zeager, and Zheng (2011). Stochastic dominance relationships also have been used in the study of auctions, e.g., in Guerre, Perrigne, and Vuong (2009). Conditional stochastic dominance implies that the relationship holds for every subgroup of the population defined by X and is useful in all of these applications. See Delgado and Escanciano (2013) for a different approach to testing conditional stochastic dominance from the one considered here.

Sometimes, one may be interested in the conditional stochastic dominance relationship among multiple distributions. For example, for $W = (Y_1, Y_2, Y_3, X)$, one would like to know whether Y_1 s-th order stochastically dominates Y_2 and Y_2 over Y_3 conditional on X. The corresponding conditional moment inequalities to be tested are as follows,

$$E_{F_0}[G_s(Y_2, \tau) - G_s(Y_1, \tau)|X] \geq 0 \text{ and}$$
$$E_{F_0}[G_s(Y_3, \tau) - G_s(Y_2, \tau)|X] \geq 0 \text{ a.s. } \forall \tau \in \mathcal{T}, \tag{2.5}$$

where $\mathcal{T}$ contains the supports of Y_1, Y_2, and Y_3. For example, the comparison of multiple distributions has been considered in Dardanoni and Forcina (1999) for Lorenz dominance.

Example 2 (Random-Coefficients Binary-Outcome Models with Instrumental Variables). Consider the random-coefficients binary-outcome model with instrumental variables (IV's) studied in Chesher and Rosen (2014) (CR):

$$Y_1 = 1\{\beta_0 + X_1'\beta_1 + Y_2'\beta_2 \geq 0\}, \tag{2.6}$$

where $\beta := (\beta_0, \beta_1', \beta_2')'$ are random coefficients that belong to the space R^{d_β}. The covariate vector X_1 is assumed to be exogenous (i.e., independent of β), while the covariate vector Y_2 may be endogenous. Let X_2 be a vector of instrumental variables that is independent of β. Suppose the parameter of interest is the marginal distribution of β, denoted by F_β. Theorem 1 of CR implies that under their Assumptions A1-A3, the sharp identified set for F_β is defined by the following moment inequalities:

$$E_{F_0}[F_\beta(\mathcal{S}) - 1\{S(Y_1, Y_2, X_1) \subset \mathcal{S}\}|X_1, X_2] \geq 0 \text{ a.s. } \forall \mathcal{S} \in \mathbf{S}, \tag{2.7}$$

where

$$S(y_1, y_2, x_1) := cl\{b = (b_0, b_1', b_2')' \in R^{d_\beta} : y_1 = 1\{b_0 + x_1'b_1 + y_2'b_2 \geq 0\}\},$$
$$\mathbf{S} := \{cl(\cup_{c \in \mathcal{C}} H(c)) : \mathcal{C} \subset R^{d_\beta}\},$$
$$H(c) := \{b \in R^{d_\beta} : b'c \geq 0\} \text{ for } c \in R^{d_\beta}, \tag{2.8}$$

cl denotes "closure," and $H(c)$ is the half-space orthogonal to $c \in R^{d_\beta}$.

Often one may wish to parameterize F_β by assuming $F_\beta(\cdot) = F_\beta(\cdot; \theta)$ for a known distribution function $F_\beta(\cdot; \cdot)$ and an unknown finite-dimensional parameter $\theta \in \Theta$. Then, the sharp identified set for θ is defined by the moment inequalities:

$$E_{F_0}[F_\beta(\mathcal{S},\theta) - 1\{S(Y_1,Y_2,X_1) \subset \mathcal{S}\}|X_1,X_2] \geq 0 \text{ a.s. } \forall \mathcal{S} \in \mathbf{S}. \tag{2.9}$$

This fits into the framework of (2.1) with $W = (Y_1, Y_2', X_1', X_2')'$, $X = (X_1', X_2')'$, $\tau = \mathcal{S}$, $\mathcal{T} = \mathbf{S}$, $p = 1$, $v = 0$, and $m(w,\theta,\tau) = F_\beta(\mathcal{S},\theta) - 1\{S(y_1,y_2,x_1) \subset \mathcal{S}\}$.

Example 3 (Convex Moment Prediction Models–Support Function Approach). Beresteanu, Molchanov, and Molinari (2010) (BMM) establish a framework to characterize the sharp identified set for a general class of incomplete models with convex moment predictions using random set theory. Examples of such models include static, simultaneous move, finite games with complete or incomplete information in the presence of multiple equilibria, best linear prediction models with interval outcome and/or regressor data, and random utility models of multinomial choice with interval regressor data. BMM show that the sharp identified set for these models can be characterized by a continuum of conditional moment inequalities using the support function of the set. For parameter inference, BMM suggest applying the procedure in this paper and they verify the high-level assumptions in an earlier version of this paper in two examples. Here, we describe their identification framework briefly.

Consider a model based on an observed random vector W and an unobserved random vector V. The model maps each value of (W,V) to a closed set $Q_\theta(W,V) \subseteq R^d$, where θ is the model parameter that belongs to a parameter space Θ, and d is a positive integer. Let X be a sub-vector of W with support contained in $\mathcal{X}$ and let $q(x) : \mathcal{X} \to R^d$ be a known function. Suppose (W,V) and W take values in some sets $\mathcal{WV}$ and $\mathcal{W}$, respectively. BMM assume that the sharp identified set of θ implied by the model is

$$\Theta_I = \{\theta \in \Theta : q(X) \in \mathbb{E}_{F_0}[Q_\theta(W,V)|X] \ a.s. \ [X]\}, \tag{2.10}$$

where $\mathbb{E}_{F_0}[\cdot]$ stands for the Aumann expectation of the random set inside the square brackets under the true distribution F_0 of (W,V). BMM show that the event $q(X) \in \mathbb{E}_{F_0}[Q_\theta(W,V)|X]$ can be written equivalently as the following set of moment inequalities

$$E_{F_0}[h(Q_\theta(W,V),u) - u'q(X)|X] \geq 0 \ a.s.[X], \ \forall u \in R^d \ s.t. \ \|u\| \leq 1, \tag{2.11}$$

where $h(Q,u)$ is the support function of Q in the direction given by u, that is, $h(Q,u) = \sup_{q\in Q} q'u$.

The inequalities (2.11) do not fall immediately into our general framework because of the unobservable V. However, in applications, one typically has that either $Q_\theta(W,V) = Q_\theta(W)$ (so that V does not appear in (2.11)) or the distribution of V given X (denoted $F_{V|X}(v|x;\theta)$) is known to the researcher up to an unknown parameter θ. In the former case, (2.11) fits the form of (2.1). In the latter case, we write (2.11) as

$$E_{F_0}\left[\int h(Q_\theta(W,v),u)dF_{V|X}(v|X;\theta) - u'q(X)|X\right] \geq 0 \ a.s.\ [X],\ \forall u \in R^d \ s.t.\ \|u\| \leq 1, \tag{2.12}$$

which fits the form of (2.1). The former case includes the best linear predictor example in BMM, and the latter case includes the entry game example in BMM.

Example 4 (IV Ordered-Choice Models). Chesher and Smolinski (2012) show that the sharp identified set for a nonparametric single equation instrumental variable (SEIV) model with ordered outcome and discrete endogenous regressors can be characterized by a finite, but potentially very large, number of moment inequalities. Consider the non-separable model

$$Y = h(Z,U), \tag{2.13}$$

where $Y \in \{1,2,...,M\}$ and $Z \in \{z_1,...,z_K\}$, the error term U is normalized to be uniformly distributed in $[0,1]$. Assume that there is a vector of instrumental variables X that is independent of U. Then, one has a SEIV model. Further, assume that h is weakly increasing in U. Then, h has a threshold crossing representation: for $m = 1,...,M$ and $z \in \{z_1,...,z_K\}$:

$$h(z,u) = m \text{ if } u \in (h_{m-1}(z), h_m(z)] \tag{2.14}$$

for some constants $0 = h_0(z) < \cdots < h_M(z) = 1$. Thus, estimating $h(z,u)$ amounts to estimating the $J = (M-1)K$ threshold parameters $\gamma = (\gamma_{11},...,\gamma_{(M-1)1},...,\gamma_{1K},...,\gamma_{(M-1)K})'$, where

$$\gamma_{mk} = h_m(z_k)\ \forall m = 1,...,M-1,\ \forall k = 1,...,K. \tag{2.15}$$

Chesher and Smolinski (2012) show that the sharp identified set for γ can be characterized by the following moment inequalities

$$E_{F_0}\left[\gamma_{\ell s}-\sum_{k=1}^{K}\sum_{m=1}^{M-1}1\{Y=m,Z=z_k,\gamma_{mk}\leq\gamma_{\ell s}\}\middle|X\right]\geq 0\text{ a.s. }[X]\text{ and}$$
$$E_{F_0}\left[\sum_{k=1}^{K}\sum_{m=1}^{M-1}1\{Y=m,Z=z_k,\gamma_{(m-1)k}<\gamma_{\ell s}\}-\gamma_{\ell s}\middle|X\right]\geq 0\text{ a.s. }[X]\ \forall\ell\leq M-1,\ \forall s\leq K,$$
$$E_{F_0}\left[\gamma_{\ell s}-\gamma_{ms}-1\{m<Y\leq\ell,Z=z_s\}|\,X\right]\geq 0\text{ a.s. }[X]$$
$$\forall\ell>m,\ \forall\ell,m\leq M-1,\ \forall s\leq K. \tag{2.16}$$

We arrange the above $N:=2(M-1)K+(M-2)(M-1)K/2$ inequalities into a column, and index them by τ for $\tau=1,\ldots,N$. Let $W=(Y,X,Z')'$ and let $m(W,\gamma,\tau)$ be the expression inside the conditional expectation in the τth inequality. Then, this example falls into the framework of (2.1) with $\theta=\gamma$.

One may wish to parameterize the threshold functions γ via $\gamma=\Gamma(\theta)$. In that case, the same set of moment inequalities as above defines the sharp identified set for θ. For example, Chesher and Smolinski (2012) show that, for the linear ordered-probit model,

$$\gamma_{mk}:=h_m(z_k)=\Phi(c_m-a_1z_k)\ \forall m=1,\ldots,M-1,\ \forall k=1,\ldots,K, \tag{2.17}$$

where $c_1,\ldots,c_{M-1}$ are the threshold values, a_1 is the slope parameter, and $\Phi(\cdot)$ is the standard normal distribution function.

Example 5 (Revealed Preference Approach in Discrete Games). Pakes, Porter, Ho, and Ishii (2015) formalize the idea of using the revealed preference principle to estimate games in which a finite number of players have a discrete set of actions to choose from. Observing the players' equilibrium play, the econometrician can write down moment inequalities that are implied by the revealed preference principle. These moment inequalities allow one to estimate the structural parameters without solving for the equilibrium. Here we describe a simplified version of their framework.

Suppose that all players make decisions based on the same information set and the econometrician observes the information set. Players make decisions based on expected utility maximization. Suppose there are J players and each player has a feasible action set A_j that is discrete (i.e., finite or countably infinite). Let $\pi_j(a_j,a_{-j},Z;\theta)$ be the utility of player j given her own action a_j and her opponents' actions a_{-j} and the covariates Z. Let X be a sub-vector of Z that generates the information set of the players. Let the boldfaced $\mathbf{a}_j$ and $\mathbf{a}_{-j}$ be the observed actions of player j and her opponents. The function π_j is known up to the finite dimensional parameter θ. Then, the moment inequalities are

$$E_{F_0}\left(\pi_j\left(\mathbf{a}_j, \mathbf{a}_{-j}, Z; \theta\right) - \pi_j\left(a_j', \mathbf{a}_{-j}, Z; \theta\right) \middle| X\right) \geq 0 \; \forall a_j' \in A_j, \; \forall j = 1, ..., J. \tag{2.18}$$

When J is large or the number of elements in A_j is large, there are many (possibly countably infinitely many) conditional moment inequalities.

2.3 Parameter Space

Let (θ, F) denote a generic value of the parameter and the distribution of W_i. Let $\mathcal{F}$ denote the parameter space for the true values (θ_0, F_0), which satisfy the conditional moment inequalities and equalities. To specify $\mathcal{F}$, we first introduce some additional notation. For each distribution F, let F_X denote the marginal distribution of X_i under F. Let $k := p + v$.

Below, we employ a "manageability" condition that regulates the complexity of $\mathcal{T}$. It ensures a functional central limit theorem (CLT) holds, which is used in the proof of the uniform coverage probability results for the CS's. The concept of manageability is from Pollard (1990) and is defined in Section B.3 of the Appendix. This concept also is used in AS1 to regulate the complexity of the set of instrumental functions. The manageability condition could be replaced by some other condition from the literature that is sufficient for a functional central limit theorem to hold for the appropriate quantities.

The test consistency results given below apply to (θ, F) pairs that do not satisfy the conditional moment inequalities and equalities. For this reason, we introduce a set $\mathcal{F}_+$ that is a superset of $\mathcal{F}$ and does not impose the inequalities and equalities. Let $\mathcal{F}_+$ be some collection of (θ, F) that satisfy the following parameter space (PS) Assumptions PS1 and PS2 for given constants $\delta > 0$ and $C_1 < \infty$ and given deterministic function of (θ, F): $\sigma_F(\theta) := (\sigma_{F,1}(\theta), \ldots, \sigma_{F,k}(\theta))'$. The function $\sigma_F(\theta)$ is useful for the standardization of certain forms of the test statistic, and is specified in greater detail in sections below.

Assumption PS1. For any $(\theta, F) \in \mathcal{F}_+$,

(a) $\theta \in \Theta$,

(b) $\{W_i : i \geq 1\}$ are i.i.d. under F,

(c) $\sigma_{F,j}(\theta) > 0$, $\forall j = 1, \ldots, k$,

(d) $|m_j(w, \theta, \tau)/\sigma_{F,j}(\theta)| \leq M(w)$, $\forall w \in R^{d_w}$, $\forall j = 1, ..., k$, $\forall \tau \in \mathcal{T}$, for some function $M : R^{d_w} \rightarrow [0, \infty)$, and

(e) $E_F M^{2+\delta}(W_i) \leq C_1$.

Assumption PS2. For all sequences $\{(\theta_n, F_n) \in \mathcal{F}_+ : n \geq 1\}$, the triangular array of processes $\{(m_j(W_{n,i}, \theta_n, \tau)/\sigma_{F_n,j}(\theta_n))_{j=1}^k : \tau \in \mathcal{T}, i \leq n, n \geq 1\}$ is manageable with respect to the envelopes $\{M(W_{n,i}) : i \leq n, n \geq 1\}$, where $\{W_{n,i} : i \leq n, n \geq 1\}$ is a row-wise i.i.d. triangular array with $W_{n,i} \sim F_n$ $\forall i \leq n, n \geq 1$.

The parameter space $\mathcal{F}$ for the conditional moment inequality model is the subset of $\mathcal{F}_+$ that satisfies:

Assumption PS3. (a) $E_F[m_j(W_i, \theta, \tau)|X_i] \geq 0$ a.s. $[F_X]$ for $j = 1, ..., p$, $\forall \tau \in \mathcal{T}$,

(b) $E_F[m_j(W_i, \theta, \tau)|X_i] = 0$ a.s. $[F_X]$ for $j = p + 1, ..., k$, $\forall \tau \in \mathcal{T}$.

3 Tests and Confidence Sets

In this section, we describe the MCMI test statistics. To do so, we first transform the conditional moment inequalities/equalities into equivalent unconditional ones using instrumental functions. The unconditional moment conditions are as follows:

$$\begin{aligned} &E_{F_0}[m_j(W_i, \theta_0, \tau) g_j(X_i)] \geq 0 \text{ for } j = 1, ..., p \text{ and} \\ &E_{F_0}[m_j(W_i, \theta_0, \tau) g_j(X_i)] = 0 \text{ for } j = p + 1, ..., k, \\ &\qquad \forall \tau \in \mathcal{T} \text{ and } \forall g = (g_1, ..., g_k)' \in \mathcal{G}_{\text{c-cube}}, \end{aligned} \tag{3.1}$$

where g is a vector of instruments that depends on X_i and $\mathcal{G}_{\text{c-cube}}$ is a collection of instrumental functions g defined below.

We construct MCMI test statistics based on (3.1). Let the sample moment functions be

$$\begin{aligned} \overline{m}_n(\theta, \tau, g) &:= n^{-1} \sum_{i=1}^n m(W_i, \theta, \tau, g) \text{ for } g \in \mathcal{G}_{\text{c-cube}} \text{ and} \\ m(W_i, \theta, \tau, g) &:= (m_1(W_i, \theta, \tau) g_1(X_i), ..., m_k(W_i, \theta, \tau) g_k(X_i))' . \end{aligned} \tag{3.2}$$

The sample variance matrix of $n^{1/2}\overline{m}_n(\theta, g, \tau)$ is useful for most versions of the test statistic and for the critical values. It is defined as

$$\widehat{\Sigma}_n(\theta,\tau,g) := n^{-1}\sum_{i=1}^{n}(m(W_i,\theta,\tau,g)-\overline{m}_n(\theta,\tau,g))(m(W_i,\theta,\tau,g)-\overline{m}_n(\theta,\tau,g))'. \qquad (3.3)$$

When the sample variance is used, we would like it to be nonsingular because it is used to Studentize the sample moment functions. However, the matrix $\widehat{\Sigma}_n(\theta,\tau,g)$ may be singular or nearly singular with non-negligible probability for some (τ,g). Thus, we add a small positive definite matrix to $\widehat{\Sigma}_n(\theta,\tau,g)$:

$$\overline{\Sigma}_n(\theta,\tau,g) := \widehat{\Sigma}_n(\theta,\tau,g)+\varepsilon\cdot \mathrm{Diag}(\widehat{\sigma}^2_{n,1}(\theta),\ldots,\widehat{\sigma}^2_{n,k}(\theta)) \text{ for } (\tau,g)\in\mathcal{T}\times\mathcal{G}_{\text{c-cube}} \text{ and } \varepsilon=1/20, \qquad (3.4)$$

where $\widehat{\sigma}_{n,j}(\theta)$ is a consistent estimator of the $\sigma_{F,j}(\theta)$ introduced just above Assumption PS1.

In practice, if the moment functions have a natural scale (say, being a probability or the difference of two probabilities), one can take $\widehat{\sigma}_{n,j}(\theta)=\sigma_{F,j}(\theta)=1$ for all j, (θ,F), and n. Otherwise, we recommend taking $\widehat{\sigma}_{n,j}(\theta)$ and $\sigma_{F,j}(\theta)$ such that $\widehat{\sigma}^{-1}_{n,j}(\theta)m_j(W_i,\theta,\tau)$ and $\sigma^{-1}_{F,j}(\theta)m_j(W_i,\theta,\tau)$ are invariant to the rescaling of the moment functions, because this yields a test with the same property. We discuss specific choices for the examples in later sections.

We assume that the estimators $\{\widehat{\sigma}_{n,j}(\theta): j\le k\}$ satisfy the following uniform consistency condition.

Assumption SIG1. For all $\zeta>0$, $\sup_{(\theta,F)\in\mathcal{F}}\Pr\left(\max_{j\le k}|\widehat{\sigma}^2_{n,j}(\theta)/\sigma^2_{F,j}(\theta)-1|>\zeta\right)\to 0$.

The functions g that we consider are hypercubes in $[0,1]^{d_X}$. Hence, we transform each element of X_i to lie in $[0,1]$. (There is no loss in information in doing so.) For notational convenience, we suppose $X_i^\dagger\in R^{d_X}$ denotes the non-transformed IV vector and we let X_i denote the transformed IV vector. We transform $X_i^\dagger$ via a shift and rotation and then apply the standard normal distribution function $\Phi(x)$. Specifically, let

$$\begin{aligned}&X_i := \Phi(\widehat{\Sigma}^{-1/2}_{X,n}(X_i^\dagger-\overline{X}_n^\dagger)),\ \text{where } \Phi(x):=(\Phi(x_1),...,\Phi(x_{d_X}))' \text{ for } x=(x_1,...,x_{d_X})',\\ &\widehat{\Sigma}_{X,n}:=n^{-1}\Sigma_{i=1}^n(X_i^\dagger-\overline{X}_n^\dagger)(X_i^\dagger-\overline{X}_n^\dagger)',\ \text{and } \overline{X}_n^\dagger:=n^{-1}\Sigma_{i=1}^nX_i^\dagger.\end{aligned} \qquad (3.5)$$

We consider the class of indicator functions of cubes with side lengths that are $(2r)^{-1}$ for all large positive integers r. The cubes partition $[0,1]^{d_x}$ for each r. This class is countable:

$$\mathcal{G}_{\text{c-cube}} := \{g_{a,r} : g_{a,r}(x) := 1\{x \in C_{a,r}\} \cdot 1_k \text{ for } C_{a,r} \in \mathcal{C}_{\text{c-cube}}\}, \text{ where}$$
$$\mathcal{C}_{\text{c-cube}} := \left\{ C_{ar} := \prod_{u=1}^{d_x} ((a_u - 1)/(2r), a_u/(2r)] \in [0,1]^{d_x} : a = (a_1, ..., a_{d_x})' \right.$$
$$\left. a_u \in \{1, 2, ..., 2r\} \text{ for } u = 1, ..., d_x \text{ and } r = r_0, r_0 + 1, ... \right\} \quad (3.6)$$

for some positive integer r_0 and $1_k := (1, ..., 1)' \in R^k$.[2] The terminology "c-cube" abbreviates countable cubes. Note that $C_{a,r}$ is a hypercube in $[0,1]^{d_x}$ with smallest vertex indexed by a and side lengths equal to $(2r)^{-1}$.

The MCMI test statistic $\overline{T}_{n,r_{1,n}}(\theta)$ is either a Cramér-von-Mises-type (CvM) or Kolmogorov-Smirnov-type (KS) statistic. The CvM statistic is

$$\overline{T}_{n,r_{1,n}}(\theta) := \sup_{\tau \in \mathcal{T}} \sum_{r=1}^{r_{1,n}} (r^2 + 100)^{-1} \sum_{a \in \{1,...,2r\}^{d_X}} (2r)^{-d_x} S(n^{1/2}\overline{m}_n(\theta, \tau, g_{a,r}), \overline{\Sigma}_n(\theta, \tau, g_{a,r})), \quad (3.7)$$

where $S = S_1, S_2, S_3$, or S_4 as defined in (3.9) below, $(r^2 + 100)^{-1}$ is a weight function, and $r_{1,n}$ is a truncation parameter. The asymptotic size and consistency results for the CS's and tests based on $\overline{T}_{n,r_{1,n}}(\theta)$ allow for more general forms of the weight function and hold whether $r_{1,n} = \infty$ or $r_{1,n} < \infty$ and $r_{1,n} \to \infty$ as $n \to \infty$. (No rate at which $r_{1,n} \to \infty$ is needed for these results.) For computational tractability, we typically take $r_{1,n} < \infty$.

The Kolmogorov-Smirnov-type (KS) statistic is

$$\overline{T}_{n,r_{1,n}}(\theta) := \sup_{\tau \in \mathcal{T}} \sup_{g_{a,r} \in \mathcal{G}_{\text{c-cube},r_{1,n}}} S(n^{1/2}\overline{m}_n(\theta, \tau, g_{a,r}), \overline{\Sigma}_n(\theta, \tau, g_{a,r})), \quad (3.8)$$

where $\mathcal{G}_{\text{c-cube},r_{1,n}} = \{g_{a,r} \in \mathcal{G}_{\text{c-cube}} : r \leq r_{1,n}\}$. For brevity, the discussion in this paper focuses on CvM statistics and all results stated concern CvM statistics. Similar results hold for KS statistics.[3]

[2] When $a_u = 1$, the left endpoint of the interval $(0, 1/(2r)]$ is included in the interval.

[3] Such results can be established by extending the results given in Section 13.1 of Appendix B of AS2 and proved in Section 15.1 of Appendix D of AS2.

The functions S_1-S_4 are defined by

$$S_1(m,\Sigma) := \sum_{j=1}^{p}[m_j/\sigma_j]_-^2 + \sum_{j=p+1}^{p+v}[m_j/\sigma_j]^2,$$
$$S_2(m,\Sigma) := \inf_{t=(t_1',0_v')':t_1\in R_{+,\infty}^p}(m-t)'\Sigma^{-1}(m-t),$$
$$S_3(m,\Sigma) := \max\{[m_1/\sigma_1]_-^2,...,[m_p/\sigma_p]_-^2,(m_{p+1}/\sigma_{p+1})^2,...,(m_{p+v}/\sigma_{p+v})^2\}, \text{ and}$$
$$S_4(m,\Sigma) := \inf_{t=(t_1',0_v')':t_1\in R_{+,\infty}^p}(m-t)'(m-t) = \sum_{j=1}^{p}[m_j]_-^2 + \sum_{j=p+1}^{p+v}m_j^2, \quad (3.9)$$

where m_j is the jth element of the vector m, σ_j^2 is the jth diagonal element of the matrix Σ, and $[x]_- := -x$ if $x < 0$ and $[x]_- := 0$ if $x \geq 0$, $R_{+,\infty} := \{x \in R : x \geq 0\} \cup \{+\infty\}$, and $R_{+,\infty}^p := R_{+,\infty} \times ... \times R_{+,\infty}$ with p copies. The functions S_1, S_2, and S_3 are referred to as the modified method of moments (MMM) or Sum function, the quasi-likelihood ratio (QLR) function, and the Max function, respectively. The function S_4 is referred to as the identity-weighted MMM function. The test statistic based on S_4 is not invariant to scale changes of the moment functions, which may be a disadvantage in some examples. But, in other examples (e.g., Examples 2 and 4 above and the $s = 1$ case of Example 1), the moment functions are naturally on a probability scale (i.e., they take values in $[-1, 1]$) and scale invariance is not an issue. In such cases, S_4 is a desirable choice for its simplicity.

4 Critical Values

In this section we define critical values based on bootstrap simulations for the MCMI test statistics. The critical values are of the generalized moment selection (GMS) type, and are obtained via the following steps.[4]

Step 1. Compute the GMS function $\overline{\varphi}_n(\theta,\tau,g_{a,r})$ for $(\tau,g_{a,r}) \in \mathcal{T} \times \mathcal{G}_{\text{c-cube},r_{1,n}}$, where $\overline{\varphi}_n(\theta,g_{a,r})$ is defined as follows. For $g = g_{a,r}$, let

$$\xi_n(\theta,\tau,g) := \kappa_n^{-1}n^{1/2}\overline{D}_n^{-1/2}(\theta,\tau,g)\overline{m}_n(\theta,\tau,g), \text{ where}$$
$$\overline{D}_n(\theta,\tau,g) := Diag(\overline{\Sigma}_n(\theta,\tau,g)), \ \kappa_n := (0.3\ln(n))^{1/2}, \quad (4.1)$$

[4] As demonstrated in Andrews and Soares (2010), Andrews and Shi (2013a, 2014), etc., the GMS-type critical value is preferable to the plug-in asymptotic (PA)-type critical value. In consequence, we omit a discussion of PA critical values.

and $\overline{\Sigma}_n(\theta,\tau,g)$ is defined in (3.4). The jth element of $\xi_n(\theta,\tau,g)$, denoted $\xi_{n,j}(\theta,\tau,g)$, measures the slackness of the moment inequality $E_F m_j(W_i,\theta,\tau,g) \geq 0$ for $j=1,...,p$. It is shrunk towards zero via κ_n^{-1} to ensure that one does not over-estimate the slackness.

Define $\overline{\varphi}_n(\theta,\tau,g) := (\overline{\varphi}_{n,1}(\theta,\tau,g),...,\overline{\varphi}_{n,p}(\theta,\tau,g),0,...,0)' \in R^k$ by

$$\begin{aligned}\overline{\varphi}_{n,j}(\theta,\tau,g) &:= \overline{\Sigma}_{n,j}^{1/2}(\theta,\tau,g)B_n 1\{\xi_{n,j}(\theta,\tau,g) > 1\} \text{ for } j \leq p \text{ and}\\ B_n &:= (0.4\ln(n)/\ln\ln(n))^{1/2},\end{aligned} \tag{4.2}$$

where $\overline{\Sigma}_{n,j}(\theta,\tau,g)$ denotes the (j,j) element of $\overline{\Sigma}_n(\theta,\tau,g)$.

Step 2. Generate B bootstrap samples $\{W_{i,s}^* : i=1,...,n\}$ for $s=1,...,B$ using the standard nonparametric i.i.d. bootstrap. That is, draw $W_{i,s}^*$ randomly with replacement from $\{W_\ell : \ell = 1,...,n\}$ for $i=1,...,n$ and $s=1,...,B$.

Step 3. For each bootstrap sample, transform the regressors as in (3.5) (using the bootstrap sample in place of the original sample) and compute $\overline{m}_{n,s}^*(\theta,\tau,g_{a,r})$ and $\overline{\Sigma}_{n,s}^*(\theta,\tau,g_{a,r})$ just as $\overline{m}_n(\theta,\tau,g_{a,r})$ and $\overline{\Sigma}_n(\theta,\tau,g_{a,r})$ are computed, but with the bootstrap sample in place of the original sample.[5]

Step 4. For each bootstrap sample, compute the bootstrap test statistic $\overline{T}_{n,r_{1,n},s}^*(\theta)$ as $\overline{T}_{n,r_{1,n}}^{CvM}(\theta)$ (or $\overline{T}_{n,r_{1,n}}^{KS}(\theta)$) is computed in (3.7) (or (3.8)) but with $n^{1/2}\overline{m}_n(\theta,\tau,g_{a,r})$ replaced by $n^{1/2}(\overline{m}_{n,s}^*(\theta,\tau,g_{a,r}) - \overline{m}_n(\theta,\tau,g_{a,r})) + \overline{\varphi}_n(\theta,\tau,g_{a,r})$ and with $\overline{\Sigma}_n(\theta,\tau,g_{a,r})$ replaced by $\overline{\Sigma}_{n,s}^*(\theta,\tau,g_{a,r})$.[6] When standardizing the instrumental variables for the bootstrap sample, the original sample mean and sample covariance matrix are used for re-centering and rescaling. Using the bootstrap sample mean and covariance matrix for re-centering and rescaling should yield similar results.

Step 5. Take the bootstrap GMS critical value $c_{n,1-\alpha}^{GMS,*}(\theta)$ to be the $1-\alpha+\eta$ sample quantile of the bootstrap test statistics $\{\overline{T}_{n,r_{1,n},s}^*(\theta) : s=1,...,B\}$ plus η, where η is an infinitesimal positive number that facilitate the proofs but is inconsequential and can be set to zero in practice.

The MCMI CvM (or KS) CS is defined in (2.3) with $T_n(\theta) = \overline{T}_{n,r_{1,n}}^{CvM}(\theta)$ (or $\overline{T}_{n,r_{1,n}}^{KS}(\theta)$) and $c_{n,1-\alpha}(\theta) = c_{n,1-\alpha}^{GMS,*}(\theta)$. The MCMI CvM test of $H_0 : \theta = \theta_*$ rejects H_0 if $\overline{T}_{n,r_{1,n}}^{CvM}(\theta_*) > c_{n,1-\alpha}^{GMS,*}(\theta_*)$. The MCMI KS test is defined likewise using $\overline{T}_{n,r_{1,n}}^{KS}(\theta_*)$ and the KS GMS critical

[5]If the test statistic uses function S_4 defined above, $\overline{\Sigma}_n^*(\theta,\tau,g_{a,r})$ does need to be computed.

[6]If the function S_4 is used, $\overline{\Sigma}_n(\theta,\tau,g_{a,r})$ does not appear in the test statistic, and thus $\overline{\Sigma}_n^*(\theta,\tau,g_{a,r})$ does not enter the calculation of the bootstrap statistic.

value.

The choices of ε, κ_n, and B_n above are the same as those used in AS1, AS2, and Andrews and Shi (2014). These choices are based on some experimentation (in the simulation results reported in AS1 and AS2). They work well in all seven of the simulation examples in those papers as well as in the two simulation examples in this paper. The asymptotic results reported in the Appendix allow for other choices. The robustness of the finite-sample properties of the tests to the choice of these tuning parameters is documented in the Appendix for the two simulation examples considered in this paper.

The number of cubes with side-edge length indexed by r is $(2r)^{d_X}$, where d_X denotes the dimension of the covariate X_i. The computation time is approximately linear in the number of cubes. Hence, it is linear in $N_g := \sum_{r=1}^{r_{1,n}} (2r)^{d_X}$.

When there are discrete variables in X_i, the sets $C_{a,r}$ can be formed by taking interactions of each value of the discrete variable(s) with cubes based on the other variable(s).

5 Correct Asymptotic Size

In this section, we show that the CS defined above has correct asymptotic size.

5.1 Main Result

First, we introduce some additional notation. Define the asymptotic variance-covariance kernel, $\{h_{2,F}(\theta, \tau, g, \tau^\dagger, g^\dagger) : (\tau, g), (\tau^\dagger, g^\dagger) \in \mathcal{T} \times \mathcal{G}_{\text{c-cube}}\}$, of $n^{1/2}\overline{m}_n(\theta, \tau, g)$ after normalization via a diagonal matrix $D_F^{-1/2}(\theta)$. That is, we define

$$\begin{aligned} h_{2,F}(\theta, \tau, g, \tau^\dagger, g^\dagger) &:= D_F^{-1/2}(\theta)\Sigma_F(\theta, \tau, g, \tau^\dagger, g^\dagger)D_F^{-1/2}(\theta), \text{ where} \\ \Sigma_F(\theta, \tau, g, \tau^\dagger, g^\dagger) &:= Cov_F((m(W_i, \theta, \tau, g), m(W_i, \theta, \tau^\dagger, g^\dagger)), \\ D_F(\theta) &:= Diag(\sigma_{F,1}^2(\theta), \ldots, \sigma_{F,k}^2(\theta)), \end{aligned} \tag{5.1}$$

and $\{\sigma_{F,j}(\theta) : j = 1, \ldots, k\}$ are specified just before Assumption PS1. For simplicity, let $h_{2,F}(\theta)$ abbreviate $\{h_{2,F}(\theta, \tau, g, \tau^\dagger, g^\dagger) : (\tau, g), (\tau^\dagger, g^\dagger) \in \mathcal{T} \times \mathcal{G}_{\text{c-cube}}\}$.

Define the set of variance-covariance kernels

$$\mathcal{H}_2 := \{h_{2,F}(\theta) : (\theta, F) \in \mathcal{F}\}, \tag{5.2}$$

where, as defined at the end of Section 2, $\mathcal{F}$ is the subset of $\mathcal{F}_+$ that satisfies Assumption PS3. On the space of $k \times k$ matrix-valued covariance kernels on $(\mathcal{T} \times \mathcal{G}_{\text{c-cube}})^2$, which is a superset of $\mathcal{H}_2$, we use the uniform metric d defined by

$$d(h_2^{(1)}, h_2^{(2)}) := \sup_{(\tau,g),(\tau^\dagger,g^\dagger)\in\mathcal{T}\times\mathcal{G}_{\text{c-cube}}} \|h_2^{(1)}(\tau, g, \tau^\dagger, g^\dagger) - h_2^{(2)}(\tau, g, \tau^\dagger, g^\dagger)\|. \tag{5.3}$$

Correct asymptotic size is established in the following theorem. The theorem is implied by Lemmas D.1 and D.2 in Appendix D, where the lemmas are also proved. We provide a brief sketch of the proof in the next subsection, highlighting the difference with the analogous result in AS1. The role of η is also explained in the next subsection.

Theorem 5.1 *Suppose Assumption* SIG1 *holds. For any compact subset $\mathcal{H}_{2,cpt}$ of $\mathcal{H}_2$, the* MCMI *confidence set CS_n satisfies*

$$\liminf_{n\to\infty} \inf_{\substack{(\theta,F)\in\mathcal{F}: \\ h_{2,F}(\theta)\in\mathcal{H}_{2,cpt}}} P_F(\theta \in CS_n) \geq 1 - \alpha.$$

Comments. 1. Theorem 5.1 shows that the MCMI CS has correct uniform asymptotic size over compact sets of covariance kernels. The uniformity results hold whether the moment conditions involve "weak" or "strong" IV's X_i. That is, weak identification of the parameter θ due to a low correlation between X_i and the functions $m_j(W_i, \theta, \tau)$ does not affect the uniformity results.

2. The proofs in the Appendix take the transformation of the IV's to be non-data dependent. One could extend the results to allow for data-dependence by considering random hypercubes as in Pollard (1979) and Andrews (1988). These results show that one obtains the same asymptotic results with random hypercubes as with nonrandom hypercubes that converge in probability to nonrandom hypercubes (in an L^2 sense). For brevity, we do not do so.

5.2 Sketch of the Proof of Theorem 5.1 and the Role of η

A sketch of the proof of Theorem 5.1. The theorem is proved using several steps. While the steps are the same as those used to prove the analogous result (Theorem 2(a)) in AS1, notational modifications, and occasionally more substantial modifications to the arguments that complete each step are needed.

First, we use the compactness of $\mathcal{H}_{2,cpt}$ and the definitions of infimum and lim inf to write $\liminf_{n\to\infty} \inf_{\substack{(\theta,F)\in\mathcal{F}: \\ h_{2,F}(\theta)\in\mathcal{H}_{2,cpt}}} P_F(\theta \in CS_n)$ as $\lim_{n\to\infty} P_{F_{a_n}}(\overline{T}_{a_n,r_{1,a_n}}(\theta_{a_n}) \leq c^{GMS,*}_{a_n,1-\alpha}(\theta_{a_n}))$, where $\{a_n\}_{n\geq 1}$ is a subsequence of $\{n\}$, and $\{(\theta_{a_n}, F_{a_n})\}$ is a sequence in $\mathcal{F}$ such that $h_{2,F_{a_n}}(\theta_{a_n}) \to h_2$ for some $h_2 \in \mathcal{H}_{2,cpt}$. This step is the same as the analogous step in the proof of Theorem 2(a) of AS1.

Next, we show an asymptotic distributional approximation for $\overline{T}_{a_n,r_{1,a_n}}(\theta_{a_n})$:

$$\liminf_{n\to\infty} \left[\Pr{}_{F_{a_n}}(\overline{T}_{a_n,r_{1,a_n}}(\theta_{a_n}) \leq x + c) - \Pr(\overline{T}_{a_n,F_{a_n}}(\theta_{a_n}) \leq x)\right] \geq 0, \tag{5.4}$$

for any $x \in R$ and $c > 0$, where

$$\begin{aligned} \overline{T}_{n,F_n}(\theta_n) = \sup_{\tau\in\mathcal{T}} \sum_{r=1}^{r_{1,n}} (r^2+100)^{-1} \sum_{a\in\{1,\ldots,2r\}^{d_X}} (2r)^{-d_x} \times \\ S(\nu_{h_{2,F_n}(\theta_n)}(\tau, g_{a,r}) + h_{1,n,F_n}(\theta_n, \tau, g_{a,r}), h^{\varepsilon}_{2,F_n}(\theta_n, \tau, g_{a,r})), \end{aligned} \tag{5.5}$$

$\nu_{h_{2,F_n}(\theta_n)}$ is a Gaussian process indexed by $(\tau, g_{a,r})$ with variance-covariance kernel $h_{2,F_n}(\theta_n, \tau_1, g_1, \tau_2, g_2)$, $h_{1,n,F_n}(\theta_n, \tau, g) = n^{1/2} E_{F_n} m(W_i, \theta_n, \tau, g)$, and $h^{\varepsilon}_{2,F_n}(\theta_n, \tau, g) = h_{2,F_n}(\theta_n, \tau, g, \tau, g) + \varepsilon I_k$. The approximation (5.4) is proved using the weak convergence of the empirical process $\{a_n^{1/2}[\overline{m}_{a_n}(\theta_{a_n}, \tau, g) - E_{F_{a_n}} m(W_i, \theta_{a_n}, \tau, g)] : (\tau, g) \in \mathcal{T} \times \mathcal{G}\}$ to the Gaussian process, the in-probability convergence of $\{\overline{\Sigma}_{a_n}(\theta_{a_n}, \tau, g) : (\tau, g) \in \mathcal{T} \times \mathcal{G}\}$, and the continuity of the S function. This step is similar to the analogous step in the proof of Theorem 2(a) in AS1 (which is composed of the proofs of Theorem 1 and Lemma A1 in AS2), but substantively differs from the latter in two places:

- Lemma A1 of AS1 establishes the weak convergence for the empirical process involved in AS1, which is indexed by g only. In the present paper, the empirical process is indexed by both τ and g. To account for the double index, we present and prove a lemma (Lemma D.5) that takes advantage of a stability formula for covering numbers.

- The proof of Theorem 2(a) of AS1 employs a dominated convergence argument that is suitable for the pure CvM-type test statistic considered in AS1. On the other hand, we consider a KS-CvM hybrid statistic that takes a supremum over τ and integrates over

g, for which the dominated convergence argument does not apply. Instead, we rewrite Assumption S2 of AS1, which is the continuity assumption on S, in an equivalent but more convenient form, and use that to establish sup-norm convergence. Detailed arguments are given at the end of the proof of Theorem D.3 in the Appendix.

The result (5.4) implies immediately that

$$\liminf_{n\to\infty} \Pr{}_{F_{a_n}}(\overline{T}_{a_n,r_{1,a_n}}(\theta_{a_n}) \leq \bar{c}_{a_n,1-\alpha}(\theta_{a_n}) + c) \geq 1-\alpha, \tag{5.6}$$

for any $c > 0$, where $\bar{c}_{n,1-\alpha}(\theta_n)$ is the $1-\alpha$ quantile of $\overline{T}_{n,F_n}(\theta_n)$. The rest of the proof shows that the bootstrap critical value satisfies

$$\limsup_{n\to\infty} \Pr{}_{F_{a_n}}(c^{GMS,0}_{a_n,1-\alpha+c}(\theta_{a_n}) \leq \bar{c}_{a_n,1-\alpha}(\theta_{a_n}) - c_1) = 0, \tag{5.7}$$

for any positive constants c, c_1, where $c^{GMS,0}_{a_n,1-\alpha+c}(\theta_{a_n})$ is defined as $c^{GMS,*}_{a_n,1-\alpha+c}(\theta_{a_n})$ is defined except with $\eta = 0$. This step is similar to the analogous step in the proof of Theorem 2(a) of AS1. However, AS1 gives explicit arguments only for the asymptotic approximation critical value and not for the bootstrap critical value. In the present paper, we prove bootstrap validity explicitly. The arguments for the bootstrap are given in Lemma D.4 in Appendix D. ■

Next we explain the role of η in the above proof. First, note that the infinitesimal number η is added to two places in the critical value: to the conditional quantile and to the confidence level. The η added to the conditional quantile is needed due to the c in (5.4) and (5.6). To make these two equations hold with $c = 0$, one would need to establish a uniform (over n) anti-concentration bound for the distribution of $\overline{T}_{a_n,F_{a_n}}(\theta_{a_n})$. While such a bound has been derived in Chernozhukov, Chetverikov, and Kato (2013, 2014a,b) for the supremum of a Gaussian process, it is to our knowledge not available for $\overline{T}_{a_n,F_{a_n}}(\theta_{a_n})$, which is not a supremum of a Gaussian process even for the KS test statistic.

The η added to the confidence level is due to the c in (5.7). There are two ways to eliminate this η. One is by imposing a uniform (over n) lower bound on the slope of the distribution function of $\overline{T}_{a_n,F_{a_n}}(\theta_{a_n})$ around its $1-\alpha$ quantile. This would make (5.7) hold with $c = 0$. However, such a bound is difficult to verify. Another way is to strengthen (5.7) so that c is replaced by $c_n \to 0$, as done in Chernozhukov, Chetverikov, and Kato (2014a).

This would require either a Berry-Esseen type distributional convergence rate result for the empirical process $\{a_n^{1/2}[\overline{m}_{a_n}(\theta_{a_n}, \tau, g) - E_{F_{a_n}} m(W_i, \theta_{a_n}, \tau, g)] : (\tau, g) \in \mathcal{T} \times \mathcal{G}\}$ or such a result for the KS or CvM test statistic. Neither is available to our knowledge.

6 Power Against Fixed Alternatives

We now show that the power of the MCMI test converges to one as $n \to \infty$ for all fixed alternatives (for which Assumptions PS1 and PS2 hold). This implies that for any fixed distribution F_0 and any parameter value θ_* *not* in the identified set Θ_{F_0}, the MCMI CS excludes θ_* with probability approaching one. In this sense, MCMI CS based on $T_n(\theta)$ fully exploits the infinite number of conditional moment inequalities/equalities. CS's based on a finite number of unconditional moment inequalities/equalities do not have this property.[7]

The null hypothesis is

$$\begin{aligned} H_0 : E_{F_0}[m_j(W_i, \theta_*, \tau)|X_i] &\geq 0 \text{ a.s. } [F_{X,0}] \text{ for } j = 1, ..., p \text{ and} \\ E_{F_0}[m_j(W_i, \theta_*, \tau)|X_i] &= 0 \text{ a.s. } [F_{X,0}] \text{ for } j = p+1, ..., k, \ \forall \tau \in \mathcal{T}, \end{aligned} \tag{6.1}$$

where θ_* denotes the null parameter value and F_0 denotes the fixed true distribution of the data. The alternative hypothesis is $H_1 : H_0$ does not hold. The following assumption specifies the properties of fixed alternatives (FA).

Let $\mathcal{F}_+$ be as defined in Section 2.3. Note that $\mathcal{F}_+$ includes (θ, F) pairs for which θ lies outside of the identified set Θ_F as well as all values in the identified set.

The set $\mathcal{X}_F(\theta, \tau)$ of values x for which the moment inequalities or equalities evaluated at θ are violated under F is defined as follows. For any $\theta \in \Theta$ and any distribution F with $E_F[\|m(W_i, \theta, \tau)\|] < \infty$, let

$$\begin{aligned} \mathcal{X}_F(\theta, \tau) := \{x \in R^{d_x} : \ & E_F[m_j(W_i, \theta, \tau)|X_i = x] < 0 \text{ for some } j \leq p \text{ or} \\ & E_F[m_j(W_i, \theta, \tau)|X_i = x] \neq 0 \text{ for some } j = p+1, ..., k\}. \end{aligned} \tag{6.2}$$

[7]This holds because the identified set based on a finite number of moment inequalities typically is larger than the identified set based on all the conditional moment inequalities. In consequence, CI's based on a finite number of inequalities include points in the difference between these two identified sets with probability whose limit infimum as $n \to \infty$ is $1 - \alpha$ or larger even though these points are not in the identified set based on the conditional moment inequalities.

The next assumption, Assumption MFA, states that violations of the conditional moment inequalities or equalities occur for the null parameter θ_* for X_i values in a set with positive probability under F_0 for some $\tau \in \mathcal{T}$. Thus, under Assumption MFA, the moment conditions specified in (6.1) do not hold.

Assumption MFA. The null value $\theta_* \in \Theta$ and the true distribution F_0 satisfy: (a) for some $\tau_* \in \mathcal{T}$, $P_{F_0}(X_i \in \mathcal{X}_{F_0}(\theta_*, \tau_*)) > 0$ and (b) $(\theta_*, F_0) \in \mathcal{F}_+$.

We employ the following assumption on the weights $\{\widehat{\sigma}^2_{n,j}(\theta) : j \leq k, n \geq 1\}$.

Assumption SIG2. For all $\zeta > 0$, $\Pr_{F_0}\left(\max_{j \leq k} |\widehat{\sigma}^2_{n,j}(\theta_*)/\sigma^2_{F_0,j}(\theta_*) - 1| > \zeta\right) \to 0$.

Note that Assumption SIG2 is not implied by Assumption SIG1 because (θ_*, F_0) does not belong to $\mathcal{F}$.

The following Theorem shows that the MCMI test is consistent against all fixed alternatives that satisfy Assumption MFA.

Theorem 6.1 *Suppose Assumptions* MFA *and* SIG2 *hold. Then, the* MCMI *test satisfies*

$$\lim_{n\to\infty} P_{F_0}(T_n(\theta_*) > c^{GMS,*}_{n,1-\alpha}(\theta_*)) = 1.$$

Theorem 6.1 is implied by Theorem E.1 in Appendix E, where the latter is proved. The proof is composed of two parts. First, we show that $n^{-1}T_n(\theta_*)$ converges in probability to a positive quantity, and second, we show that the critical value is $O_p(1)$. The first part combines the proofs of the fixed alternative results for the KS and CvM cases in AS1 and AS2. The second part is the same as the analogous part in AS2 up to notational changes.

7 Example 1: Conditional Stochastic Dominance

In this section, we apply the general theory developed above to Example 1. We first establish primitive sufficient conditions for Assumptions PS1 and PS2 for this example, and then carry out a simple Monte Carlo experiment for testing first-order stochastic dominance.

7.1 Verification of Assumptions

We treat the first-order stochastic dominance case separately in our discussion from the higher-order stochastic dominance case because it allows for weaker assumptions on the distributions of Y_1 and Y_2.

7.1.1 First-Order Stochastic Dominance

Recall that the conditional moment inequalities implied by first-order conditional stochastic dominance are

$$E_{F_0}[1\{Y_2 \leq \tau\} - 1\{Y_1 \leq \tau\}|X] \geq 0 \text{ a.s. } \forall \tau \in \mathcal{T}. \tag{7.1}$$

The moment conditions for this model do not depend on a parameter θ. Hence, to fit the notation with that of the general theory, we set $\Theta = \{0\}$ (without loss of generality). Also observe that $p = k = 1$ in this example.

For this example, we use $\sigma_{F,1}(0) = \widehat{\sigma}_{n,1}(0) = 1$ for all F because the moment function has a natural scale. Hence, Assumptions SIG1 and SIG2 hold.

Lemma 7.1 *Let $\mathcal{F}_+$ be the set of $(0, F)$ such that $\{(Y_{1,i}, Y_{2,i}, X_i')' : i \geq 1\}$ are i.i.d. under F. Then, $\mathcal{F}_+$ satisfies Assumptions* PS1 and PS2 *with $M(w) = 1$, $\delta > 0$, and $C_1 = 1$.*

The proof of the lemma is given in Appendix F. The core part of the proof is the verification of Assumption PS2, which is done via the pseudo-dimension bound on the covering numbers of the set $\{1\{y_2 \leq \tau\} - 1\{y_1 \leq \tau\} : \tau \in \mathcal{T}\}$ and the fact that the pseudo-dimension of the set is at most one (by Lemma 4.4 of Pollard (1990)).

7.1.2 Higher-Order Stochastic Dominance

The conditional moment inequalities implied by sth-order conditional stochastic dominance for $s > 1$ are

$$E_{F_0}[(\tau - Y_2)^{s-1} 1\{Y_2 \leq \tau\} - (\tau - Y_1)^{s-1} 1\{Y_1 \leq \tau\}|X] \geq 0 \text{ a.s. } \forall \tau \in \mathcal{T}. \tag{7.2}$$

As above, we set $\Theta = \{0\}$. In this example, $p = k = 1$.

For this example, we use $\sigma^2_{F,1}(0) = E_F[(Y_1 - E(Y_1))^2] + E_F[(Y_2 - E(Y_2))^2]$ and $\widehat{\sigma}^2_{n,1}(0) = n^{-1}\sum_{i=1}^n [(Y_{1,i} - \overline{Y}_{1,n})^2 + (Y_{2,i} - \overline{Y}_{2,n})^2]$, where $\overline{Y}_{j,n} = n^{-1}\sum_{i=1}^n Y_{j,n}$ for $j = 1, 2$.

Lemma 7.2 *Suppose $s > 1$. Let $\underline{\sigma} > 0$ and $B \in (0, \infty)$ be constants. Let $\mathcal{F}_+$ be the set of (θ, F) for which* (i) $\theta \in \Theta$, (ii) $\{(Y_{1,i}, Y_{2,i}, X_i')' : i \geq 1\}$ *are i.i.d. under* F, (iii) $\sigma^2_{F,1}(0) \geq \underline{\sigma}^2$, *and* (iv) $\mathcal{T} \subset [-B, B]$. *Then,*

(a) $\mathcal{F}_+$ *satisfies Assumptions* PS1 and PS2 *with* $M(w) = [(B - y_2)^{s-1} + (B - y_1)^{s-1}]/\sigma_{F,1}(0)$, $\delta > 0$, *and* $C_1 = 2^{s(2+\delta)} B^{(s-1)(2+\delta)} \underline{\sigma}^{-(2+\delta)}$, *and*

(b) *Assumptions* SIG1 *and* SIG2 *hold.*

The verification of Assumption PS2 in this case also uses the pseudo-dimension bound on the covering numbers. Unlike in Lemma 7.1, the pseudo-dimension of the set of standardized moment functions is not obvious. We prove that the pseudo-dimension is at most one.

7.2 Monte Carlo Results

In this subsection, we report Monte Carlo results for testing the first-order conditional stochastic dominance between the conditional distributions of Y_1 and Y_2 given X. That is, we test the null hypothesis:

$$E_{F_0}[1\{Y_2 \leq \tau\} - 1\{Y_1 \leq \tau\}|X] \geq 0 \text{ a.s. } \forall \tau \in \mathcal{T} \equiv R, \tag{7.3}$$

where Y_1, Y_2, X are scalar random variables. We consider the MCMI tests proposed above based on the CvM and KS test statistics combined with the GMS critical value. For comparative purposes, we also consider the CvM and KS test statistics combined with sub-sampling critical values, as well as the two-step multiplier bootstrap method (CCK-MB) and the two-step empirical bootstrap method (CCK-EB) proposed in Chernozhukov, Chetverikov, and Kato (2014a, CCK hereafter).[8]

In this example, we take the instrument X to have the uniform $[0, 1]$ distribution and take Y_1 and Y_2 to have log-normal distributions given X:

$$\begin{aligned} Y_1 &= \exp(\sigma_1(X) Z_1 + \mu_1(X)) \text{ and} \\ Y_2 &= \exp(\sigma_2(X) Z_2 + \mu_2(X)), \end{aligned} \tag{7.4}$$

where $\sigma_1(X)$, $\mu_1(X)$, $\sigma_2(X)$, and $\mu_2(X)$ determine whether and how the null hypothesis that

[8] These methods have the best power among the six one-step and two-step methods proposed in CCK. The three-step methods proposed in CCK are not applicable in this model.

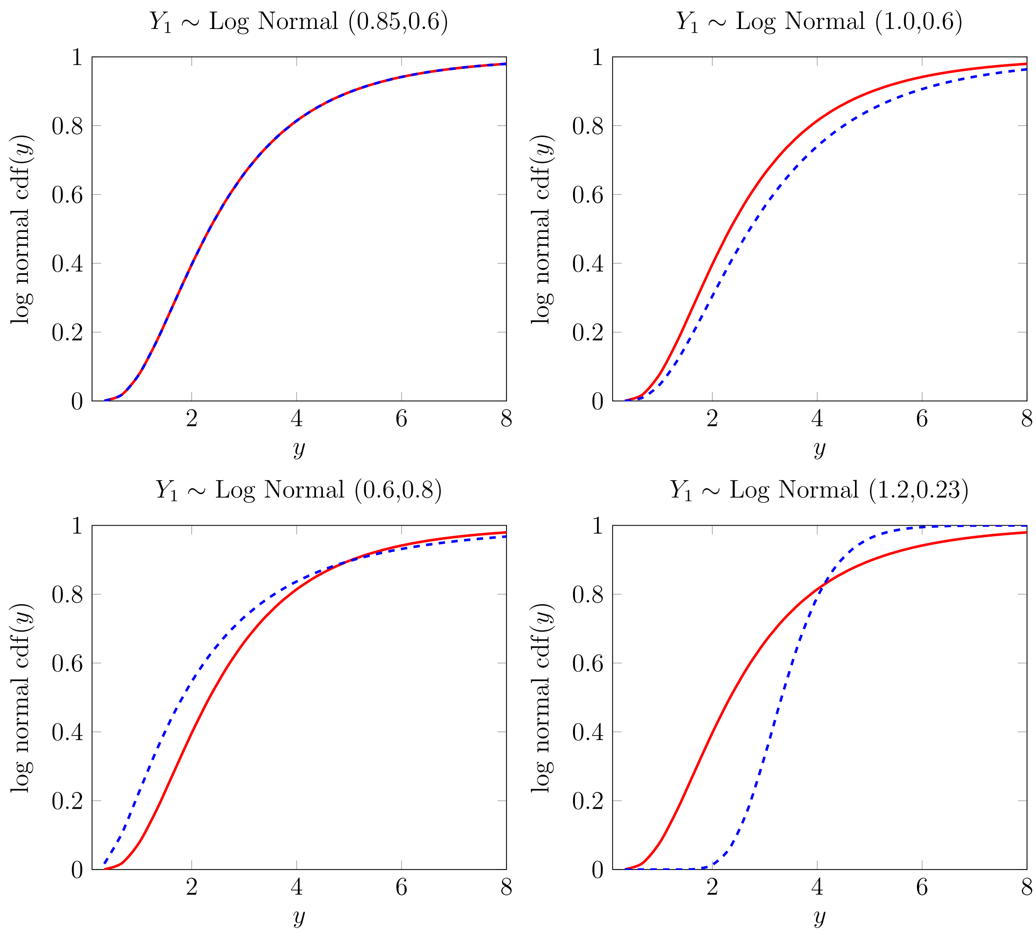

Figure 1: Conditional CDF's of Y_1 (dashed blue) and Y_2 (solid red) given $X = 1$. In all graphs, $Y_2 \sim$ Log Normal $(0.85, 0.6)$.

Y_1 first-order stochastically dominates Y_2 given X is violated.

To generate the simulated data, we let $\mu_1(X) = c_1 X + c_3$, $\sigma_1(X) = c_2 X + c_4$, $\mu_2(X) = 0.85$, and $\sigma_2(X) = 0.6$. These data-generating processes (DGPs) are adapted from Barrett and Donald (2003). Four values of $\mathbf{c} := (c_1, c_2, c_3, c_4)$ are considered: $\mathbf{c}_A = (0, 0, 0.85, 0.6)$, $\mathbf{c}_B = (0.15, 0, 0.85, 0.6)$, $\mathbf{c}_C = (-0.25, 0.2, 0.85, 0.6)$, and $\mathbf{c}_D = (0.35, 0, 0.85, 0.23)$. With $\mathbf{c}_A$ and $\mathbf{c}_B$, the null that Y_1 first-order stochastically dominate Y_2 conditional on X holds, while with $\mathbf{c}_C$ and $\mathbf{c}_D$, the null hypothesis is violated. To visualize the nature of the DGPs, we draw in Figure 1 the conditional cdf's of Y_1 and Y_2 given $X = 1$ at these four $\mathbf{c}$ values.

Note that with $\mathbf{c}_A$, Y_1 and Y_2 have identical distributions conditional on X. In this case, all of the moment inequalities are binding. The test should ideally have rejection probability equal to its nominal level in this boundary case. For this reason, we use this DGP to size-correct the rejection probabilities under the two alternative DGP's $\mathbf{c}_C$ and $\mathbf{c}_D$.

In the implementation of the tests, we compute the supremum over $\mathcal{T}$ by discretization. Specifically, we approximate $\mathcal{T}$ by N_τ points in $\mathcal{T}$ for a positive integer N_τ. The N_τ points on $\mathcal{T}$ are chosen as follows: first pool the n observations of Y_1 and those of Y_2 to get a sample of size $2n$. Then use as grid points the $1/(N_\tau+1), 2/(N_\tau+1), \ldots, N_\tau/(N_\tau+1)$ percentiles of this $2n$ sample.

For the sample size and the tuning parameters of all tests considered, we consider a base case with the sample size $n = 250$, the hypercube parameter $r_{1,n} = 3$, and $N_\tau = 25$. Then, for comparison, we also consider three variations of the base case where each differs from the base case in only one dimension.[9] We set η to zero in all cases for our methods. For the sub-sampling critical values, we use a subsample size of 20. For the CCK methods, we take the tuning parameters from CCK's Monte Carlo simulations.

Table 1: Null Rejection Probabilities for Nominal .05 First-Order Stochastic Dominance Tests

	CvM/GMS	KS/GMS	CvM/Sub	KS/Sub	CCK-MB	CCK-EB
	Null 1: $(c_1, c_2, c_3, c_4) = (0, 0, 0.85, 0.6)$					
Base case: ($n = 250$, $r_{1,n} = 3$, $N_\tau = 25$)	.057	.064	.071	.213	.035	.018
$n = 500$	.049	.052	.079	.212	.032	.029
$r_{1,n} = 4$	.059	.055	.098	.282	.027	.010
$N_\tau = 30$	.062	.068	.085	.239	.034	.014
	Null 2: $(c_1, c_2, c_3, c_4) = (0.15, 0, 0.85, 0.6)$					
Base case: ($n = 250$, $r_{1,n} = 3$, $N_\tau = 25$)	.014	.019	.029	.131	.011	.006
$n = 500$	.009	.014	.017	.089	.013	.010
$r_{1,n} = 4$	.014	.019	.039	.192	.007	.006
$N_\tau = 30$	.018	.019	.037	.137	.011	.007

Note: For computation reasons, not all subsamples are used. The bootstrap and sub-sampling critical values both use 1000 repetitions to simulate the critical values. The two-step version of CCK's MB and EB methods are used.

[9] More variations are considered in the additional Monte Carlo exercise in Appendix G.

Simulated rejection probabilities based on 1000 simulation repetitions are reported in Tables 1 and 2. Table 1 reports the rejection probabilities under the two null DGP's and Table 2 reports the size-corrected rejection probabilities under the two alternative DGP's. As the tables show, the CvM/GMS test performs the best overall in that it has the most accurate size and the highest power. The KS/GMS test has somewhat worse power perhaps due to the DGP designs. The CvM/Sub-sampling test has greater over-rejections than, and comparable power to, the CvM/GMS test, while the KS/Sub-sampling test exhibit severe over-rejections. The CCK tests have good size control, but somewhat lower power than the KS/GMS test and significantly lower (size-corrected) power than the CvM/GMS test.

Table 2: Size-corrected Power for Nominal .05 First-order Stochastic Dominance Tests

	CvM/GMS	KS/GMS	CvM/Sub	KS/Sub	CCK-MB	CCK-EB
	Alternative 1: $(c_1, c_2, c_3, c_4) = (-0.25, 0.2, 0.85, 0.6)$					
Base case: ($n = 250$, $r_{1,n} = 3$, $N_\tau = 25$)	.505	.379	.463	.281	.301	.210
$n = 500$	.809	.689	.806	.603	.596	.525
$r_{1,n} = 4$	.509	.367	.475	.272	.254	.148
$N_\tau = 30$	.470	.405	.443	.297	.309	.202
	Alternative 2: $(c_1, c_2, c_3, c_4) = (0.35, 0, 0.85, 0.23)$					
Base case: ($n = 250$, $r_{1,n} = 3$, $N_\tau = 25$)	.581	.295	.622	.346	.204	.178
$n = 500$	.942	.768	.946	.767	.670	.665
$r_{1,n} = 4$	.609	.246	.643	.335	.168	.131
$N_\tau = 30$	.539	.309	.598	.350	.208	.172

Note: The bootstrap and sub-sampling critical values use 1000 repetitions to simulate the critical values. Size correction is carried out using the null DGP with $(c_1, c_2, c_3, c_4) = (0, 0, 0.85, 0.6)$.

8 Example 2: Random-Coefficients Binary-Outcome Models with Instrumental Variables

We focus on the model given in (2.9), and restated here for the reader's convenience:

$$E_{F_0}[F_\beta(\mathcal{S},\theta) - 1\{S(Y_1, Y_2, X_1) \subset \mathcal{S}\}|X_1, X_2] \geq 0 \text{ a.s. } \forall \mathcal{S} \in \mathbf{S}, \tag{8.1}$$

where Y_1 is the binary dependent variable, Y_2 is a d_2-dimensional endogenous covariate, X_1 is a d_1-dimensional exogenous covariate, and X_2 is a vector of instruments.

8.1 Verification of Assumptions

We first note that, when $d_1 + d_2 > 1$, the manageability assumption, Assumption PS2, does not hold in general because the Vapnik-Chervonenkis (VC) dimension of the set $\{(1\{S(Y_{1,i}, Y_{2,i}, X_{1,i}) \subset \mathcal{S}\})_{i=1}^n : \mathcal{S} \in \mathbf{S}\}$ typically diverges to infinity as $n \to \infty$. Thus, we need to restrict attention to a subset of $\mathbf{S}$. Fortunately, in many applications, restriction to an appropriate subset of $\mathbf{S}$ (specified below) does not affect the set identification power of the model. We apply our general theory to such applications.

For a positive integer m, we consider subsets of $\mathbf{S}$ of the form: $\mathbf{S}_m := \{\cup_{j=1}^m H(c_j) : c_j \in R^{d_\beta} \backslash \{0^{d_\beta}\}\}$. That is, $\mathbf{S}_m$ is the collection of at most m unions of half-spaces in R^{d_β} through the origin. Let $\Theta_F(\mathbf{S}_m) := \{\theta \in \Theta : E_F[F_\beta(\mathcal{S},\theta) - 1\{S(Y_1, Y_2, X_1) \subset \mathcal{S}\}|X_1, X_2] \geq 0 \text{ a.s. } \forall \mathcal{S} \in \mathbf{S}_m\}$. Define $\Theta_F(\mathbf{S})$ analogously. The applications we consider are required to satisfy the following assumption. This assumption is satisfied in Example 2 of CR with $m = 2$ and Example 3 of CR with $m = 4$. This assumption is always satisfied when $d_1 + d_2 = 1$ because in that case $\mathbf{S}_m = \mathbf{S}$ for $m = 2$.

Assumption V1. $\boldsymbol{\Theta_{F_0}(\mathbf{S_m}) = \Theta_{F_0}(\mathbf{S})}$.

Under this assumption, we can base inference on the conditional moment inequality model:

$$E_F[F_\beta(\mathcal{S},\theta) - 1\{S(Y_1, Y_2, X_1) \subset \mathcal{S}\}|X_1, X_2] \geq 0 \text{ a.s. } \forall \mathcal{S} \in \mathbf{S}_m. \tag{8.2}$$

We first write $S(y_1, y_2, x_1)$ in the canonical form of a half-space:

$$S(y_1, y_2, x_1) = cl\{b = (b_0, b_1', b_2')' \in R^{d_\beta} : y_1 = 1\{b_0 + b_1' x_1 + b_2' y_2 \geq 0\}\}$$
$$= H((y_1 - 1/2)(1, x_1', y_2')'). \quad (8.3)$$

The following lemma yields a convenient representation of the event $\{S(Y_1, Y_2, X_1) \subset \mathcal{S}\}$ for $\mathcal{S} \in \mathbf{S}_m$.

Lemma 8.1 *For any $c_1, ..., c_m \in R^{d_\beta} \backslash \{0^{d_\beta}\}$ (not necessarily distinct from each other), there exists a $d_\beta \times M$ real matrix $B(c_1, ..., c_m)$ with $M = \max_{j \in \{1,...,d_\beta\}} \left[\binom{m}{\min\{j,m\}-1} + 2(d_\beta - j)\right]$ such that, for any $\bar{c} \in R^{d_\beta} \backslash \{0^{d_\beta}\}$, the following statements are equivalent:*

(a) $H(\bar{c}) \subset \cup_{j=1}^m H(c_j)$,

(b) $\bar{c} = \sum_{j=1}^m \lambda_j c_j$ *for some* $\lambda_1, ..., \lambda_m \geq 0$, *and*

(c) $B(c_1, ..., c_m)' \bar{c} \geq 0^M$.

The lemma implies that the conditional moment inequality model (8.2) has the following equivalent representation:

$$E_F[F_\beta(\mathcal{S}(\tau), \theta) - 1\{(Y_1 - 1/2) B(\tau)'(1, X_1', Y_2')' \geq 0\} | X_1, X_2] \geq 0 \text{ a.s. } \forall \tau \in \mathcal{T}, \quad (8.4)$$

where $\mathcal{T} = \{\tau = (c_1, ..., c_m) : c_1, ..., c_m \in R^{d_\beta} \backslash \{0^{d_\beta}\}\}$, $B(\tau) := B(c_1, ..., c_m)$, and $\mathcal{S}(\tau) = \cup_{j=1}^m H(c_j)$.

The equivalent representation just given is instrumental in proving the lemma below, which verifies the high-level conditions for this example. Note that in this example, $p = k = 1$. We use $\sigma_{F,1}(\theta) = \widehat{\sigma}_{n,1}(\theta) = 1$ for all (θ, F) because the moment function has a natural scale. Hence, Assumptions SIG1 and SIG2 hold.

Lemma 8.2 *For the model in (8.2), let $\mathcal{F}_+$ be the set of (θ, F) such that $\theta \in \Theta$ and $\{(Y_{1,i}, Y_{2,i}', X_i')' : i \geq 1\}$ are i.i.d. under F. Then $\mathcal{F}_+$ satisfies Assumptions* PS1 and PS2 *with $M(w) = 1$, $\delta > 0$, and $C_1 = 1$.*

The main part of the proof of Lemma 8.2 is the verification of Assumption PS2. To verify this assumption, we use a pseudo-dimension bound for covering numbers (specifically, Lemma 4.1 of Pollard (1990)). We show that the pseudo-dimension is finite by applying Lemma 4.4 of Pollard (1990) to the equivalent representation in (8.4)[10]

[10]Note that the representation (8.4) is simply a technical device useful for the theory and for intuitive

8.2 Monte Carlo Results

In this subsection, we report Monte Carlo results for a binary choice model similar to the numerical example in CR. The model has one endogenous regressor (Y_2), one instrument variable (X), and no exogenous regressors. That is,

$$Y_1 = 1\{\beta_0 + \beta_1 Y_2 < 0\} \text{ with } (\beta_0, \beta_1) \perp X. \tag{8.5}$$

Further, we take β_0 and β_1 to be jointly normally distributed: $\beta_0 = \alpha_0 + U_0$ and $\beta_1 = \alpha_1 + U_1$, where

$$\begin{pmatrix} U_0 \\ U_1 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \gamma_0 \\ \gamma_0 & \gamma_1 + \gamma_0^2 \end{pmatrix}\right).$$

Thus, the model contains the unknown parameter $\theta = (\alpha_0, \alpha_1, \gamma_0, \gamma_1)'$. CR show that the sharp identified set for θ is characterized by the following conditional moment inequalities:

$$E_{F_0}[F_\beta(\mathcal{S}, \theta) - 1\{S(Y_1, Y_2) \subset \mathcal{S}\}|X] \geq 0 \text{ a.s. } \forall \mathcal{S} \in \mathbf{S}, \tag{8.6}$$

where, using the half-space notation $H(\cdot)$ defined in (2.8),

$$\begin{aligned} S(y_1, y_2) &= H((y_1 - 1/2)(1, y_2)'), \\ \mathbf{S} &= \{\mathcal{S}_{\tau_1,\tau_2} = H(\cos\tau_1, \sin\tau_1) \cup H(\cos\tau_2, \sin\tau_2) : 0 \leq \tau_1 \leq \tau_2 \leq 2\pi\}, \\ F_\beta(\mathcal{S}_{\tau_1,\tau_2}, \theta) &= 1 - \Phi(-m_1, -m_2, \rho), \end{aligned} \tag{8.7}$$

$\Phi(x_1, x_2, \rho)$ is cdf of the bivariate normal $N(0, [1, \rho; \rho, 1])$, m_j is the mean divided by the standard deviation of $\beta_0 \cos\tau_j + \beta_1 \sin\tau_j$, for $j = 1, 2$, and ρ is the correlation coefficient between $\beta_0 \cos\tau_1 + \beta_1 \sin\tau_1$ and $\beta_0 \cos\tau_2 + \beta_1 \sin\tau_2$.

To generate the data, we let

$$Y_2 = \delta_1 X + \delta_2 U_0 + \delta_3 U_1 + \delta_4 V, \tag{8.8}$$

where $X \sim N(0, 1)$ is independent of (U_0, U_1) and $V \sim N(0, 1)$ is independent of (X, U_0, U_1).

understanding, and is *not* needed in practice. Thus, we do not need to know the form of the mapping $B(\cdot)$. This is important because its form is typically complicated. Mathematically, each column of B is the polar of a facet of the convex (pointed) polyhedral cone spanned by $c_1, ..., c_m$. Algebraic representations of facets of convex polyhedral cones are complicated.

Let $\theta = (0, -1, -1, 1)'$, and $\delta := (\delta_1, \delta_2, \delta_3, \delta_4)' = (1, 0.577, -0.577, 0.577)'$.[11]

We compute the probabilities that the CS's for θ cover given values of θ. For the given values of θ, we consider $\theta = (0, \alpha_1, -1, 1)'$, where α_1 runs from -1 to 1.4. Note that $(0, -1, 1)$ is the true value of $(\alpha_0, \gamma_0, \gamma_1)$, and -1 is the true value of α_1. Thus, $(0, -1, -1, 1)$ is in the identified set, and should ideally be covered by the CS's with at least the nominal coverage probability. Numerical calculation of the boundary of the identified set shows that $(0, \alpha_1, -1, 1)$ is outside the identified set for any $\alpha_1 > -0.8274$ and, hence, it is desirable that CS's cover such α_1 values with low probabilities.[12]

We consider CS's based on the CvM and KS statistics and the GMS and sub-sampling critical values. For comparative purposes, we also consider the two-step CCK-EB and two-step CCK-SN (self-normalizing) based CS's.[13] For all CS's, we choose $r_{1,n} = 3$ and approximate $\mathcal{T}$ by grid points.[14] For the GMS CS's, we set η to zero. For the sub-sampling CS's, we set the subsample size to 20. For the CCK-EB CS, we take the tuning parameter values from CCK's Monte Carlo simulations. We use 1001 repetitions to simulate the bootstrap critical values and we use 1001 subsamples to construct the sub-sampling critical values. We employ 1000 Monte Carlo repetitions to obtain the simulated coverage probabilities of given points of θ.

Figure 2 provides coverage probability graphs for sample sizes n = 250, 500, 1000, and

[11]This value of δ is the weak-identification specification in CR. Since identification strength is irrelevant for evaluating the property of the MCMI tests, we focus on this weak-identification specification and do not consider other specifications.

[12]Specifically, the way we compute the boundary is as follows. First we construct the criterion function $Q(\theta) = \min_{x \in \mathcal{X}_{N_x}} \min_{\tau \in \mathcal{T}_{N_\tau}} F_\beta(\mathcal{S}(\tau), \theta) - E[1\{S(Y_1, Y_2) \subseteq \mathcal{S}_{\tau_1, \tau_2}\}|X = x]$, where $\mathcal{X}_{N_x}$ is the set of $N_x = 20$ equally-spaced grid points in the interval $[-4, 4]$, $\mathcal{T}_{N_\tau}$ is the approximation of $\mathcal{T}$ described in the next footnote, $F_\beta(\mathcal{S}_\tau, \theta)$ is computed using the bivariate-normal cdf function in Aptech Gauss, and $E[1\{S(Y_1, Y_2) \subseteq \mathcal{S}_{\tau_1, \tau_2}\}|X = x]$ is computed using i.i.d. Monte Carlo simulations with 10^7 simulation repetitions. Then we fix $\alpha_0, \gamma_0, \gamma_1$ at their true values, and search for $a_1 > -1$ that makes $Q(\alpha_0, a_1, \gamma_0, \gamma_1)$ zero. The function $Q(\alpha_0, \cdot, \gamma_0, \gamma_1)$ appears to be monotonically decreasing in the range $[-1, 2]$ and changes signs from one end point to the other.

[13]The two-step CCK methods perform better than the one-step CCK methods in this example. The performance of the CCK-MB method lies in between that of the CCK-EB and the CCK-SN. We do not consider the three-step CCK methods for two reasons. First, those methods require the derivative of $F_\beta(\mathcal{S}, \theta)$ with respect to θ, the analytical form of which is complicated because $F_\beta(\mathcal{S}, \theta)$ is a quadrant probability of a bivariate normal with both the mean and the variance-covariance matrix dependent on θ. Second, the potential gain of using the three-step CCK methods is likely small because, for every $\mathcal{S}$, we expect $F_\beta(\mathcal{S}, \theta)$ to depend strongly on the mean and the variance of the bivariate normal, and hence on θ. CCK also do not provide any simulation results for their three-step methods.

[14]We consider N_{τ_2} equally-spaced grid points for τ_2 in $[0, 2\pi]$, and grid points for τ_1 in $[0, \tau_2]$ with the same spacing. We set $N_{\tau_2} = 15$ for our CS's, which results in 120 points in $\{(\tau_1, \tau_2) \in [0, 2\pi] : \tau_1 \leq \tau_2\}$. We set $N_{\tau_2} = 14$ for the CCK-EB CS, because when $N_{\tau_2} = 15$, some of the moments have very small variance, which causes the CCK-EB CS to have a zero coverage probability for the true value.

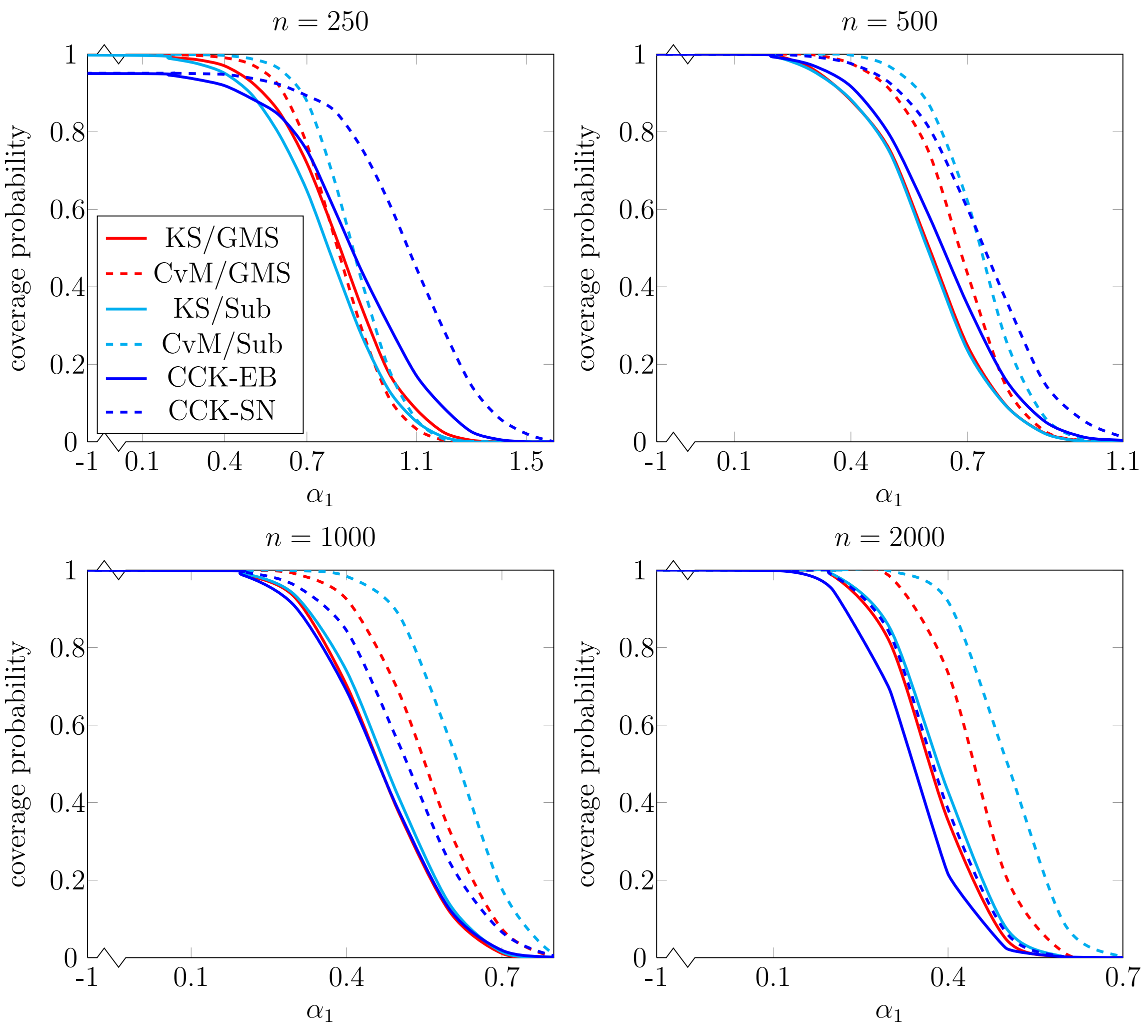

Figure 2: Coverage Probabilities in the IV Random-Coefficients Binary-Outcome Model. (Nominal size = .95, $(\alpha_0, \gamma_0, \gamma_1) = (0, -1, 1)$, and $(\alpha_0, \alpha_1, \gamma_0, \gamma_1)$ is in the identified set if and only if $\alpha_1 \leq -0.8274$.)

2000. As the figure shows, the coverage probabilities of the CS's equal one at the boundary ($\alpha_1 = -0.8274$) of the identified set for all of the CS's except the CCK-EB CS for the case of $n = 250$. This is probably due to the fact that the boundary of the identified set is determined by X values in a set with Lebesgue measure zero. For $n = 250$, the coverage probability of the CCK-EB CS is closer to the nominal size 0.95 at the boundary of the identified set than the other CS's, but its coverage probabilities decrease more slowly than those of the other CS's as α_1 deviates from the identified set (i.e., as α_1 increases beyond -0.8274).

The coverage probabilities of all of the CS's for points outside the identified set decrease with the sample size (with the exception of the CCK CS's for points close to the identified set) and with the magnitude of the deviation from the identified set, as expected. The best performing CS's (lowest curves) are the KS/GMS and KS/sub-sampling CS's at $n = 500$, where the coverage probability curves of the KS/GMS and KS/Sub-sampling CS's overlap completely and form the lowest curve in the graph. At $n = 1000$, the KS/GMS and the CCK-EB curves overlap and form the lowest curve in the graph. At $n = 2000$, the CCK-EB performs better than the other CS's. For each of the four sample sizes considered, the CvM-based CS's do not perform as well as the other CS's.

9 Examples 3-5

In this section, we verify the high-level assumptions for Examples 3, 4, and 5.

9.1 Example 3: Convex Moment Prediction Models–Support Function Approach

As mentioned above, Beresteanu, Molchanov, and Molinari (2010) verify a version of the high-level conditions given in an earlier version of our paper for the best linear predictor and entry-game applications of this example. In this subsection, we verify our current high-level conditions for the general BMM framework in (2.12).

We focus on the moment inequality model in (2.12) because it includes the case where $Q_\theta(W, V) = Q_\theta(W)$ as a special case. For this model, $p = k = 1$. For simplicity, we take $\widehat{\sigma}_{n,1}(\theta) = \sigma_{F,1}(\theta) = 1$ for all (θ, F) and all n, and hence Assumptions SIG1 and SIG2 hold. Alternatively, one could choose $\sigma_{F,1}(\theta)$ and $\widehat{\sigma}_{n,1}(\theta)$ that are scale equi-variant in the spirit of those in Section 7.1.2.

Lemma 9.1 *For the model in* (2.12)*, let* $\mathcal{F}_+$ *be the set of* (θ, F) *such that* (i) $\theta \in \Theta$, (ii) $\{W_i : i \geq 1\}$ *are i.i.d. under* F, (iii) $Q_\theta(w, v) \subseteq \{q \in R^d : \|q\| \leq M(w)/2\}$ *for some measurable function* $M(w)$ $\forall (w, v) \in \mathcal{WV}$, (iv) $\|q(x)\| \leq M(w)/2$ $\forall x \in \mathcal{X}$, $\forall w \in \mathcal{W}$, *and* (v) $E_F[M(W)^{2+\delta}] \leq C_1$ *for some* $\delta > 0$ *and* $C_1 < \infty$. *Then,* $\mathcal{F}_+$ *satisfies Assumptions* PS1 *and* PS2 *with* $M(w)$, δ, *and* C_1 *as defined immediately above.*

The verification of Assumption PS2 in this case relies on a direct calculation of the covering numbers using the Lipschitz continuity of the moment function with respect to the index u.

9.2 Examples 4 and 5: Countable Conditional Moment Inequalities

In this subsection, we verify the high-level assumptions for models with countably many conditional moment inequalities. Examples 4 and 5 are of this type.

Suppose that the identification theory implies the following moment inequality model:

$$E_{F_0}[\widetilde{m}(W, \theta, \tau)|X] \geq 0, \text{ for } \tau = 1, 2, 3, \ldots, \tag{9.1}$$

where $\widetilde{m}(W, \theta, \tau)$ is a real-valued moment function. For example, these moment conditions could be the ones in (2.16) or (2.18).

In general, the raw moment functions $\widetilde{m}(W, \theta, \tau)$ may not satisfy Assumption PS2. Thus, we rescale them with weights that decrease with τ. Let $w_{\mathcal{T}}(\tau) : [1, \infty) \to (0, 1]$ be a strictly decreasing, positive, weight function with inverse function $\lambda_{\mathcal{T}}(\xi) : (0, 1] \to [1, \infty)$ that satisfies $\int_0^1 \sqrt{\log(\lambda_{\mathcal{T}}(\xi))} d\xi < \infty$. Then, we let

$$m(W, \theta, \tau) = w_{\mathcal{T}}(\tau)\widetilde{m}(W, \theta, \tau) \ \forall \tau = 1, 2, \ldots \ . \tag{9.2}$$

In consequence, the moment inequality model (9.1) is equivalent to

$$E_{F_0}[m(W, \theta, \tau)|X] \geq 0 \ \forall \tau = 1, 2, \ldots \ . \tag{9.3}$$

We verify the high-level assumptions given above for this rescaled form of the moment inequalities.

For this model, $p = k = 1$, and we use $\sigma^2_{F,1}(\theta) = Var_F(m(W, \theta, 1))$ and $\widehat{\sigma}^2_{n,1}(\theta) = n^{-1}\sum_{i=1}^n [m(W_i, \theta, 1) - \overline{m}_n(\theta, 1)]^2$, where $\overline{m}_n(\theta, 1) = n^{-1}\sum_{i=1}^n m(W_i, \theta, 1)$.

Lemma 9.2 *For the model in* (9.3), *let* $\mathcal{F}_+$ *be the set of* (θ, F) *such that* (i) $\theta \in \Theta$, (ii) $\{W_i : i \geq 1\}$ *are i.i.d. under* F, (iii) $\sigma^2_{F,1}(\theta) \geq \underline{\sigma}^2$ *for some constant* $\underline{\sigma}^2 > 0$, (iv) $|\widetilde{m}(w, \theta, \tau)| \leq B(w)$ $\forall w \in \mathcal{W}$, $\forall \tau \in \mathcal{T}$, $\forall \tau \in \Theta$, *for some measurable function* $B(w)$, *and* (v) $E[(B(W)/\underline{\sigma})^{2+\delta}] \leq C_1$ *for some* $\delta > 0$ *and* $C_1 < \infty$. *Let* $w_{\mathcal{T}}(\tau)$ *be a weight function that satisfies the definition above. Then,*

(a) $\mathcal{F}_+$ *satisfies Assumptions* PS1 *and* PS2 *with* $M(w) = B(w)/\underline{\sigma}$ *and with* C_1 *and* δ *defined immediately above, and*

(b) *Assumptions* SIG1 *and* SIG2 *hold.*

The verification of Assumption PS2 in this case relies on a direct calculation of the covering numbers. The covering numbers are properly bounded due to the decreasing weight $w_{\mathcal{T}}(\tau)$.

We note that the weighting scheme requires an ordering of the moment conditions. A natural ordering of the moment conditions is often available. For example, in Example 4, suppose that M (the number of values that the dependent variable Y can take) is small but the number of values that Z can take is large, one natural order of the moment conditions is according to the empirical probability $Z = z_s$, while treating moment conditions with the same s but different ℓ, m (indices for the value of Y) as ties in the ordering. In Example 5, one can order the actions according to how close they are to the optimal (observed) action. A similar ordering may be used for the dynamic model of imperfect competition in Example 3 of CCK.

When there are no ties in the ordering, an example of the weight is $w_{\mathcal{T}}(\tau) = \tau^{-b}$ for some $b > 0$. Then $\lambda_{\mathcal{T}}(\xi) = \xi^{-1/b}$ and

$$\int_0^1 \sqrt{\log(\xi^{-1/b})}d\xi = \sqrt{1/b}\int_0^1 \sqrt{\log(\xi^{-1})}d\xi = b^{-1/2}\int_0^\infty 2x^2 e^{-x^2} dx < \infty, \tag{9.4}$$

where the last equality holds by change of variables with $x = \sqrt{\log(\xi^{-1})}$ (or, equivalently, $\xi = e^{-x^2}$). When there are ties, one can consider the tied moment conditions as one, assign the decreasing weights as just described, and give equal weights to the tied observations.

10 Conclusion

In this paper, we construct confidence sets for models defined by many conditional moment inequalities/equalities. The conditional moment restrictions in the models can be finite, countably infinite, or uncountably infinite. To deal with the complication brought about by the vast number of moment restrictions, we exploit the manageability (Pollard (1990)) of the class of moment functions. We verify the manageability condition in five examples from the recent partial-identification literature.

The proposed confidence sets are constructed by inverting joint tests that employ all of the moment restrictions. The confidence sets are shown to have correct asymptotic size in a uniform sense and to exclude parameter values outside the identified set with probability approaching one. Monte Carlo experiments for a conditional stochastic dominance example and a random-coefficients binary-outcome example support the theoretical results.

References

Anderson, G., 1996. Nonparametric tests of stochastic dominance in income distributions. Econometrica 64, 1183–1193.

Anderson, G., 2004. Toward an empirical analysis of polarization. Journal of Econometrics 122, 1–26.

Andrews, D. W. K., 1988. Chi-Square Diagnostic Tests for Econometric Models: Theory. Econometrica 56, 1419–1453.

Andrews, D. W. K., Shi, X., 2009. Inference based on conditional moment inequalities. Cowles Foundation, Yale University. Unpublished manuscript.

Andrews, D. W. K., Shi, X., 2013a. Inference based on conditional moment inequalities. Econometrica 81, 609–666.

Andrews, D. W. K., Shi, X., 2013b. Supplemental material to 'Inference based on conditional moment inequalities.' Available at Econometrica Supplemental Material 81, http://www.econometricsociety.org/ecta/supmat/9370_Proofs.pdf.

Andrews, D. W. K., Shi, X., 2014. Nonparametric inference based on conditional moment inequalities. Journal of Econometrics 179, 31–45. Appendix of proofs available as Cowles Foundation Discussion Paper No. 1840RR, Yale University, 2011.

Andrews, D. W. K., G. Soares, 2010. Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection. Econometrica 78, 119–157.

Armstrong, T. B., 2014a. On the choice of test statistic for conditional moment inequalities. Cowles Foundation Discussion Paper No. 1960, Yale University.

Armstrong, T. B., 2014b. Weighted KS statistics for inference on conditional moment inequalities. Journal of Econometrics 181, 92–116.

Armstrong, T. B., 2015. Asymptotically exact inference in conditional moment inequality models. Journal of Econometrics 186, 51–85.

Armstrong, T. B., Chan, H. P., 2013. Multiscale adaptive inference on conditional moment inequalities. Cowles Foundation Discussion Paper No. 1885R, Yale University.

Barrett, G. F., Donald, S. G., 2003. Consistent tests for stochastic dominance. Econometrica 71, 71–104.

Barrett, G. F., Donald, S. G., Bhattacharya, D., 2014. Consistent nonparametric tests for Lorenz dominance. Journal of Business and Economic Statistics 32, 1–13.

Beare, B. K., Moon, J.-M., 2015. Nonparametric tests of density ratio ordering. Econometric Theory 31, 471–492.

Beare, B. K., Shi, X., 2015. An improved bootstrap test of density ratio ordering. Department of Economics, University of Wisconsin. Unpublished manuscript.

Beresteanu, A., Molchanov, I., Molinari, F., 2010. Sharp identification regions in models with convex predictions. Econometrica 79, 1785–1821.

Bishop, J. A., Zeager, L. A., Zheng, B., 2011. Interdistributional inequality, stochastic dominance, and poverty. Department of Economics, East Carolina University. Unpublished manuscript.

Chernozhukov, V., Chetverikov, D., Kato, K., 2013. Comparison and Anti-concentration Bounds for Maxima of Gaussian Random Vectors. Department of Economics, MIT. Unpublished manuscript.

Chernozhukov, V., Chetverikov, D., Kato, K., 2014a. Testing many moment inequalities. Department of Economics, MIT. Unpublished manuscript.

Chernozhukov, V., Chetverikov, D., Kato, K., 2014b. Anti-Concentration and Honest Adaptive Confidence Bands. The Annals of Statistics 42, 1787-1818.

Chernozhukov, V., Lee, S., Rosen, A., 2013. Intersection bounds: estimation and inference. Econometrica 81, 667–737.

Chesher, A., Rosen, A. M., 2014. An instrumental variable random-coefficients model for binary outcomes. Econometrics Journal 17, S1–S19.

Chesher, A., Smoliński, K., 2012. IV models of ordered choice. Journal of Econometrics 166, 33–48.

Chetverikov, D., 2012. Adaptive test of conditional moment inequalities. Department of Economics, UCLA. Unpublished manuscript.

Dardanoni, V., Forcina, A., 1999. Inference for Lorenz curve orderings. Econometrics Journal 2, 49–75.

Davidson, R., Duclos, J.-Y., 2000. Statistical inference for stochastic dominance and for the measurement of poverty and inequality. Econometrica 68, 1435–1464.

Delgado, M. A., Escanciano, J. C., 2013. Conditional stochastic dominance testing. Journal of Business and Economic Statistics 31, 16–28.

Galichon, A., Henry, M., 2009. A test of non-identifying restrictions and confidence regions for partially identified parameters. Journal of Econometrics 152, 186–196.

Gandhi, A., Lu, Z., Shi, X., 2013. Estimating demand for differentiated products with error in market shares. Department of Economics, University of Wisconsin. Unpublished manuscript.

Guerre, E., Perrigne, I., Vuong, Q., 2009. Nonparametric identification of risk aversion in first-price auctions under exclusion restrictions. Econometrica 77, 1193–1227.

Kahn, S., Tamer, E., 2009. Inference on endogenously censored regression models using conditional moment inequalities. Journal of Econometrics 152, 104–119.

Lee, S., Song, K., Whang, Y.-J., 2013. Testing functional inequalities. Journal of Econometrics 172, 14–32.

Linton, O., Song, K., Whang, Y.-J., 2010. An improved bootstrap test of stochastic dominance. Journal of Econometrics 154, 186–202.

Pakes, A., Porter, J., Ho, K., Ishii, J., 2015. Moment inequalities and their application. Econometrica 83, 315–334.

Pollard, D., 1979. General Chi-Square Goodness-of-Fit Tests with Data-Dependent Cells. Z. Wahrscheinlichkeitstheorie 50, 317–331.

Pollard, D., 1990. Empirical Process Theory and Application, NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. II. Institute of Mathematical Statistics.

This Appendix is not to be published. It will be made available on the web.

Appendix

to

Inference Based on Many Conditional Moment Inequalities

Donald W. K. Andrews
Cowles Foundation for Research in Economics
Yale University

Xiaoxia Shi
Department of Economics
University of Wisconsin, Madison

June 2010
Revised: April 26, 2016

A Outline

This Appendix provides proofs of Theorems 5.1 and 6.1 of Andrews and Shi (2010) "Inference Based on Many Conditional Moment Inequalities," referred to hereafter as ASM. In fact, the results given here cover a much broader class of test statistics than is considered in ASM. We let AS1 abbreviate Andrews and Shi (2013a) and AS2 abbreviate Andrews and Shi (2013b).

This Appendix is organized as follows. Section B defines the class of test statistics that are considered. This class includes the statistics that are considered in ASM. Section B also provides the definition of manageability that is used in Assumption PS2. Section C introduces the critical values, the confidence sets (CS's), and the tests. Section D establishes the correct asymptotic size of the CS's. Theorem 5.1 of ASM is a corollary to Lemmas D.1 and D.2, which are given in Section D. Section E establishes that the CS's contain fixed parameter values outside the identified set with probability that goes to zero. Equivalently, the tests upon which the CS's are constructed are shown to be consistent tests. Theorem 6.1 of ASM is a corollary to Theorem E.1, which is given in Section E. Section F provides proofs of Lemma 7.1-8.2 of ASM, which verify Assumptions PS1, PS2, SIG1, and SIG2 in the examples given in ASM. Section G provides additional Monte Carlo simulation results for the two simulation examples considered in ASM. These results are designed to analyze the robustness of the tests and CS's to the tuning parameters that are used.

B General Form of the Test Statistic

B.1 Test Statistic

Here we define the general form of the test statistic $T_n(\theta)$ that is used to construct a CS. We transform the conditional moment inequalities/equalities given X_i into equivalent unconditional moment inequalities/equalities by choosing appropriate weighting functions of X_i, i.e., X_i instruments. Then, we construct a test statistic based on the instrumented moment conditions.

The instrumented moment conditions are of the form:

$$E_{F_0}[m_j(W_i,\theta_0,\tau)\,g_j(X_i)] \geq 0 \text{ for } j=1,...,p \text{ and} \tag{B.1}$$
$$E_{F_0}[m_j(W_i,\theta_0,\tau)\,g_j(X_i)] = 0 \text{ for } j=p+1,...,k, \text{ for } g=(g_1,...,g_k)' \in \mathcal{G} \text{ and } \tau \in \mathcal{T},$$

where θ_0 and F_0 are the true parameter and distribution, respectively, g is the instrument vector that depends on the conditioning variables X_i, and $\mathcal{G}$ is a collection of instruments. Typically $\mathcal{G}$ contains an infinite number of elements.

The identified set $\Theta_{F_0}(\mathcal{G})$ of the model defined by (B.1) is

$$\Theta_{F_0}(\mathcal{G}) := \{\theta \in \Theta : \text{(B.1) holds with } \theta \text{ in place of } \theta_0\}. \tag{B.2}$$

The collection $\mathcal{G}$ is chosen so that $\Theta_{F_0}(\mathcal{G}) = \Theta_{F_0}$, where Θ_{F_0} is the identified set based on the conditional moment inequalities and equalities defined in (2.2) of ASM. Section B.4 provides conditions for this equality and shows that the instruments defined in (3.6) of ASM satisfy the conditions. Additional sets $\mathcal{G}$ are given in AS1 and AS2.

We construct test statistics based on (B.1). The sample moment functions are defined in (3.2) in ASM. The sample variance-covariance matrix of $n^{1/2}\overline{m}_n(\theta,\tau,g)$ is defined in (3.3) in ASM. The matrix $\widehat{\Sigma}_n(\theta,\tau,g)$ may be singular with non-negligible probability for some $g \in \mathcal{G}$. This is undesirable because the inverse of $\widehat{\Sigma}_n(\theta,\tau,g)$ needs to be consistent for its population counterpart uniformly over $g \in \mathcal{G}$ for the test statistics considered below. Thus, we employ a modification of $\widehat{\Sigma}_n(\theta,\tau,g)$, denoted by $\overline{\Sigma}_n(\theta,\tau,g)$ and defined in (3.4) in ASM, such that the smallest eigenvalue of $\overline{\Sigma}_n(\theta,\tau,g)$ is bounded away from zero.

The test statistic $T_n(\theta)$ is either a Cramér-von-Mises-type (CvM) or a Kolmogorov-Smirnov-type (KS) statistic. The CvM statistic is

$$T_n(\theta) := \sup_{\tau \in \mathcal{T}} \int_{\mathcal{G}} S(n^{1/2}\overline{m}_n(\theta,\tau,g), \overline{\Sigma}_n(\theta,\tau,g))dQ(g), \tag{B.3}$$

where S is a non-negative function and Q is a weight function (i.e., probability measure) on $\mathcal{G}$. The functions S and Q are discussed in Sections B.2 and B.5 below, respectively.

The Kolmogorov-Smirnov-type (KS) statistic is

$$T_n(\theta) := \sup_{\tau \in \mathcal{T}} \sup_{g \in \mathcal{G}} S(n^{1/2}\overline{m}_n(\theta,\tau,g), \overline{\Sigma}_n(\theta,\tau,g)). \tag{B.4}$$

For brevity, the discussion in this Appendix focusses on CvM statistics and all results stated concern CvM statistics. Similar results hold for KS statistics. Such results can be established by extending the results given in Section 13.1 of Appendix B of AS2 and proved in Section 15.1 of Appendix D of AS2.

B.2 S Function Assumptions

Let $m_I := (m_1, ..., m_p)'$ and $m_{II} := (m_{p+1}, ..., m_k)'$. Let Δ be the set of $k \times k$ positive-definite diagonal matrices. Let $\mathcal{W}$ be the set of $k \times k$ positive definite matrices.

Assumption S1. $\forall (m, \Sigma) \in \{(m, \Sigma) : m \in (-\infty, \infty]^p \times R^v, \Sigma \in \mathcal{W}\}$,

(a) $S(Dm, D\Sigma D) = S(m, \Sigma)\ \forall D \in \Delta$,

(b) $S(m_I, m_{II}, \Sigma)$ is non-increasing in each element of m_I,

(c) $S(m, \Sigma) \geq 0$,

(d) S is continuous, and

(e) $S(m, \Sigma + \Sigma_1) \leq S(m, \Sigma)$ for all $k \times k$ positive semi-definite matrices Σ_1.

It is worth pointing out that Assumption S1(d) requires S to be continuous in m at all points m in the extended vector space $(-\infty, \infty]^p \times R^v$, not only for points in R^{p+v}.

Let $\mathcal{M}$ denote a bounded subset of R^k. Let $\mathcal{W}_{cpt}$ denote a compact subset of $\mathcal{W}$.

Assumption S2. $S(m, \Sigma)$ is uniformly continuous in the sense that

$$\lim_{\delta \downarrow 0} \sup_{\mu \in R_+^p \times \{0\}^v} \sup_{\substack{m, m^* \in \mathcal{M} \\ \|m - m^*\| \leq \delta}} \sup_{\substack{\Sigma, \Sigma^* \in \mathcal{W}_{cpt} \\ \|\Sigma - \Sigma^*\| \leq \delta}} |S(m + \mu, \Sigma) - S(m^* + \mu, \Sigma^*)| = 0.$$ [15]

Assumption S3. $S(m, \Sigma) > 0$ if and only if $m_j < 0$ for some $j = 1, ..., k$, where $m = (m_1, ..., m_k)'$ and $\Sigma \in \mathcal{W}$.

Assumption S4. For some $\chi > 0$, $S(am, \Sigma) = a^\chi S(m, \Sigma)$ for all scalar $a > 0$, $m \in R^k$, and $\Sigma \in \mathcal{W}$.

[15] It is important that the supremum is only over μ vectors with non-negative elements μ_j for $j \leq p$. Without this restriction on the μ vectors, Assumption S2 would not hold for typical S functions of interest. Also note that Assumption S2 here is Assumption S2$'$, rather than Assumption S2, in AS1. Although Assumption S2 in AS1 is seemingly weaker than Assumption S2$'$, the former implies the latter, i.e. the two assumptions are equivalent. The equivalence can be established by adapting the proof of the well-known result that continuous functions defined on compact sets are uniformly continuous.

It is shown in Lemma 1 of AS1 that the functions S_1-S_3 in (3.9) satisfy Assumptions S1-S4. The function S_4 also does by similar arguments.

B.3 Definition of Manageability

Here we introduce the concept of manageability from Pollard (1990) that is used in Assumption PS2 in ASM and Assumption M that is introduced in the following section. This condition is used to regulate the complexity of $\mathcal{T} \times \mathcal{G}$. It ensures that $\{n^{1/2}(\overline{m}_n(\theta, \tau, g) - E_{F_n}\overline{m}_n(\theta, \tau, g)) : (\tau, g) \in \mathcal{T} \times \mathcal{G}\}$ satisfies a functional central limit theorem (FCLT) under drifting sequences of distributions $\{F_n : n \geq 1\}$. The latter is utilized in the proof of the uniform coverage probability results for the CS's. See Pollard (1990) and Appendix E of AS2 for more about manageability.

Definition (Pollard, 1990, Definition 3.3). The *packing number* $D(\xi, \rho, V)$ for a subset V of a metric space $(\mathcal{V}, \rho)$ is defined as the largest b for which there exist points $v^{(1)}, ..., v^{(b)}$ in V such that $\rho(v^{(s)}, v^{(s')}) > \xi$ for all $s \neq s'$. The *covering number* $N(\xi, \rho, V)$ is defined to be the smallest number of closed balls with ρ-radius ξ whose union covers V.

It is easy to see that $N(\xi, \rho, V) \leq D(\xi, \rho, V) \leq N(\xi/2, \rho, V)$.

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be the underlying probability space equipped with probability distribution $\mathbf{P}$. Let $\{f_{n,i}(\cdot, \tau) : \Omega \to R : \tau \in \mathcal{T}, i \leq n, n \geq 1\}$ be a triangular array of random processes. Let

$$\mathcal{F}_{n,\omega} := \{(f_{n,1}(\omega, \tau), ..., f_{n,n}(\omega, \tau))' : \tau \in \mathcal{T}\}. \tag{B.5}$$

Because $\mathcal{F}_{n,\omega} \subset R^n$, we use the Euclidean metric $\| \cdot \|$ on this space. For simplicity, we omit the metric argument in the packing number function, i.e., we write $D(\xi, V)$ in place of $D(\xi, \| \cdot \|, V)$ when $V \subset \mathcal{F}_{n,\omega}$.

Let $\odot$ denote the element-by-element product. For example for $a, b \in R^n$, $a \odot b = (a_1 b_1, ..., a_n b_n)'$. Let *envelope functions* of a triangular array of processes $\{f_{n,i}(\omega, \tau) : \tau \in \mathcal{T}, i \leq n, n \geq 1\}$ be an array of functions $\{F_n(\omega) = (F_{n,1}(\omega), ..., F_{n,n}(\omega))' : n \geq 1\}$ such that $|f_{n,i}(\omega, \tau)| \leq F_{n,i}(\omega)$ $\forall i \leq n, n \geq 1, \tau \in \mathcal{T}, \omega \in \Omega$.

Definition (Pollard, 1990, Definition 7.9). A triangular array of processes $\{f_{n,i}(\omega, \tau) : \tau \in \mathcal{T}, i \leq n, n \geq 1\}$ is said to be *manageable* with respect to the envelopes $\{F_n(\omega) : n \geq 1\}$ if there exists a deterministic real function λ on $(0, 1]$ for which (i) $\int_0^1 \sqrt{\log \lambda(\xi)} d\xi < \infty$ and

(ii) $D(\xi\|\alpha\odot F_n(\omega)\|,\alpha\odot\mathcal{F}_{n,\omega})\leq\lambda(\xi)$ for $0<\xi\leq 1$, all $\omega\in\Omega$, all n-vectors α of nonnegative weights, and all $n\geq 1$.

B.4 X Instruments

The collection of instruments $\mathcal{G}$ needs to satisfy the following condition in order for the unconditional moments $\{E_F[m(W_i,\theta,\tau,g)]:(\tau,g)\in\mathcal{T}\times\mathcal{G}\}$ to incorporate the same information as the conditional moments $\{E_F[m(W_i,\theta,\tau)|X_i=x]:x\in R^{d_x}\}$.

For any $\theta\in\Theta$ and any distribution F with $E_F[\|m(W_i,\theta,\tau)\|]<\infty$, $\forall\tau\in\mathcal{T}$, let $\mathcal{X}_F(\theta,\tau)$ be defined as in (6.2) in ASM.

Assumption CI. For any $\theta\in\Theta$ and distribution F for which $E_F[\|m(W_i,\theta,\tau)\|]<\infty$, $\forall\tau\in\mathcal{T}$, if $P_F(X_i\in\mathcal{X}_F(\theta,\tau_*))>0$ for some $\tau_*\in\mathcal{T}$, then there exists some $g\in\mathcal{G}$ such that

$$E_F[m_j(W_i,\theta,\tau_*)g_j(X_i)]<0 \text{ for some } j\leq p \text{ or}$$
$$E_F[m_j(W_i,\theta,\tau_*)g_j(X_i)]\neq 0 \text{ for some } j>p.$$

Note that CI abbreviates "conditionally identified." The following Lemma indicates the importance of Assumption CI. The proof of the lemma is the same as the proof of Lemma 2 in AS1, which is given in AS2, and in consequence, is omitted.

Lemma B.1 *Assumption* CI *implies that* $\Theta_F(\mathcal{G})=\Theta_F$ *for all* F *with* $\sup_{\theta\in\Theta}E_F[\|m(W_i,\theta,\tau)\|]<\infty$.

Collections $\mathcal{G}$ that satisfy Assumption CI contain non-negative functions whose supports are cubes, boxes, or other sets which are arbitrarily small.

The collection $\mathcal{G}$ also must satisfy the following "manageability" condition.

Assumption M. (a) $0\leq g_j(x)\leq G$ $\forall x\in R^{d_x},\forall j\leq k,\forall g\in\mathcal{G}$, for some constant $G<\infty$, and

(b) the processes $\{g_j(X_{n,i}):g\in\mathcal{G},\ i\leq n,n\geq 1\}$ are manageable with respect to the constant function G for $j=1,...,k$, where $\{X_{n,i}:i\leq n,n\geq 1\}$ is a row-wise i.i.d. triangular array with $X_{n,i}\sim F_{X,n}$ and $F_{X,n}$ is the distribution of $X_{n,i}$ under F_n for some $(\theta_n,F_n)\in\mathcal{F}_+$ for $n\geq 1$.[16]

[16] The asymptotic results given in the paper hold with Assumption M replaced by any alternative assumption that is sufficient to obtain the requisite empirical process results given in Lemma D.2 below.

Lemma 3 of AS1 establishes Assumptions CI and M for $\mathcal{G}_{\text{c-cube}}$ defined in (3.6) of ASM.[17]

B.5 Weight Function Q

The weight function Q can be any probability measure on $\mathcal{G}$ whose support is $\mathcal{G}$. This support condition is needed to ensure that no functions $g \in \mathcal{G}$, which might have set-identifying power, are "ignored" by the test statistic $T_n(\theta)$. Without such a condition, a CS based on $T_n(\theta)$ would not necessarily shrink to the identified set as $n \to \infty$. Section E below introduces the support condition formally and shows that the probability measure Q considered here satisfies it.

We now give an example of a weight function Q on $\mathcal{G}_{\text{c-cube}}$.

Weight Function Q for $\mathcal{G}_{\text{c-cube}}$. There is a one-to-one mapping $\Pi_{\text{c-cube}} : \mathcal{G}_{\text{c-cube}} \to AR := \{(a,r) : a \in \{1,...,2r\}^{d_x} \text{ and } r = r_0, r_0+1, ...\}$. Let Q_{AR} be a probability measure on AR. One can take $Q = \Pi_{\text{c-cube}}^{-1} Q_{AR}$. A natural choice of measure Q_{AR} is uniform on $a \in \{1,...,2r\}^{d_x}$ conditional on r combined with a distribution for r that has some probability mass function $\{w(r) : r = r_0, r_0+1, ...\}$. This yields the test statistic

$$\sup_{\tau \in \mathcal{T}} \sum_{r=r_0}^{\infty} w(r) \sum_{a \in \{1,...,2r\}^{d_x}} (2r)^{-d_x} S(n^{1/2}\overline{m}_n(\theta,\tau,g_{a,r}), \overline{\Sigma}_n(\theta,\tau,g_{a,r})), \tag{B.6}$$

where $g_{a,r}(x) := 1(x \in C_{a,r}) \cdot 1_k$ for $C_{a,r} \in \mathcal{C}_{\text{c-cube}}$.

The weight function Q_{AR} with $w(r) := (r^2+100)^{-1}$ is used in the test statistics in ASM, see (3.7).

B.6 Computation of Sums, Integrals, and Suprema

The test statistic $T_n(\theta)$ given in (B.6) involves an infinite sum. A collection $\mathcal{G}$ with an uncountable number of functions g yields a test statistic $T_n(\theta)$ that is an integral with respect to Q. This infinite sum or integral can be approximated by truncation, simulation, or quasi-Monte Carlo (QMC) methods. If $\mathcal{G}$ is countable, let $\{g_1,...,g_{s_n}\}$ denote the first s_n functions g that appear in the infinite sum that defines $T_n(\theta)$. Alternatively, let $\{g_1,...,g_{s_n}\}$ be s_n i.i.d. functions drawn from $\mathcal{G}$ according to the distribution Q. Or, let $\{g_1,...,g_{s_n}\}$ be

[17] Lemma 3 of AS1 and Lemma B2 of AS2 also establish Assumptions CI and M of this Appendix for the collections $\mathcal{G}_{\text{box}}$, $\mathcal{G}_{\text{B-spline}}$, $\mathcal{G}_{\text{box,dd}}$, and $\mathcal{G}_{\text{c/d}}$ defined there.

the first s_n terms in a QMC approximation of the integral with respect to (wrt) Q. Then, an approximate test statistic obtained by truncation, simulation, or QMC methods is

$$\overline{T}_{n,s_n}(\theta) := \sup_{\tau \in \mathcal{T}} \sum_{\ell=1}^{s_n} w_{Q,n}(\ell) S(n^{1/2}\overline{m}_n(\theta,\tau,g_\ell), \overline{\Sigma}_n(\theta,\tau,g_\ell)), \tag{B.7}$$

where $w_{Q,n}(\ell) := Q(\{g_\ell\})$ when an infinite sum is truncated, $w_{Q,n}(\ell) := s_n^{-1}$ when $\{g_1, ..., g_{s_n}\}$ are i.i.d. draws from $\mathcal{G}$ according to Q, and $w_{Q,n}(\ell)$ is a suitable weight when a QMC method is used. For example, in (B.6), the outer sum can be truncated at $r_{1,n}$, in which case, $s_n := \sum_{r=r_0}^{r_{1,n}} (2r)^{d_x}$ and $w_{Q,n}(\ell) := w(r)(2r)^{-d_x}$ for ℓ such that g_ℓ corresponds to $g_{a,r}$ for some a. The test statistics in (3.7) of ASM are of this form when $r_{1,n} < \infty$.[18]

It can be shown that truncation at s_n, simulation based on s_n simulation repetitions, or QMC approximation based on s_n terms, where $s_n \to \infty$ as $n \to \infty$, is sufficient to maintain the asymptotic validity of the tests and CS's as well as the asymptotic power results under fixed alternatives.

The KS form of the test statistic requires the computation of a supremum over $g \in \mathcal{G}$. For computational ease, this can be replaced by a supremum over $g \in \mathcal{G}_n$, where $\mathcal{G}_n \uparrow \mathcal{G}$ as $n \to \infty$, in the test statistic and in the definition of the critical value (defined below). The same asymptotic size results and asymptotic power results under fixed alternatives hold for KS tests with $\mathcal{G}_n$ in place of $\mathcal{G}$. For results of this sort for the tests considered in AS1 and AS2, see Section 13.1 of Appendix B in AS2 and Section 15.1 of Appendix D in AS2.

C GMS Confidence Sets

C.1 Bootstrap GMS Critical Values

It is shown in Theorem D.3 in Section D.3.1 below that when θ is in the identified set the "uniform asymptotic distribution" of $T_n(\theta)$ is the distribution of $T(h_n)$, where $T(h)$ is defined below, $h_n := (h_{1,n}, h_2)$, $h_{1,n}(\cdot)$ is a function from $\mathcal{T} \times \mathcal{G}$ to $R^p_{[+\infty]} \times \{0\}^v$ that depends on the slackness of the moment inequalities and on n, where $R_{[+\infty]} := R \cup \{+\infty\}$, and $h_2(\tau, g, \tau^\dagger, g^\dagger)$ is a $k \times k$ matrix-valued covariance kernel on $(\mathcal{T} \times \mathcal{G})^2$.

[18] Typically, the supremum over τ is obtained through smooth optimization techniques and there is no need to approximate $\mathcal{T}$ by a finite set. However, when smooth optimization is not applicable, we can also approximate $\mathcal{T}$ with a finite subset in the same way as approximating $\mathcal{G}$ by a finite subset.

For $h := (h_1, h_2)$, define

$$T(h) := \sup_{\tau \in \mathcal{T}} \int_{\mathcal{G}} S(\nu_{h_2}(\tau, g) + h_1(\tau, g), h_2^{\varepsilon}(\tau, g)) dQ(g), \tag{C.1}$$

where $h_2^{\varepsilon}(\tau, g) = h_2(\tau, g, \tau, g) + \varepsilon I_k$, and

$$\{\nu_{h_2}(\tau, g) : (\tau, g) \in \mathcal{T} \times \mathcal{G}\} \tag{C.2}$$

is a mean zero R^k-valued Gaussian process with covariance kernel $h_2(\cdot, \cdot)$ on $(\mathcal{T} \times \mathcal{G})^2$, $h_1(\cdot)$ is a function from $\mathcal{T} \times \mathcal{G}$ to $R^p_{[+\infty]} \times \{0\}^v$, and ε is as in the definition of $\overline{\Sigma}_n(\theta, \tau, g)$ in (3.4).[19] The definition of $T(h)$ in (C.1) applies to CvM test statistics. For the KS test statistic, one replaces $\int_{\mathcal{G}} \cdots\ dQ(g)$ by $\sup_{g \in \mathcal{G}} \cdots$.

We are interested in tests of nominal level α and CS's of nominal level $1 - \alpha$. Let

$$c_0(h, 1 - \alpha) \ (= c_0(h_1, h_2, 1 - \alpha)) \tag{C.3}$$

denote the $1 - \alpha$ quantile of $T(h)$. If $h_n := (h_{1,n}, h_2)$ was known, we would use $c_0(h_n, 1 - \alpha)$ as the critical value for the test statistic $T_n(\theta)$. However, h_n is not known and $h_{1,n}$ cannot be consistently estimated. In consequence, we replace h_2 in $c_0(h_{1,n}, h_2, 1 - \alpha)$ by a uniformly consistent estimator $\widehat{h}_{2,n}(\theta)$ $(:= \widehat{h}_{2,n}(\theta, \cdot, \cdot))$ of the covariance kernel h_2 and we replace $h_{1,n}$ by a data-dependent GMS function $\varphi_n(\theta)$ $(:= \varphi_n(\theta, \cdot))$ on $\mathcal{T} \times \mathcal{G}$ (defined in Section C.2 below) that is constructed to be less than or equal to $h_{1,n}(\tau, g)$ for all $(\tau, g) \in \mathcal{T} \times \mathcal{G}$ with probability that goes to one as $n \to \infty$. Because $S(m, \Sigma)$ is non-increasing in m_I by Assumption S1(b), where $m := (m_I', m_{II}')'$ and $m_I \in R^p$, the latter property yields a test with asymptotic level less than or equal to the nominal level α. The quantities $\widehat{h}_{2,n}(\theta)$ and $\varphi_n(\theta)$ are defined below.

Using $\widehat{h}_{2,n}(\theta)$ and $\varphi_n(\theta)$, in principle, one can obtain an approximation of $c_0(h_1, h_2, 1 - \alpha)$ using $c_0(\varphi_n(\theta), \widehat{h}_{2,n}(\theta), 1 - \alpha)$. However, computing $c_0(\varphi_n(\theta), \widehat{h}_{2,n}(\theta), 1 - \alpha)$ in practice is not easy because it involves the simulation of the Gaussian process $\{\nu_{\widehat{h}_2(\theta)}(\tau, g) : \mathcal{T} \times \mathcal{G}\}$. Although we approximate $\mathcal{G}$ by a finite set in ASM, we may not always do so for $\mathcal{T}$. Even when we also use a finite approximation for $\mathcal{T}$, the number of pairs (τ, g) under consideration often is large. That creates difficulty for simulating the Gaussian process.

[19] The sample paths of $\nu_{h_2}(\cdot, \cdot)$ are concentrated on the set $U^k_{\rho_{h_2}}(\mathcal{T} \times \mathcal{G})$ of bounded uniformly ρ_{h_2}-continuous R^k-valued functions on $\mathcal{T} \times \mathcal{G}$, where ρ_{h_2} is the pseudo-metric on $\mathcal{T} \times \mathcal{G}$ defined by $\rho^2_{h_2}(\iota, \iota^{\dagger}) := tr(h_2(\iota, \iota) - h_2(\iota, \iota^{\dagger}) - h_2(\iota^{\dagger}, \iota) + h_2(\iota^{\dagger}, \iota^{\dagger}))$, where $\iota := (\tau, g)$ and $\iota^{\dagger} := (\tau^{\dagger}, g^{\dagger})$.

Thus, we recommend using a bootstrap version of the critical value instead.

The bootstrap GMS critical value is[20]

$$c^*(\varphi_n(\theta), \widehat{h}^*_{2,n}(\theta), 1-\alpha) := c^*_0(\varphi_n(\theta), \widehat{h}^*_{2,n}(\theta), 1-\alpha+\eta) + \eta, \tag{C.4}$$

where $c^*_0(h, 1-\alpha)$ is the $1-\alpha$ conditional quantile of $T^*(h)$ and $T^*(h)$ is defined as in (C.1) but with $\{\nu_{h_2}(\tau, g) : (\tau, g) \in \mathcal{T} \times \mathcal{G}\}$ replaced by the bootstrap empirical process $\{\nu^*_n(\theta, \tau, g) : (\tau, g) \in \mathcal{T} \times \mathcal{G}\}$. The bootstrap empirical process is defined to be

$$\nu^*_n(\theta, \tau, g) := n^{-1/2} \widehat{D}_n(\theta)^{-1/2} \sum_{i=1}^{n} (m(W^*_i, \theta, \tau, g) - \overline{m}_n(\theta, \tau, g)), \tag{C.5}$$

where $\{W^*_i : i \leq n\}$ is an i.i.d. bootstrap sample drawn from the empirical distribution of $\{W_i : i \leq n\}$ and $\widehat{D}_n(\theta)$ is defined in (C.10). The function $\widehat{h}^*_{2,n}(\theta, \tau, g, \tau^\dagger, g^\dagger)$ is defined as

$$\widehat{h}^*_{2,n}(\theta, \tau, g, \tau^\dagger, g^\dagger) = \widehat{D}_n^{-1/2}(\theta) \widehat{\Sigma}^*_n(\theta, \tau, g, \tau^\dagger, g^\dagger) \widehat{D}_n^{-1/2}(\theta), \text{where} \tag{C.6}$$

$\widehat{D}_n(\theta)$ is defined in (C.10) below, and

$$\widehat{\Sigma}^*_n(\theta, \tau, g, \tau^\dagger, g^\dagger) := n^{-1} \sum_{i=1}^{n} (m(W^*_i, \theta, \tau, g) - \overline{m}^*_n(\theta, \tau, g)) \left(m(W^*_i, \theta, \tau^\dagger, g^\dagger) - \overline{m}^*_n(\theta, \tau^\dagger, g^\dagger)\right)', \tag{C.7}$$

and $\overline{m}^*_n(\theta, \tau, g) = n^{-1} \sum_{i=1}^{n} m(W^*_i, \theta, \tau, g)$. Note that we do not recompute $\widehat{D}_n(\theta)$ for the bootstrap samples, which simplifies the theoretical derivations below. Also note that the variance-covariance kernel $\widehat{h}^*_{2,n}(\theta, \tau, g, \tau^\dagger, g^\dagger)$ only enters $c(\varphi_n(\theta), \widehat{h}^*_{2,n}(\theta), 1-\alpha)$ via indices $(\tau, g, \tau^\dagger, g^\dagger)$ such that $(\tau, g) = (\tau^\dagger, g^\dagger)$.

The nominal level $1-\alpha$ GMS CS is given by

$$CS_n := \{\theta \in \Theta : T_n(\theta) \leq c^*_{n,1-\alpha}(\theta)\}, \tag{C.8}$$

[20]The constant η is an *infinitesimal uniformity factor* (IUF) that is employed to circumvent problems that arise due to the presence of the infinite-dimensional nuisance parameter $h_{1,n}$ that affects the distribution of the test statistic in both small and large samples. The IUF obviates the need for complicated and difficult-to-verify uniform continuity and strict monotonicity conditions on the large sample distribution functions of the test statistic.

where the critical value $c^*_{n,1-\alpha}(\theta)$ abbreviates $c^*(\varphi_n(\theta), \widehat{h}^*_{2,n}(\theta), 1-\alpha)$.

When the test statistic, $\overline{T}_{n,s_n}(\theta)$, is a truncated sum, simulated integral, or a QMC quantity, a bootstrap approximate-GMS critical value can be employed. It is defined analogously to the bootstrap GMS critical value but with $T^*(h)$ replaced by $T^*_{s_n}(h)$, where $T^*_{s_n}(h)$ has the same definition as $T^*(h)$ except that a truncated sum, simulated integral, or QMC quantity appears in place of the integral with respect to Q, as in Section B.6. The same functions $\{g_1, ..., g_{s_n}\}$ are used in all bootstrap critical value calculations as in the test statistic $\overline{T}_{n,s_n}(\theta)$.

Next, we define the asymptotic covariance kernel, $\{h_{2,F}(\theta, \tau, g, \tau^\dagger, g^\dagger) : (\tau, g), (\tau^\dagger, g^\dagger) \in \mathcal{T} \times \mathcal{G}\}$, of $n^{1/2}(\overline{m}_n(\theta, \tau, g) - E_F\overline{m}_n(\theta, \tau, g))$ after normalization via a diagonal matrix $D_F^{-1/2}(\theta)$. Define

$$
\begin{aligned}
h_{2,F}(\theta, \tau, g, \tau^\dagger, g^\dagger) &:= D_F^{-1/2}(\theta)\Sigma_F(\theta, \tau, g, \tau^\dagger, g^\dagger)D_F^{-1/2}(\theta), \text{ where} \\
\Sigma_F(\theta, \tau, g, \tau^\dagger, g^\dagger) &:= Cov_F(m(W_i, \theta, \tau, g), m(W_i, \theta, \tau^\dagger, g^\dagger)'), \qquad \text{(C.9)} \\
D_F(\theta) &:= Diag(\sigma^2_{F,1}(\theta), \ldots, \sigma^2_{F,k}(\theta)),
\end{aligned}
$$

and $\sigma^2_{F,j}(\theta)$ is introduced above Assumption PS1.

Correspondingly, the sample covariance kernel $\widehat{h}_{2,n}(\theta)$ $(= \widehat{h}_{2,n}(\theta, \cdot, \cdot))$, which is an estimator of $h_{2,F}(\theta, \tau, g, \tau^\dagger, g^\dagger)$, is defined by

$$
\begin{aligned}
\widehat{h}_{2,n}(\theta, \tau, g, \tau^\dagger, g^\dagger) &:= \widehat{D}_n^{-1/2}(\theta)\widehat{\Sigma}_n(\theta, \tau, g, \tau^\dagger, g^\dagger)\widehat{D}_n^{-1/2}(\theta), \text{ where} \\
\widehat{\Sigma}_n(\theta, \tau, g, \tau^\dagger, g^\dagger) &:= n^{-1}\sum_{i=1}^n (m(W_i, \theta, \tau, g) - \overline{m}_n(\theta, \tau, g)) \left(m(W_i, \theta, \tau^\dagger, g^\dagger) - \overline{m}_n(\theta, \tau^\dagger, g^\dagger)\right)', \\
\widehat{D}_n(\theta) &:= Diag(\widehat{\sigma}^2_{n,1}(\theta), \ldots, \widehat{\sigma}^2_{n,k}(\theta)), \qquad \text{(C.10)}
\end{aligned}
$$

and $\widehat{\sigma}^2_{n,j}(\theta)$ is a consistent estimator of $\sigma^2_{F,j}(\theta)$ introduced below (3.4).

Note that $\widehat{\Sigma}_n(\theta, \tau, g)$, defined in (3.3), equals $\widehat{\Sigma}_n(\theta, \tau, g, \tau, g)$.

C.2 Definition of $\varphi_{\mathbf{n}}(\theta)$

Next, we define $\varphi_n(\theta)$. As discussed above, $\varphi_n(\theta)$ is constructed such that $\varphi_n(\theta, \tau, g) \leq h_{1,n}(\tau, g)$ $\forall(\tau, g) \in \mathcal{T} \times \mathcal{G}$ with probability that goes to one as $n \to \infty$ uniformly over

$(\theta, F) \in \mathcal{F}$. Let

$$\xi_n(\theta, \tau, g) := \kappa_n^{-1} n^{1/2} \overline{D}_n^{-1/2}(\theta, \tau, g) \overline{m}_n(\theta, \tau, g), \text{ where } \overline{D}_n(\theta, \tau, g) := Diag(\overline{\Sigma}_n(\theta, \tau, g)), \tag{C.11}$$

$\overline{\Sigma}_n(\theta, \tau, g)$ is defined in (3.4), and $\{\kappa_n : n \geq 1\}$ is a sequence of constants that diverges to infinity as $n \to \infty$. The jth element of $\xi_n(\theta, \tau, g)$, denoted by $\xi_{n,j}(\theta, \tau, g)$, measures the slackness of the moment inequality $E_F m_j(W_i, \theta, \tau, g) \geq 0$ for $j = 1, ..., p$.

Define $\varphi_n(\theta, \tau, g) := (\varphi_{n,1}(\theta, \tau, g), ..., \varphi_{n,p}(\theta, \tau, g), 0, ..., 0)' \in R^k$ via, for $j \leq p$,

$$\begin{aligned} \varphi_{n,j}(\theta, \tau, g) &:= \overline{h}_{2,n,j}^{1/2}(\theta, \tau, g) B_n 1(\xi_{n,j}(\theta, \tau, g) > 1), \\ \overline{h}_{2,n}(\theta, \tau, g) &:= \widehat{D}_n^{-1/2}(\theta) \overline{\Sigma}_n(\theta, \tau, g) \widehat{D}_n^{-1/2}(\theta), \text{ and} \\ \overline{h}_{2,n,j}(\theta, \tau, g) &:= [\overline{h}_{2,n}(\theta, \tau, g)]_{jj}. \end{aligned} \tag{C.12}$$

We assume:

Assumption GMS1.(a) $\overline{\varphi}_n(\theta, \tau, g)$ satisfies (C.12) and $\{B_n : n \geq 1\}$ is a nondecreasing sequence of positive constants, and

(b) $\kappa_n \to \infty$ and $B_n / \kappa_n \to 0$ as $n \to \infty$.

In ASM and Andrews and Shi (2014), we use $\kappa_n = (0.3 \ln(n))^{1/2}$ and $B_n = (0.4 \ln(n) / \ln \ln(n))^{1/2}$, which satisfy Assumption GMS1.

The multiplicand $\overline{h}_{2,n,j}^{1/2}(\theta, \tau, g)$ in (C.12) is an "ε-adjusted" standard deviation estimator for the jth normalized sample moment based on g (see (3.4) for the ε-adjustment in $\overline{\Sigma}_n(\theta, \tau, g)$). It provides a suitable scaling for $\varphi_n(\theta, \tau, g)$.

D Asymptotic Size

In this section, we show that the bootstrap GMS CS's have correct uniform asymptotic coverage probabilities, i.e., correct asymptotic size.

D.1 Notation

First, define

$$
\begin{aligned}
h_{1,n,F}(\theta,\tau,g) &:= n^{1/2} D_F^{-1/2}(\theta) E_F m(W_i,\theta,\tau,g), \\
h_{n,F}(\theta,\tau,g,\tau^\dagger,g^\dagger) &:= (h_{1,n,F}(\theta,\tau,g), h_{2,F}(\theta,\tau,g,\tau^\dagger,g^\dagger)), \\
\widehat{h}_{2,n,F}(\theta,\tau,g,\tau^\dagger,g^\dagger) &:= D_F^{-1/2}(\theta)\widehat{\Sigma}_n(\theta,\tau,g,\tau^\dagger,g^\dagger) D_F^{-1/2}(\theta), \\
\overline{h}_{2,n,F}(\theta,\tau,g) &:= \widehat{h}_{2,n,F}(\theta,\tau,g,\tau,g) + \varepsilon D_F^{-1/2}(\theta)\widehat{D}_n(\theta) D_F^{-1/2}(\theta) \\
&= D_F^{-1/2}(\theta)\overline{\Sigma}_n(\theta,\tau,g) D_F^{-1/2}(\theta), \text{ and} \qquad \text{(D.1)} \\
\nu_{n,F}(\theta,\tau,g) &:= n^{-1/2}\sum_{i=1}^{n} D_F^{-1/2}(\theta)[m(W_i,\theta,\tau,g) - E_F m(W_i,\theta,\tau,g)],
\end{aligned}
$$

where $m(W_i,\theta,\tau,g)$, $\widehat{\Sigma}_n(\theta,\tau,g,\tau^\dagger,g^\dagger)$, $\overline{\Sigma}_n(\theta,\tau,g)$, $D_F(\theta)$, and $\widehat{D}_n(\theta)$ are defined in (3.2), (3.3), (3.4), and (5.1) of ASM, and (C.10), respectively.

Below we write $T_n(\theta)$ as a function of the quantities in (D.1). As defined, (i) $h_{1,n,F}(\theta,\tau,g)$ is the k-vector of normalized means of the moment functions for $(\tau,g)\in\mathcal{T}\times\mathcal{G}$, which measures the slackness of the population moment conditions under (θ,F), and it has the very useful feature that it is non-negative when $(\theta,F)\in\mathcal{F}$ by (2.1) of ASM, (ii) $h_{n,F}(\theta,\tau,g,\tau^\dagger,g^\dagger)$ contains the approximation to the normalized means of $D_F^{-1/2}(\theta)m(W_i,\theta,\tau,g)$ and the covariances of $D_F^{-1/2}(\theta)m(W_i,\theta,\tau,g)$ and $D_F^{-1/2}(\theta)m(W_i,\theta,\tau^\dagger,g^\dagger)$, (iii) $\widehat{h}_{2,n,F}(\theta,\tau,g,\tau^\dagger,g^\dagger)$ and $\overline{h}_{2,n,F}(\theta,\tau,g)$ are hybrid quantities—part population, part sample—based on the matrices $\widehat{\Sigma}_n(\theta,\tau,g,\tau^\dagger,g^\dagger)$ and $\overline{\Sigma}_n(\theta,\tau,g)$, respectively, and (iv) $\nu_{n,F}(\theta,\tau,g)$ is the sample average of the moment functions $D_F^{-1/2}(\theta)m(W_i,\theta,\tau,g)$ normalized to have mean zero and variance that is $O(1)$, but not $o(1)$. Note that $\nu_{n,F}(\theta,\cdot,\cdot)$ is an empirical process indexed by $(\tau,g)\in\mathcal{T}\times\mathcal{G}$ with covariance kernel given by $h_{2,F}(\theta,\tau,g,\tau^\dagger,g^\dagger)$.

The normalized sample moments $n^{1/2}\overline{m}_n(\theta,\tau,g)$ can be written as

$$n^{1/2}\overline{m}_n(\theta,\tau,g) = D_F^{1/2}(\theta)(\nu_{n,F}(\theta,\tau,g) + h_{1,n,F}(\theta,\tau,g)). \qquad \text{(D.2)}$$

The test statistic $T_n(\theta)$, defined in (B.3), can be written as

$$T_n(\theta) = \sup_{\tau\in\mathcal{T}} \int_{\mathcal{G}} S(\nu_{n,F}(\theta,\tau,g) + h_{1,n,F}(\theta,\tau,g), \overline{h}_{2,n,F}(\theta,\tau,g)) dQ(g). \qquad \text{(D.3)}$$

Note the close resemblance between $T_n(\theta)$ and $T(h)$ (defined in (C.1)).

Let $\mathcal{H}_1$ denote the set of all functions from $\mathcal{T} \times \mathcal{G}$ to $R^p_{[+\infty]} \times \{0\}^v$.

For notational simplicity, for any function of the form $r_F(\theta, \tau, g)$ for $(\tau, g) \in \mathcal{T} \times \mathcal{G}$, let $r_F(\theta)$ denote the function $r_F(\theta, \cdot, \cdot)$ on $\mathcal{T} \times \mathcal{G}$. Correspondingly, for any function of the form $r_F(\theta, \tau, g, \tau^\dagger, g^\dagger)$ for $(\tau, g), (\tau^\dagger, g^\dagger) \in \mathcal{T} \times \mathcal{G}$, let $r_F(\theta)$ denote the function $r_F(\theta, \cdot, \cdot, \cdot, \cdot)$ on $(\mathcal{T} \times \mathcal{G})^2$. Thus, $h_{2,F}(\theta)$ abbreviates the asymptotic covariance kernel $\{h_{2,F}(\theta, \tau, g, \tau^\dagger, g^\dagger) : (\tau, g), (\tau^\dagger, g^\dagger) \in \mathcal{T} \times \mathcal{G}\}$ defined in (C.9). Define

$$\mathcal{H}_2 := \{h_{2,F}(\theta) : (\theta, F) \in \mathcal{F}\}, \tag{D.4}$$

where, as defined at the end of Section 2, $\mathcal{F}$ is the subset of $\mathcal{F}_+$ that satisfies Assumption PS3. On the space of $k \times k$ matrix-valued covariance kernels on $(\mathcal{T} \times \mathcal{G})^2$, which is a superset of $\mathcal{H}_2$, we use the uniform metric d defined by

$$d(h_2^{(1)}, h_2^{(2)}) := \sup_{(\tau,g),(\tau^\dagger,g^\dagger) \in \mathcal{T} \times \mathcal{G}} \|h_2^{(1)}(\tau, g, \tau^\dagger, g^\dagger) - h_2^{(2)}(\tau, g, \tau^\dagger, g^\dagger)\|. \tag{D.5}$$

Let $\Rightarrow$ denote weak convergence. Let $\{a_n\}$ denote a subsequence of n. Let $\rho_{h_2}(\theta)$ be the intrinsic pseudometric on $\mathcal{T} \times \mathcal{G}$ for the tight Gaussian process $\nu_{h_2}(\theta)$ with variance-covariance kernel h_2:

$$\begin{aligned} &\rho_{h_2}(\tau, g, \tau^\dagger, g^\dagger) \\ &:= tr\left(h_2(\tau, g, \tau, g) - h_2(\tau, g, \tau^\dagger, g^\dagger) - h_2(\tau^\dagger, g^\dagger, \tau, g) + h_2(\tau^\dagger, g^\dagger, \tau^\dagger, g^\dagger)\right). \end{aligned} \tag{D.6}$$

D.2 Proof of Theorem 5.1

Theorem 5.1 of ASM is a result of two lemmas. The two lemmas together imply the uniform validity of the GMS CS over $\mathcal{F}$ under Assumptions M, S1,, S2, andGMS1.

The first lemma below establishes the uniform asymptotic size under two high-level assumptions (given below). The second lemma verifies these two assumptions under Assumptions M, S1, and S2.

Assumption PS4. For any subsequence $\{a_n\}$ of $\{n\}$ and any sequence $\{(\theta_{a_n}, F_{a_n}) \in \mathcal{F}_+ : n \geq 1\}$ for which

$$\lim_{n\to\infty} d(h_{2,F_{a_n}}(\theta_{a_n}), h_2) = 0 \tag{D.7}$$

for some $k \times k$ matrix-valued covariance kernel $h_2(\tau, g, \tau^\dagger, g^\dagger)$ on $(\mathcal{T} \times \mathcal{G})^2$, we have

(i) $\nu_{a_n,F_{a_n}}(\theta_{a_n}) \Rightarrow \nu_{h_2}(\cdot)$ and

(ii) $d(\widehat{h}_{2,a_n,F_{a_n}}(\theta_{a_n}), h_2) \to_p 0$ as $n \to \infty$, where $\widehat{h}_{2,a_n,F_{a_n}}(\theta_{a_n})$ is defined in (D.1).

Assumption PS5. For any subsequence $\{a_n\}$ of $\{n\}$ and any sequence $\{(\theta_{a_n}, F_{a_n}) \in \mathcal{F}_+ : n \geq 1\}$, conditional on any sample path ω for which

$$\lim_{n\to\infty} d(\widehat{h}_{2,a_n,F_{a_n}}(\theta_{a_n})(\omega), h_2) = 0, \tag{D.8}$$

for some $k \times k$ matrix-valued covariance kernel $h_2(\tau, g, \tau^\dagger, g^\dagger)$ on $(\mathcal{T} \times \mathcal{G})^2$, we have (i) $\nu^*_{a_n}(\theta) \Rightarrow \nu_{h_2}$ and (ii) $d(\widehat{h}^*_{2,a_n}(\theta_{a_n}), h_2) \to_p 0$.

Lemma D.1 *Suppose Assumptions* PS4, PS5, S1, S2, *and* SIG1 *hold, and Assumption* GMS1 *holds when considering* GMS *critical values. Then, for any compact subset* $\mathcal{H}_{2,cpt}$ *of* $\mathcal{H}_2$*, the* GMS *CS satisfies:*

$$\liminf_{n\to\infty} \inf_{\substack{(\theta,F)\in\mathcal{F}: \\ h_{2,F}(\theta)\in\mathcal{H}_{2,cpt}}} P_F(\theta \in CS_n) \geq 1 - \alpha.$$

Lemma D.2 *Suppose Assumptions* M, S1, *and* S2 *hold. Then,*

(a) *Assumption* PS4 *holds and*

(b) *Assumption* PS5 *holds.*

Comments. 1. Lemma D.1(a) shows that GMS CS has correct uniform asymptotic size. The uniformity results hold whether the moment conditions involve "weak" or "strong" IV's X_i.

2. Theorem 5.1 of ASM for the case $r_{1,n} = \infty$ is proved by verifying the conditions of Lemma D.2 (that is, by showing that Assumptions M, S1, S2, and GMS1 hold for the $\mathcal{G}_{\text{c-cube}}$ set and the S functions considered in ASM).[21] The functions S_1, S_2, and S_3 in (3.9) of ASM satisfy Assumptions S1 and S2 by Lemma 1 of AS1 and the function S_4 of ASM satisfy Assumptions S1 and S2 by similar arguments. Lemma 3 of AS1 establishes Assumption M for $\mathcal{G}_{\text{c-cube}}$ defined in (3.6) of ASM. Assumption GMS1 holds immediately for κ_n and B_n used in (4.1) and (4.2) of ASM, respectively. Theorem 5.1 of ASM holds for $r_{1,n}$ such that $r_{1,n} < \infty$ and $r_{1,n} \to \infty$ as $n \to \infty$ by minor alterations to the proofs.

[21] The quantity $r_{1,n}$ is the test statistic truncation value that appears in (3.7) of ASM. It satisfies either $r_{1,n} = \infty$ for all $n \geq 1$ or $r_{1,n} < \infty$ and $r_{1,n} \to \infty$ as $n \to \infty$.

D.3 Proof of Lemma D.1

D.3.1 Theorem D.3

The following Theorem provides a uniform asymptotic distributional result for the test statistic $T_n(\theta)$. It is an analogue of Theorem 1 in AS1. It is used in the proof of Lemma D.1.

Theorem D.3 *Suppose Assumptions* PS4, S1, S2, *and* SIG1 *hold. Then, for all compact subsets $\mathcal{H}_{2,cpt}$ of $\mathcal{H}_2$, for all constants $x_{h_{n,F}(\theta)} \in R$ that may depend on (θ, F) and n through $h_{n,F}(\theta)$, and all $\delta > 0$, we have*

(a) $\limsup_{n\to\infty} \sup_{\substack{(\theta,F)\in\mathcal{F}: \\ h_{2,F}(\theta)\in\mathcal{H}_{2,cpt}}} \left[P_F(T_n(\theta) > x_{h_{n,F}(\theta)}) - P\left(T(h_{n,F}(\theta)) + \delta > x_{h_{n,F}(\theta)}\right)\right] \leq 0$ *and*

(b) $\liminf_{n\to\infty} \inf_{\substack{(\theta,F)\in\mathcal{F}: \\ h_{2,F}(\theta)\in\mathcal{H}_{2,cpt}}} \left[P_F(T_n(\theta) > x_{h_{n,F}(\theta)}) - P\left(T(h_{n,F}(\theta)) - \delta > x_{h_{n,F}(\theta)}\right)\right] \geq 0,$

where $T(h)$ is the function defined in (C.1).

Proof of Theorem D.3. Theorem D.3 is similar to Theorem 1 in AS1. The proof of the latter theorem goes through with the following modifications:

(i) Redefine $SubSeq(h_2)$ to be the set of subsequences $\{(\theta_{a_n}, F_{a_n}) \in \mathcal{F} : n \geq 1\}$ where $\{a_n\}$ is a subsequence of $\{n\}$, such that (D.7) holds.

(ii) Replace $\int \cdots dQ(g)$ by $\sup_{\tau\in\mathcal{T}} \int_{\mathcal{G}} \cdots dQ(g)$. In other instances where g and $\mathcal{G}$ appear, replace g with (τ, g) and $\mathcal{G}$ with $\mathcal{T} \times \mathcal{G}$.

(iii) Replace "by Lemma A1" with "by Assumption PS4."

(iv) Change the paragraph at the bottom of p. 6 of AS2 to the following:

"Given this and Assumption SIG1, by the almost sure representation theorem, e.g., see Pollard (1990, Thm. 9.4), there exists a probability space and random quantities $\tilde{\nu}_{a_n}(\cdot)$, $\tilde{h}_{2,a_n}(\cdot)$, $\tilde{V}_{a_n}$, and $\tilde{\nu}_0(\cdot)$ defined on it such that (i) $(\tilde{\nu}_{a_n}(\cdot), \tilde{h}_{2,a_n}(\cdot), \tilde{V}_{a_n})$ has the same distribution as $(\nu_{a_n,F_{a_n}}(\theta_{a_n}, \cdot), \widehat{h}_{2,a_n,F_{a_n}}(\theta_{a_n}, \cdot), D^{-1/2}_{F_{a_n}}(\theta_{a_n})\widehat{D}_{a_n}(\theta_{a_n})D^{-1/2}_{F_{a_n}}(\theta_{a_n}))$, (ii) $(\tilde{\nu}_0(\cdot))$ has the same distribution as $\nu_{h_{2,0}}(\cdot)$, and

$$\text{(iii)} \quad \sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}} \left\| \begin{pmatrix} \tilde{\nu}_{a_n}(\tau,g) \\ \tilde{h}_{2,a_n}(\tau,g) \\ vec(\tilde{V}_{a_n}) \end{pmatrix} - \begin{pmatrix} \tilde{\nu}_0(\tau,g) \\ h_{2,0}(\tau,g) \\ vec(I_k) \end{pmatrix} \right\| \to 0 \text{ as } n\to\infty, \text{ a.s.} \qquad \text{(D.9)}$$

(v) Replace $Diag(\tilde{h}_{2,a_n}(1_k))$ by $\tilde{V}_{a_n}$.

With the above modifications, the proof of Theorem 1 in AS2 up to the proof of (12.7) of AS2 goes through. The proof of (12.7) in AS2, which relies on a dominated convergence argument, does not go through because the test statistic considered in this paper is not of the pure CvM type, and thus, $\tilde{T}_{a_n}$ and $\tilde{T}_{a_n,0}$ are not integrals with respect to (τ, g).

We change the proof of (12.7) in AS2 to the following.

As in the proof of (12.7) in AS2, we fix a sample path ω at which $(\tilde{\nu}_{a_n}(\tau, g), \tilde{h}_{2,a_n}(\tau, g))(\omega)$ converges to $(\tilde{\nu}_0(\tau, g), h_{2,0}(\tau, g))(\omega)$ uniformly over $(\tau, g) \in \mathcal{T} \times \mathcal{G}$ as $n \to \infty$ and $\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}} \|\tilde{\nu}_0(\tau, g)(\omega)\| < \infty$. Let $\tilde{\Omega}$ be the collection of such sample paths. By (D.9), $P(\tilde{\Omega}) = 1$. For a fixed $\omega \in \tilde{\Omega}$, by Assumption S2, we have

$$\sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}} \sup_{\mu\in[0,\infty)^p\times\{0\}^v} |S(\tilde{\nu}_{a_n}(\tau, g)(\omega) + \mu, \tilde{h}^{\varepsilon}_{2,a_n}(\tau, g)(\omega)) - S(\tilde{\nu}_0(\tau, g)(\omega) + \mu, h^{\varepsilon}_{2,0}(\tau, g))| \to 0, \tag{D.10}$$

as $n \to \infty$, where $\tilde{h}^{\varepsilon}_{2,n}(\tau, g) := \tilde{h}_{2,n}(\tau, g) + \varepsilon\tilde{V}_{a_n}$, and $h^{\varepsilon}_{2,0}(\tau, g) := h_{2,0}(\tau, g) + \varepsilon I_k$. Thus, for every $\omega \in \tilde{\Omega}$,

$$\begin{aligned} |\tilde{T}_{a_n}(\omega) - \tilde{T}_{a_n,0}(\omega)| \le \sup_{(\tau,g)\in\mathcal{T}\times\mathcal{G}} & |S(\tilde{\nu}_{a_n}(\tau, g)(\omega) + h_{1,a_n,F_{a_n}}(\theta_{a_n}, \tau, g), \tilde{h}^{\varepsilon}_{2,a_n}(\tau, g)(\omega)) - \\ & - S(\tilde{\nu}_0(\tau, g)(\omega) + h_{1,a_n,F_{a_n}}(\theta_{a_n}, \tau, g), h^{\varepsilon}_{2,0}(\tau, g))| \\ & \to 0 \text{ as } n \to \infty. \end{aligned} \tag{D.11}$$

This verifies (12.7) in AS2. □

D.3.2 Proof of Lemma D.1

Lemma D.1 is similar to Theorem 2(a) of AS1 and we modify the proof of the latter in AS2 to fit the context of Lemma D.1. In addition to notational changes, a substantial modification is needed because Theorem 2 of AS1 does not cover bootstrap critical values.

Specifically, the proof of Theorem 2(a) in AS2 with the following modifications provides the proof of Lemma D.1.

(i) Replace all references to "Assumption M" of AS1 by references to "Assumption PS4" stated above and Assumptions S1 and S2 of AS1 by Assumptions S1 and S2 stated above. Replace $\int \cdots dQ(g)$ by $\sup_{\tau\in\mathcal{T}} \int_{\mathcal{G}} \cdots dQ(g)$. In other instances where g and $\mathcal{G}$ appear, replace g with (τ, g) and $\mathcal{G}$ with $\mathcal{T} \times \mathcal{G}$. Let $\widehat{D}_{a_n}(\theta_{a_n})$ be defined as in (C.10) above, rather than as in AS1 and AS2.

(ii) Replace references to "Theorem 1(a)" of AS1 with references to "Theorem D.3(a)" stated above.

(iii) Redefine $SubSeq(h_2)$ to be the set of subsequences $\{(\theta_{a_n}, F_{a_n}) \in \mathcal{F} : n \geq 1\}$ for which (D.7) holds, where $\{a_n\}$ is a subsequence of $\{n\}$.

(iv) Replace references to "Lemma A1" of AS2 to references to "Assumption PS4" stated above.

(v) In both the statement and the proof of Lemma A3 in AS2, replace $c(\varphi_n(\theta), \widehat{h}_{2,n}(\theta), 1-\alpha)$ with $c_0(\varphi_n(\theta), h_{2,F}(\theta), 1-\alpha)$, and $c(h_{1,n,F}(\theta), \widehat{h}_{2,n}(\theta), 1-\alpha)$ with $c_0(h_{1,n,F}(\theta), h_{2,F}(\theta), 1-\alpha)$. The proof of Lemma A3 given in AS2 goes through with the following changes:

In the 6th and 7th last lines of the proof of Lemma A3 in AS2, delete "$\varepsilon^{-1/2} h_{2,0,j}^{-1/2}(1_k, 1_k)(1 + o_p(1)) =$", and change "by Lemma A1(b) and (5.2)" to "by Assumption SIG1 and (D.1)."

(vi) Replace Lemma A4 in AS2 with Lemma D.4 given immediately below. The proof of the Lemma D.4 given below is self-contained and does not rely on an analogue of Lemma A5 of AS2.

No other changes are needed in the proof of Theorem 2(a) in AS2. □

The following lemma is used in the proof of Lemma D.1 given immediately above.

Lemma D.4 *Suppose Assumptions* PS4, PS5, S1, S2, *and* GMS1 *hold. Then, for all* $\delta \in (0, \eta)$, *where* $\eta > 0$ *is defined in* (C.4),

$$\lim_{n\to\infty} \sup_{\substack{(\theta,F)\in\mathcal{F}: \\ h_{2,F}(\theta)\in\mathcal{H}_{2,cpt}}} P_F\left(c^*(\varphi_n(\theta), \widehat{h}^*_{2,n}(\theta), 1-\alpha) < c_0(\varphi_n(\theta), h_{2,F}(\theta), 1-\alpha) + \delta\right) = 0.$$

Prove of Lemma D.4. The result of the Lemma is equivalent to

$$\lim_{n\to\infty} \sup_{\substack{(\theta,F)\in\mathcal{F}: \\ h_{2,F}(\theta)\in\mathcal{H}_{2,cpt}}} P_F(c_0^*(\varphi_n(\theta), \widehat{h}^*_{2,n}(\theta), 1-\alpha+\eta) + \eta < c_0(\varphi_n(\theta), h_{2,F}(\theta), 1-\alpha) + \delta) = 0. \tag{D.12}$$

By considering a sequence $\{(\theta_n, F_n) \in \mathcal{F} : n \geq 1\}$ that is within $\zeta_n \to 0$ of the supremum in the above display for all $n \geq 1$, it suffices to show that

$$\lim_{n\to\infty} P_{F_n}(c_0^*(\varphi_n(\theta_n), \widehat{h}^*_{2,n}(\theta_n), 1-\alpha+\eta) + \eta < c_0(\varphi_n(\theta_n), h_{2,F_n}(\theta_n), 1-\alpha) + \delta) = 0. \tag{D.13}$$

Given any subsequence $\{u_n\}$ of $\{n\}$, there exists a further subsequence $\{w_n\}$ such that $d(h_{2,F_{w_n}}(\theta_{w_n}), h_{2,0}) \to 0$ as $n \to \infty$ for some matrix-valued covariance function $h_{2,0}$ by the compactness of $\mathcal{H}_{2,cpt}$. It suffices to show that (D.13) holds with w_n in place of n.

By Assumption PS4(ii), $d(h_{2,F_{w_n}}(\theta_{w_n}), h_{2,0}) \to 0$ implies that $d(\widehat{h}_{2,w_n,F_{w_n}}(\theta_{w_n}), h_{2,0}) \to_p 0$, which then implies

$$d(\widehat{h}_{2,w_n}(\theta_{w_n}), h_{2,0}) \to_p 0, \tag{D.14}$$

where $\widehat{h}_{2,n}(\theta)$ and $\widehat{h}_{2,n,F}(\theta)$ are defined in (C.10) and (D.1), respectively. Then, by a general convergence in probability result, given any subsequence of $\{w_n\}$ there exists a further subsequence $\{a_n\}$ such that

$$d(\widehat{h}_{2,a_n}(\theta_{a_n}), h_{2,0}) \to 0 \text{ a.s.} \tag{D.15}$$

Hence, it suffices to show (D.13) with a_n in place of n. Let $\bar{\Omega}$ be the set of sample paths ω such that $d(\widehat{h}_{2,a_n}(\theta_{a_n})(\omega), h_2) \to 0$. The above display implies that $P(\bar{\Omega}) = 1$.

Consider an arbitrary sample path $\omega \in \bar{\Omega}$. Below we show that for all constants $x_n \in R$ (possibly dependent on ω) and all $\xi > 0$,

$$\begin{aligned}&\limsup_{n\to\infty} \Big[P\left(T^*(\varphi_{a_n}(\theta_{a_n}), \widehat{h}^*_{2,a_n}(\theta_{a_n})) \le x_{a_n}|\omega\right) - \\ &P\left(T(\varphi_{a_n}(\theta_{a_n}), h_{2,F_{a_n}}(\theta_{a_n})) \le x_{a_n} + \xi|\omega\right)\Big] \le 0,\end{aligned} \tag{D.16}$$

where in the first line $P(\cdot|\omega)$ denotes bootstrap probability conditional on the original sample path ω, in the second line $P(\cdot|\omega)$ denotes $\nu_{h_{2,F_{a_n}}(\theta_{a_n})}(\cdot)$ probability conditional on the original sample path ω, and $\nu_{h_{2,F_{a_n}}(\theta_{a_n})}(\cdot)$ is the Gaussian process defined in (C.2) with $h_2 = h_{2,F_{a_n}}(\theta_{a_n})$, which is taken to be independent of the original sample $\{W_i : i \le n\}$ and, hence, is independent of $\varphi_{a_n}(\theta_{a_n})$.

The interval $(0, \eta - \delta)$ is non-empty because $\delta \in (0, \eta)$ by assumption. Using (D.16), we obtain, for all $\xi \in (0, \eta - \delta)$,

$$\begin{aligned}&\limsup_{n\to\infty} P\left(T^*(\varphi_{a_n}(\theta_{a_n}), \widehat{h}^*_{2,a_n}(\theta_{a_n})) \le c_0(\varphi_{a_n}(\theta_{a_n}), h_{2,F_{a_n}}(\theta_{a_n}), 1-\alpha) + \delta - \eta|\omega\right) \\ &\le \limsup_{n\to\infty} P\left(T(\varphi_{a_n}(\theta_{a_n}), h_{2,F_{a_n}}(\theta_{a_n})) \le c_0(\varphi_{a_n}(\theta_{a_n}), h_{2,F_{a_n}}(\theta_{a_n}), 1-\alpha) + \delta - \eta + \xi|\omega\right) \\ &\le 1 - \alpha,\end{aligned} \tag{D.17}$$

where the second inequality holds because $\delta - \eta + \xi < 0$ for $\xi \in (0, \eta - \delta)$. For any df F with $1 - \alpha + \eta$ quantile denoted by $q_{1-\alpha+\eta}$, we have $F(q_{1-\alpha+\eta}) \geq 1 - \alpha + \eta$. Hence, if $F(x) < 1 - \alpha + \eta$, then $x < q_{1-\alpha+\eta}$. Combining this with the result in (D.17) implies that given any $\delta \in (0, \eta)$, for n sufficiently large,

$$c_0(\varphi_{a_n}(\theta_{a_n})(\omega), h_{2,F_{a_n}}(\theta_{a_n}), 1-\alpha) + \delta - \eta < c_0^*(\varphi_{a_n}(\theta_{a_n}), h_{2,F_{a_n}}(\theta_{a_n}), 1-\alpha+\eta)(\omega), \quad (D.18)$$

where the indexing by ω denotes that the result holds for fixed $\omega \in \bar{\Omega}$. Because (D.18) holds for all $\omega \in \bar{\Omega}$ and $P(\bar{\Omega}) = 1$, the bounded convergence theorem applies and establishes (D.13).

It remains to prove the result in (D.16). This result follows from an analogous argument to that used to prove Theorem D.3(b). Note the common structure of the original sample and bootstrap sample test statistics:

$$\begin{aligned} T_n(\theta_n) &= \sup_{\tau \in \mathcal{T}} \int S(\nu_{n,F_n}(\theta_n, \tau, g) + h_{1,n,F_n}(\theta_n, \tau, g), \overline{h}_{2,n,F_n}(\theta_n, \tau, g)) dQ(g), \\ T_n^*(\varphi_n(\theta_n), \widehat{h}_{2,n}^*(\theta_n)) &= \sup_{\tau \in \mathcal{T}} \int S(\nu_n^*(\theta_n, \tau, g) + \varphi_n(\theta_n, \tau, g), \overline{h}_{2,n}^*(\theta_n, \tau, g)) dQ(g), \text{ where} \\ \overline{h}_{2,n}^*(\theta, \tau, g) &:= \widehat{h}_{2,n}^*(\theta, \tau, g) + \varepsilon I_k, \end{aligned}$$

$\nu_{n,F}$, $h_{1,n,F}$, and $\overline{h}_{2,n,F}$ are defined in (D.1), $\varphi_n(\theta)$ is defined in (C.12), and $\widehat{h}_{2,n}^*(\theta)$ is defined following (C.5) using (C.10) with W_i^* in place of W_i.

The result of Theorem D.3(b) with $T_n(\theta_n)$ replaced by $T_n^*(\varphi_n(\theta_n), \widehat{h}_{2,n}^*(\theta_n))$, with $T(h_{n,F}(\theta))$ replaced by $T(\varphi_n(\theta_n), h_{2,F_n}(\theta_n))$, and with δ replaced by ξ, when applied to the subsequence $\{(\theta_{a_n}, F_{a_n}) : n \geq 1\}$ is the result of (D.16). The result in (D.16) follows by the same argument as that for Theorem D.3(b) with $\nu_{a_n,F_{a_n}}(\theta_{a_n}, \cdot)$ replaced by $\nu_{a_n}^*(\theta_{a_n}, \cdot)(\omega)$, where $\nu_{a_n}^*(\theta_{a_n}, \cdot)(\omega)$ denotes the bootstrap empirical process given the sample path ω of the original sample, with $\widehat{h}_{2,a_n,F_{a_n}}(\theta_{a_n}, \cdot, \cdot)$ replaced by $\widehat{h}_{2,a_n}^*(\theta_{a_n}, \cdot, \cdot)(\omega)$, and with Assumption PS4 replaced by Assumption PS5, which guarantees that $\nu_{a_n}^*(\theta_{a_n})(\omega) \Rightarrow \nu_{h_{2,0}}$ and $d(\widehat{h}_{2,a_n}^*(\theta_{a_n})(\omega), h_{2,0}) \to_p 0$. □

D.4 Proof of Lemma D.2

The verification of Assumption PS4 is the same as the proof of Lemma A1 given in Appendix E of AS2 except with some notation changes and with Lemma D.5 below replacing Lemma E1(a) in AS2 in the proof. (Lemma A1 of AS2 is stated in Appendix A of AS2.) The verification of Assumption PS5 is the same as that of Assumption PS4 except that all arguments are conditional on the sample path ω (specified in Assumption PS5). Details are omitted for brevity.

Lemma D.5 *Let* $(\Omega, \mathbb{F}, P)$ *be a probability space and let* ω *denote a generic element in* Ω. *Suppose that the row-wise i.i.d. triangular arrays of random processes* $\{\phi_{n,i}(\omega, g) : g \in \mathcal{G}, i \leq n, n \geq 1\}$ *and* $\{c_{n,i}(\omega, \tau) : \tau \in \mathcal{T}, i \leq n, n \geq 1\}$ *are manageable with respect to the envelopes* $\{F_n(\omega) : \Omega \to R^n : n \geq 1\}$ *and* $\{C_n(\omega) : \Omega \to R^n : n \geq 1\}$, *respectively. Then,* $\{\phi_{n,i}(\omega, g)c_{n,i}(\omega, \tau) : (\tau, g) \in \mathcal{T} \times \mathcal{G}, i \leq n, n \geq 1\}$ *is manageable with respect to the envelopes* $\{F_n(\omega) \odot C_n(\omega) : n \geq 1\}$, *where* $\odot$ *stands for the coordinate-wise product.*

Proof of Lemma D.5. For a positive number ξ and a Euclidean space G, the packing number $D(\xi, G)$ is defined in Section B.3. For each $\omega \in \Omega$ and each $n \geq 1$, let $\mathcal{F}_{n,\omega} := \{(\phi_{n,1}(\omega, g), ..., \phi_{n,n}(\omega, g))' : g \in \mathcal{G}\}$, and let $\mathcal{C}_{n,\omega} := \{(c_{n,1}(\omega, \tau), ..., c_{n,n}(\omega, \tau))' : \tau \in \mathcal{T}\}$. Let $\lambda_\phi(\varepsilon)$ and $\lambda_c(\varepsilon)$ be the deterministic functions that (i) bound from above $D(\varepsilon||\alpha \odot F_n(\omega)||, \alpha \odot \mathcal{F}_{n,\omega})$ and $D(\varepsilon||\alpha \odot C_n(\omega)||, \alpha \odot \mathcal{C}_{n,\omega})$, respectively, for an arbitrary nonnegative n-vector α, and (ii) satisfy $\int_0^1 \sqrt{\log \lambda_\phi(x)}dx < \infty$ and $\int_0^1 \sqrt{\log \lambda_c(x)}dx < \infty$. Such functions exist by the assumed manageability of the triangular arrays of random processes in the lemma.

For an arbitrary $\varepsilon > 0$, construct a bound for $D(\varepsilon||\alpha \odot F_n(\omega) \odot C_n(\omega)||, \alpha \odot \mathcal{F}_{n,\omega} \odot \mathcal{C}_{n,\omega})$ as follows:

$$
\begin{aligned}
& D(\varepsilon||\alpha \odot F_n(\omega) \odot C_n(\omega)||, \alpha \odot \mathcal{F}_{n,\omega} \odot \mathcal{C}_{n,\omega}) \\
\leq\ & D((\varepsilon/4)||\alpha \odot F_n(\omega) \odot C_n(\omega)||, \alpha \odot F_n(\omega) \odot \mathcal{C}_{n,\omega}) \\
& \times D((\varepsilon/4)||\alpha \odot F_n(\omega) \odot C_n(\omega)||, \alpha \odot C_n(\omega) \odot \mathcal{F}_{n,\omega}) \\
\leq\ & \sup_{\alpha^* \in R^n_+} D((\varepsilon/4)||\alpha^* \odot C_n(\omega)||, \alpha^* \odot \mathcal{C}_{n,\omega}) \sup_{\alpha^* \in R^n_+} D((\varepsilon/4)||\alpha^* \odot F_n(\omega)||, \alpha^* \odot \mathcal{F}_{n,\omega}) \\
\leq\ & \lambda_\phi(\varepsilon/4)\lambda_c(\varepsilon/4),
\end{aligned} \tag{D.19}
$$

where the first inequality holds by the displayed equation following (5.2) in Pollard (1990), the second inequality holds because $\alpha \odot F_n(\omega), \alpha \odot C_n(\omega) \in R_+^n$, and the last inequality holds by the definitions of $\lambda_\phi(\varepsilon)$ and $\lambda_c(\varepsilon)$.

Then, the manageability of $\{\phi_{n,i}(\omega, g)c_{n,i}(\omega, g) : (\tau, g) \in \mathcal{T} \times \mathcal{G}, i \leq n, n \geq 1\}$ with respect to the envelopes $\{F_n(\omega) \odot C_n(\omega) : n \geq 1\}$ is proved by the following calculations:

$$\begin{aligned}
\int_0^1 \sqrt{\log(\lambda_\phi(x/4)\lambda_c(x/4))}dx &\leq \int_0^1 \sqrt{\log \lambda_\phi(x/4)}dx + \int_0^1 \sqrt{\log \lambda_c(x/4)}dx \\
&= 4\int_0^{1/4} \sqrt{\log \lambda_\phi(y)}dy + 4\int_0^{1/4} \sqrt{\log \lambda_c(y)}dy \\
&\leq 4\int_0^1 \sqrt{\log \lambda_\phi(y)}dy + 4\int_0^1 \sqrt{\log \lambda_c(y)}dy \\
&< \infty, \qquad \text{(D.20)}
\end{aligned}$$

where the first inequality holds by the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b > 0$, the equality holds by a change of variables, the second inequality holds because the integrands are nonnegative on $(1/4, 1]$, and the last inequality holds by the definitions of $\lambda_\phi(\varepsilon)$ and $\lambda_c(\varepsilon)$. $\square$

E Power Against Fixed Alternatives

We now show that the power of the GMS test converges to one as $n \to \infty$ for all fixed alternatives. Thus, the test is a consistent test.

Recall that the null hypothesis is

$$\begin{aligned}
H_0 : &E_{F_0}[m_j(W_i, \theta_*, \tau)|X_i] \geq 0 \text{ a.s. } [F_{X,0}] \text{ for } j = 1, ..., p \text{ and} \\
&E_{F_0}[m_j(W_i, \theta_*, \tau)|X_i] = 0 \text{ a.s. } [F_{X,0}] \text{ for } j = p+1, ..., k, \ \forall \tau \in \mathcal{T}, \qquad \text{(E.1)}
\end{aligned}$$

where θ_* denotes the null parameter value and F_0 denotes the fixed true distribution of the data. The alternative is that H_0 does not hold. Assumption MFA of ASM specifies the properties of fixed alternatives. For convenience, we restate this assumption here. Recall that $\mathcal{X}_F(\theta, \tau)$, defined in (6.2), is the set of points $x \in R^{d_x}$ such that under F there is a violation of some conditional moment inequality or equality, evaluated at (θ, τ), conditional on $X_i = x$.

Assumption MFA. The value $\theta_* \in \Theta$ and the true distribution F_0 satisfy: (a) for some

$\tau_* \in \mathcal{T}$, $P_{F_0}(X_i \in \mathcal{X}_{F_0}(\theta_*, \tau_*)) > 0$ and (b) $(\theta_*, F_0) \in \mathcal{F}_+$.

The following assumption requires the measure Q on $\mathcal{G}$ to have full support. For each $(\theta, F, \tau) \in \mathcal{F}_+ \times \mathcal{T}$, define a pseudometric on $\mathcal{G}$: $d_{(\theta,,F,\tau)}(g, g^\dagger) = \|E_F[m(W_i, \theta, \tau)(g(X_i) - g^\dagger(X_i))]\|$ for $g, g^\dagger \in \mathcal{G}$. Let $\mathcal{B}_{d_{(\theta,F,\tau)}}(g_0, \delta) = \{g \in \mathcal{G} : d_{(\theta,F,\tau)}(g, g_0) \leq \delta\}$.

Assumption MQ. The support of Q under $d_{(\theta,F,\tau)}$ is $\mathcal{G}$ for all $(\theta, F, \tau) \in \mathcal{F}_+ \times \mathcal{T}$. That is, $\forall(\theta, F, \tau) \in \mathcal{F}_+ \times \mathcal{T}$, $\forall\delta > 0$, and $\forall g_0 \in \mathcal{G}$, $Q(B_{d(\theta,F,\tau)}(g_0, \delta)) > 0$.

The following theorem shows that the GMS test is consistent against all fixed alternatives defined in Assumption MFA.

Theorem E.1 *Suppose Assumptions* PS4, PS5, MFA, CI, MQ, S1, S3, S4, *and* SIG2 *hold. Then,*

(a) $\lim_{n\to\infty} P_{F_0}(T_n(\theta_*) > c^*(\varphi_n(\theta_*), \widehat{h}^*_{2,n}(\theta_*), 1-\alpha) = 1$, *and*

(b) $\lim_{n\to\infty} P_{F_0}(T_n(\theta_*) > c^*(0^k_{\mathcal{T}\times\mathcal{G}}, \widehat{h}^*_{2,n}(\theta_*), 1-\alpha) = 1$.

Comments. 1. Theorem 6.1 of ASM for the case $r_{1,n} = \infty$ is proved by verifying that the conditions of Theorem E.1 (except Assumption MFA) hold for the $\mathcal{G}_{\text{c-cube}}$ set, the S functions, and the measure Q_{AR} defined as in ASM. (See Section B.5 for the definition of Q_{AR} with weight function $w(r) := (r^2 + 100)^{-1}$.) Assumption CI holds for $\mathcal{G}_{\text{c-cube}}$ defined in (3.6) of ASM by Lemma 3 of AS1. Assumption MQ holds for $\mathcal{G}_{\text{c-cube}}$ and Q_{AR} because $\mathcal{G}_{\text{c-cube}}$ is countable and Q_{AR} has a probability mass function that is positive at each element in $\mathcal{G}_{\text{c-cube}}$. Assumptions S1-S4 hold for the functions S_1, S_2, and S_3 defined in (3.9) of ASM by Lemma 1 of AS1, and for S_4 in (3.9) by similar arguments. Assumptions PS4 and PS5 hold by Lemma D.2 provided Assumption M holds. Assumption M holds for $\mathcal{G}_{\text{c-cube}}$ by Lemma 3 of AS1. (Note that Assumption M with F_0 in place of F_n in part (b) holds because $\mathcal{C}_{\text{c-cube}}$ is a Vapnik-Cervonenkis class of sets.)

2. Theorem 6.1 of ASM holds for $r_{1,n}$ such that $r_{1,n} < \infty$ and $r_{1,n} \to \infty$ as $n \to \infty$ by making some alterations to the proof of Theorem E.1. The alterations required are the same as those described for A-CvM tests in the proof of Theorem B2 in Appendix D of AS2.[22]

[22]The proof of Theorem B2 describes alterations to the proof of Theorem 3 of AS1, which is given in Appendix C of AS2, to accommodate A-CvM tests based on truncation, simulation, or quasi-Monte Carlo computation and KS tests. Theorem 3 of AS1 establishes that the tests in AS1 have asymptotic power equal to one for fixed alternative distributions.

Proof of Theorem E.1. Because $\varphi_n(\theta_*) \geq 0$ and $c^*(\cdot, \widehat{h}^*_{2,n}(\theta_*), 1-\alpha)$ is weakly decreasing by definition, we have that part (a) is implied by part (b). Therefore, it suffices to show part (b) only.

Let

$$A(\theta_*) := \sup_{\tau \in \mathcal{T}} \int_{\mathcal{G}} S(E_{F_0}[m(W_i, \theta_*, \tau) g(X_i)], \overline{\Sigma}_{F_0}(\theta_*, \tau, g) dQ(g). \tag{E.2}$$

First, we show that

$$|n^{-\chi/2} T_n(\theta_*) - A(\theta_*)| \rightarrow_p 0. \tag{E.3}$$

For any $\delta > 0$,

$$\begin{aligned}
& P_{F_0}(|n^{-\chi/2} T_n(\theta_*) - A(\theta_*)| > \delta) \\
\leq\ & P_{F_0}\left(\sup_{\substack{\mu \in [0,\infty) \\ (\tau,g) \in \mathcal{T} \times \mathcal{G}}} \left| S(n^{-1/2} \nu_{n,F_0}(\theta_*, \tau, g) + \mu, \widehat{h}^{\varepsilon}_{2,n,F_0}(\theta_*, \tau, g)) - S(\mu, h^{\varepsilon}_{2,F_0}(\theta_*, \tau, g)) \right| > \delta \right) \\
\leq\ & P_{F_0}\left(\sup_{(\tau,g) \in \mathcal{T} \times \mathcal{G}} ||n^{-1/2} \nu_{n,F_0}(\theta_*, \tau, g)|| + \sup_{(\tau,g) \in \mathcal{T} \times \mathcal{G}} ||\widehat{h}^{\varepsilon}_{2,n,F_0}(\theta_*, \tau, g) - h^{\varepsilon}_{2,F_0}(\theta_*, \tau, g)|| > \xi_\delta \right) \\
\rightarrow\ & 0 \text{ as } n \rightarrow \infty,
\end{aligned} \tag{E.4}$$

where $\widehat{h}^{\varepsilon}_{2,n,F_0}(\theta, \tau, g) := \widehat{h}_{2,n,F_0}(\theta, \tau, g) + \varepsilon D^{-1/2}_{F_0}(\theta_n) \widehat{D}_0(\theta_n) D^{-1/2}_{F_0}(\theta_n)$, $h^{\varepsilon}_{2,F_0}(\theta_*, \tau, g) := h_{2,F_0}(\theta_*, \tau, g) + \varepsilon I_k$, the first inequality uses Assumption S4, and the second inequality holds for some $\xi_\delta > 0$ by Assumptions PS4, S2 and SIG2. This establishes (E.3).

Next we show that $A(\theta_*) > 0$. By Assumption MFA, there exists a $\tau_* \in \mathcal{T}$ and either a $j_0 \leq p$ such that $P_{F_0}(E_{F_0}[m_{j_0}(W_i, \theta_*, \tau_*)|X_i] < 0) > 0$ or a $j_0 > p$ such that $P_{F_0}(E_{F_0}[m_{j_0}(W_i, \theta_*, \tau_*)|X_i] \neq 0) > 0$. Without loss of generality, assume that $j_0 \leq p$. By Assumption CI, there is a $g_* \in \mathcal{G}$ such that $E_{F_0}[m_{j_0}(W_i, \theta_*, \tau_*) g_{*j_0}(X_i)] < 0$, where $g_{*j_0}(X_i)$ denotes the j_0th element of $g_*(X_i)$.

Because $E_{F_0}[m_{j_0}(W_i, \theta_*, \tau_*) g_{*j_0}(X_i)]$ is continuous in g with respect to the pseudo-metric $d_{(\theta_*, F_0, \tau_*)}$, there exists a $\delta > 0$ such that $\forall g \in \mathcal{B}_{d(\theta_*, F_0, \tau_*)}(g_*, \delta)$, $E_{F_0}[m_{j_0}(W_i, \theta_*, \tau_*) g_{j_0}(X_i)]$ has the same sign as $E_{F_0}[m_{j_0}(W_i, \theta_*, \tau_*) g_{*j_0}(X_i)]$, i.e., $E_{F_0}[m_{j_0}(W_i, \theta_*, \tau_*) g_{j_0}(X_i)] < 0$, $\forall g \in \mathcal{B}_{d(\theta_*, F_0, \tau_*)}(g_*, \delta)$. By Assumption MQ, $Q(\mathcal{B}_{d(\theta_*, F_0, \tau_*)}(g_*, \delta)) > 0$. Therefore,

$$A(\theta_*) \geq \int_{\mathcal{B}_{d(\theta_*, F_0, \tau_*)}(g_*, \delta)} S(E_{F_0}[m(W_i, \theta_*, \tau_*) g(X_i)], \overline{\Sigma}_{F_0}(\theta_*, \tau_*, g) dQ(g) > 0, \tag{E.5}$$

where the second inequality holds by Assumption S3 and $Q(\mathcal{B}_{d(\theta_*,F_0,\tau_*)}(g_*,\delta)) > 0$.

Analogous arguments to those used to establish (14.34) of AS2 show

$$c^*(0^k_{\mathcal{T}\times\mathcal{G}}, \widehat{h}^*_{2,n}(\theta_*), 1-\alpha) = O_p(1). \tag{E.6}$$

Equations (E.3), (E.5), and (E.6) give

$$\begin{aligned} & P_{F_0}(T_n(\theta_*) > c^*(0^k_{\mathcal{T}\times\mathcal{G}}, \widehat{h}^*_{2,n}(\theta_*), 1-\alpha)) \\ = \ & P_{F_0}(n^{-\chi/2}T_n(\theta_*) > n^{-\chi/2}c^*(0^k_{\mathcal{T}\times\mathcal{G}}, \widehat{h}^*_{2,n}(\theta_*), 1-\alpha)) \\ = \ & P_{F_0}(A(\theta_*) + o_p(1) > o_p(1)) \\ \to \ & 1 \text{ as } n \to \infty, \end{aligned} \tag{E.7}$$

which establishes part (b). □

F Proofs of Results Concerning the Examples

Proof of Lemma 7.1. Assumption PS1(a) holds because $\Theta = \{0\}$. Assumptions PS1(b) holds by the condition given in the lemma. Assumption PS1(c) holds because $\sigma^2_{F,1}(0) = 1$. Assumption PS1(d) holds because $|1\{Y_2 \leq \tau\} - 1\{Y_1 \leq \tau\}| \leq 1$. Assumption PS1(e) holds because

$$E_F M^{2+\delta}(W) = 1/\sigma_F^{2+\delta}(0) = 1. \tag{F.1}$$

Next, we verify Assumption PS2. For $j = 1, 2$, consider the set $\mathcal{M}_{n,j,y_j} := \{(-1\{y_{j,i} \leq \tau\})_{i=1}^n \in R^n : \tau \in \mathcal{T}\}$ for an arbitrary realization $\{y_{j,i} : i \leq n\}$ of the random vector $\{Y_{j,i} : i \leq n\}$. The set has pseudo-dimension (defined on p. 15 of Pollard (1990)) at most one by Lemma 4.4 of Pollard (1990). Then, by Corollary 4.10 of Pollard (1990), there exist constants $c_1 \geq 1$ and $c_2 > 0$ (not depending on j, n, ε, or $\{y_{j,i} : i \leq n\}$) such that

$$D(\varepsilon||\alpha||, \alpha \odot \mathcal{M}_{n,j,y_j}) \leq c_1\varepsilon^{-c_2} \tag{F.2}$$

for $\varepsilon \in (0, 1]$, every rescaling vector $\alpha \in R^n_+$, and $j = 1, 2$. In consequence, by the stability of the L_2 packing numbers (see Pollard (1990, p. 22)), we have

$$\begin{aligned} & D(2\varepsilon||\alpha||, (\alpha \odot \mathcal{M}_{n,1,y_1}) \oplus (\alpha \odot \mathcal{M}_{n,2,y_2})) \\ \leq\ & D(\varepsilon||\alpha||, \alpha \odot \mathcal{M}_{n,1,y_1}) D(\varepsilon||\alpha||, \alpha \odot \mathcal{M}_{n,2,y_2}) \\ \leq\ & c_1^2 \varepsilon^{-2c_2}, \end{aligned} \tag{F.3}$$

where $A \oplus B = \{a + b : a \in A, b \in B\}$ for any two sets $A, B \subset R^n$.

Now consider the set $\mathcal{M}_{n,y_1,y_2} := \{(1\{y_{2,i} \leq \tau\} - 1\{y_{1,i} \leq \tau\})_{i=1}^n \in R^n : \tau \in \mathcal{T}\}$. By definition, $\alpha \odot \mathcal{M}_{n,y_1,y_2} \subset (\alpha \odot \mathcal{M}_{n,1,y_1}) \oplus (\alpha \odot \mathcal{M}_{n,2,y_2})$. Thus,

$$D(2\varepsilon||\alpha||, \alpha \odot \mathcal{M}_{n,y_1,y_2}) \leq c_1^2 \varepsilon^{-2c_2}. \tag{F.4}$$

Lastly, because c_1 and c_2 do not depend on n or $\{(Y_{1i}, Y_{2i}) : i \leq n\}$, the manageability of $\{1\{Y_{2,i} \leq \tau\} - 1\{Y_{1,i} \leq \tau\} : \tau \in \mathcal{T}, i \leq n, n \geq 1\}$ holds by the following calculations:

$$\int_0^1 \sqrt{\log(c_1^2 \varepsilon^{-2c_2})} d\varepsilon = \int_{\sqrt{\log(A)}}^{\infty} (2A^{1/W}/W) x^2 e^{-x^2/W} dx < \infty, \tag{F.5}$$

where $A := c_1^2$, $W := 2c_2$, $\log(A) \geq 0$ because $c_1 \geq 1$, and the equality holds by change of variables with $x = \sqrt{\log(A\varepsilon^{-W})}$ or, equivalently, $\varepsilon = A^{1/W} e^{-x^2/W}$, which yields $d\varepsilon = (2A^{1/W}/W) x e^{-x^2/W} dx$. This completes the proof. $\square$

Proof of Lemma 7.2. We prove part (a) first. Assumption PS1(a) holds because $\Theta = \{0\}$. Assumptions PS1(b) and PS1(c) hold by conditions (i) and (ii) of the lemma, respectively. Assumption PS1(d) holds because

$$\begin{aligned} & |(\tau - Y_2)^{s-1} 1\{Y_2 \leq \tau\} - (\tau - Y_1)^{s-1}) 1\{Y_1 \leq \tau\}| \\ \leq\ & (\tau - Y_2)^{s-1} 1\{Y_2 \leq \tau\} + (\tau - Y_1)^{s-1}) 1\{Y_1 \leq \tau\} \\ \leq\ & (B - Y_2)^{s-1} + (B - Y_1)^{s-1}. \end{aligned} \tag{F.6}$$

Assumption PS1(e) holds because

$$M(W) \leq 2(B - (-B))^{s-1} / \sigma_{F,1}(0) \leq 2^s B^{s-1} / \underline{\sigma}. \tag{F.7}$$

Next, we verify Assumption PS2. Consider the set $\mathcal{M}_{n,1,y_1} := \{(-(\tau - y_{1,i})^{s-1}1\{y_{1,i} \leq \tau\})_{i=1}^n \in R^n : \tau \in \mathcal{T}\}$ for an arbitrary realization $\{y_{1,i} : i \leq n\}$ of the random vector $\{Y_{1,i} : i \leq n\}$. First, we show that this set has pseudo-dimension (as defined in Pollard (1990, p. 15)) at most one. Suppose not. Then, there exists a vector $x = (x_1, x_2)' \in R^2$ and a pair (i, i') such that $\{(-(\tau - y_{1,i})^{s-1}1\{y_{1,i} \leq \tau\}, (-(\tau - y_{1,i'})^{s-1}1\{y_{1,i'} \leq \tau\}) : \tau \in \mathcal{T}\}$ surrounds x.[23] Thus, there exists $\tau_1, \tau_2 \in \mathcal{T}$ such that

$$\begin{aligned}
(\tau_1 - y_{1,i})^{s-1}1\{y_{1,i} \leq \tau_1\} &> x_1, \\
(\tau_1 - y_{1,i'})^{s-1}1\{y_{1,i'} \leq \tau_1\} &< x_2, \\
(\tau_2 - y_{1,i})^{s-1}1\{y_{1,i} \leq \tau_2\} &< x_1, \text{ and} \\
(\tau_2 - y_{1,i'})^{s-1}1\{y_{1,i'} \leq \tau_2\} &> x_2. \qquad \text{(F.8)}
\end{aligned}$$

This yields

$$\begin{aligned}
(\tau_1 - y_{1,i})^{s-1}1\{y_{1,i} \leq \tau_1\} &> (\tau_2 - y_{1,i})^{s-1}1\{y_{1,i} \leq \tau_2\} \text{ and} \\
(\tau_1 - y_{1,i'})^{s-1}1\{y_{1,i'} \leq \tau_1\} &< (\tau_2 - y_{1,i'})^{s-1}1\{y_{1,i'} \leq \tau_2\}. \qquad \text{(F.9)}
\end{aligned}$$

Due to the monotonicity of the function $G_s(y, \tau) := (\tau - y)^{s-1}1\{y \leq \tau\}$ in τ for any y, the first inequality in the equation above implies that $\tau_1 > \tau_2$, and the second inequality implies that $\tau_1 < \tau_2$, which is a contradiction. Therefore, $\mathcal{M}_{n,1,y_1}$ has pseudo-dimension at most one.

The remainder of the proof of part (a) is the same as the corresponding part of the proof of Lemma 7.1 and, hence, for brevity, is omitted.

To prove part (b), consider an arbitrary sequence $\{F_n : n \geq 1\}$ such that $(0, F_n) \in \mathcal{F}_+$

[23] As defined in Pollard (1990, p. 15), a set $A \subset R^2$ *surrounds* x if there exists points $a, b, c, d \in A$, where $a = (a_1, a_2)'$ etc., such that $a_1 > x_1$, $a_2 > x_2$, $b_1 > x_1$, $b_2 < x_2$, $c_1 < x_1$, $c_2 > x_2$, $d_1 < x_1$, and $d_2 < x_2$.

for all n. Under this sequence, we have for any $\zeta > 0$ and $j = 1, 2$,

$$\begin{aligned}\Pr_{F_n}(|\overline{Y}_{j,n} - E_{F_n}(Y_j)| > \zeta) &\le \frac{E_{F_n}(\overline{Y}_{j,n} - E_{F_n}(Y_j))^2}{\zeta^2} \\ &= \frac{E_{F_n}(Y_j - E_{F_n}(Y_j))^2}{n\zeta^2} \\ &\le \frac{E_{F_n}(Y_j^2)}{n\zeta} \\ &\le B^2/(n\zeta) \to 0 \text{ as } n \to \infty, \end{aligned} \tag{F.10}$$

where the last inequality holds because the support of Y_j is contained in $\mathcal{T}$ and $\mathcal{T}$ is contained in $[-B, B]$ by condition (iii) of the lemma. Similarly, we have under the sequence $\{F_n : n \ge 1\}$,

$$n^{-1}\sum_{i=1}^{n}(Y_{j,i} - E_{F_n}(Y_j))^{2(s-1)} - E_{F_n}(Y_j - E_{F_n}(Y_j))^{2(s-1)} \to_p 0, \tag{F.11}$$

for $j = 1, 2$. Therefore, we have

$$\begin{aligned}& n^{-1}\sum_{i=1}^{n}(Y_{j,i} - \overline{Y}_{j,n})^{2(s-1)} - E_{F_n}(Y_j - E_{F_n}(Y_j))^{2(s-1)} \\ &= n^{-1}\sum_{i=1}^{n}(Y_{j,i} - E_{F_n}(Y_j))^{2(s-1)} - E_{F_n}(Y_j - E_{F_n}(Y_j))^{2(s-1)} \\ &\quad + \sum_{b=0}^{2(s-1)-1}\binom{2(s-1)}{b}\left[n^{-1}\sum_{i=1}^{n}(Y_{j,i} - E_{F_n}(Y_j))^b\right](E_{F_n}(Y_j) - \overline{Y}_{j,n})^{2s-2-b}, \\ &= o_p(1) \\ &\quad + \sum_{b=0}^{2(s-1)-1}\binom{2(s-1)}{b}\left[n^{-1}\sum_{i=1}^{n}(Y_{j,i} - E_{F_n}(Y_j))^b\right](E_{F_n}(Y_j) - \overline{Y}_{j,n})^{2s-2-b}, \\ &= o_p(1), \end{aligned} \tag{F.12}$$

where the second equality holds by (F.11), and the last equality holds by (F.10) and the boundedness of Y_j.

Therefore,

$$
\begin{aligned}
|\widehat{\sigma}_{n,1}^2(0) - \sigma_{F_n,1}^2(0)|/\sigma_{F_n,1}^2(0) &\leq \underline{\sigma}^{-2}|\widehat{\sigma}_{n,1}^2(0) - \sigma_{F_n,1}^2(0)| \\
&\leq \underline{\sigma}^{-2}\sum_{j=1}^{2}\left|n^{-1}\sum_{i=1}^{n}(Y_{j,i} - \overline{Y}_{j,n})^{2(s-1)} - E_{F_n}(Y_j - E_{F_n}(Y_j))^{2(s-1)}\right| \\
&\to_p 0, \text{ as } n \to \infty. \qquad \text{(F.13)}
\end{aligned}
$$

Because this holds for an arbitrary sequence $\{F_n : n \geq 1\}$ such that $(0, F_n) \in \mathcal{F}_+$, it establishes both Assumptions SIG1 and SIG2. Thus, part (b) holds. □

Proof of Lemma 8.1. We show that parts (a) and (b) are equivalent by solving a linear programming problem. We show that parts (b) and (c) are equivalent by employing the convex polyhedral cone representation of linear inequalities developed in Gale (1951).

First, we show the equivalence between parts (a) and (b).

For a set $A \subset R^{d_\beta}$, let A^c denote the complement of A in R^{d_β}. By basic set operations, the statement in part (a) is equivalent to

$$\cap_{j=1}^{m} H(c_j)^c \subset H(\overline{c})^c. \qquad \text{(F.14)}$$

Because m is finite and $H(c)^c$ is an open set for any $c \in R^{d_\beta}\backslash\{0\}$, (F.14) is equivalent to

$$\cap_{j=1}^{m} cl(H(c_j)^c) \subset cl(H(\overline{c})^c). \qquad \text{(F.15)}$$

Note that $cl(H(c)^c) = \{b \in R^{d_\beta} : b'c \leq 0\}$ and (F.15) is equivalent to the redundancy of the inequality restriction $b'\overline{c} \leq 0$ on b relative to the system of linear inequalities $b'c_j \leq 0$ for $j = 1, ..., m$. In turn, the latter is equivalent to the statement that $V = 0$, where

$$V := \max_{b \in R^{d_\beta}} b'\overline{c} \text{ subject to } b'c_j \leq 0 \text{ for } j = 1, ..., m \text{ and } b'\overline{c} \leq 1. \qquad \text{(F.16)}$$

Now we solve the linear programming problem in (F.16) using the Lagrange multiplier method. It is well known that

$$V = \min_{\lambda_j \geq 0 : j=1,...,m+1} \max_{b \in R^{d_\beta}} \left(b'\overline{c} - \sum_{j=1}^{m}\lambda_j b'c_j - \lambda_{m+1}(b'\overline{c} - 1)\right). \qquad \text{(F.17)}$$

Because the maximization over b is unconstrained, for any $\lambda_1, ..., \lambda_{m+1} \geq 0$ such that $(1 - \lambda_{m+1})\overline{c} - \sum_{j=1}^m \lambda_j c_j \neq 0$, the maximum is infinite. But, $V \leq 1$ by the inequality $b'\overline{c} \leq 1$ in (F.16). Thus, the optimal $\lambda_1, ..., \lambda_{m+1}$ must satisfy

$$(1 - \lambda_{m+1})\overline{c} - \sum_{j=1}^{m} \lambda_j c_j = 0. \tag{F.18}$$

This implies that

$$V = \min_{\lambda_j \geq 0 : j=1,...,m+1} \lambda_{m+1} \text{ subject to } (1 - \lambda_{m+1})\overline{c} - \sum_{j=1}^{m} \lambda_j c_j = 0. \tag{F.19}$$

Now we show that $V = 0$ iff there exist $\lambda_1, ..., \lambda_m \geq 0$ such that

$$\overline{c} = \sum_{j=1}^{m} \lambda_j c_j, \tag{F.20}$$

which establishes the equivalence between parts (a) and (b). Suppose that there exists $\lambda_1, ..., \lambda_m \geq 0$ such that (F.20) holds, then $V \leq \min\{\lambda_{m+1} \geq 0 : \lambda_{m+1}\overline{c} = 0^{d_\beta}\} = 0$ by (F.19). However, by (F.16) (with $b = 0^{d_\beta}$), $V \geq 0$. Thus, $V = 0$. Conversely, suppose that $V = 0$, then there exists $\lambda_1, ..., \lambda_m \geq 0$ such that $(1 - 0)\overline{c} - \sum_{j=1}^m \lambda_j c_j = 0^{d_\beta}$ by (F.19), which implies (F.20).

Next, we establish the equivalence between parts (b) and (c). Using the terminology in Gale (1951), let $P(c_1, ..., c_m)$ denote the convex polyhedral cone generated by the vectors $(c_1, ..., c_m)$. That is,

$$P(c_1, ..., c_m) := \left\{ c \in R^{d_\beta} : c = \sum_{j=1}^{m} \lambda_j c_j \text{ for some } \lambda_1, ..., \lambda_m \geq 0 \right\}.$$

Then, part (b) is equivalent to $\overline{c} \in P(c_1, ..., c_m)$.

If $rk([c_1, ..., c_m]) = d_\beta$, then by Weyl's Theorem (see Theorem 1 of Gale (1951)), $P(c_1, ..., c_m)$ is an intersection of at most $\binom{m}{d_\beta - 1}$ half-spaces or, in other words, there exist $b^1, ..., b^N \in R^{d_\beta}$, where $N := \binom{m}{d_\beta - 1}$, such that $P(c_1, ..., c_m) = \{c : [b^1, \cdots, b^N]'c \geq 0\}$. Then, the equivalence between parts (b) and (c) is established with $B(c_1, ..., c_m) := [b^1, \cdots, b^N, 0^{d_\beta \times (M-N)}]$ for an arbitrary $[b^1, \cdots, b^N]$ that satisfies $P(c_1, ..., c_m) = \{c : [b^1, \cdots, b^N]'c \geq$

$0\}$.

If $rk([c_1, ..., c_m]) < d_\beta$, let $L(c_1, ..., c_m)$ be the linear subspace spanned by $c_1, ..., c_m$. Let the dimension of this linear subspace be d_L. Applying Weyl's Theorem on $L(c_1, ..., c_m)$, we have that there exist $b^1, ..., b^{N_1} \in R^{d_\beta}$, where $N_1 := \binom{m}{d_L - 1}$, such that $P(c_1, ..., c_m) = \{c : [b^1, \cdots, b^{N_1}]'c \geq 0\} \cap L(c_1, ..., c_m)$. Moreover, by the property of linear subspaces, there exist $b^{N_1+1}, ..., b^{N_2} \in R^{d_\beta}$, where $N_2 := N_1 + d_\beta - d_L$, such that $L(c_1, ..., c_m) = \{c : [b^{N_1+1}, \cdots, b^{N_2}]'c = 0\}$. Therefore,

$$P(c_1, ..., c_m) = \{c : [b^1, \cdots, b^{N_2}, -b^{N_1+1}, \cdots, -b^{N_2}]'c \geq 0\}. \tag{F.21}$$

Then, the equivalence between parts (b) and (c) holds with $B(c_1, ..., c_m) := [b^1, \cdots, b^{N_2}, -b^{N_1+1}, \cdots, -b^{N_2}, 0^{d_\beta \times (M - 2N_2 + N_1)}]$ for arbitrary $b^1, ..., b^{N_2}$ that satisfy (F.21). $\square$

Proof of Lemma 8.2. Assumption PS1(a) holds because $\theta \in \Theta$. Assumptions PS1(b) holds by the i.i.d. condition in the lemma. Assumption PS1(c) holds by $\sigma^2_{F,1}(\theta) = 1$. Assumption PS1(d) holds because $|F_\beta(\mathcal{S}, \theta) - 1\{S(Y_1, Y_2, X_1) \subset \mathcal{S}\}| \leq 1$. Assumption PS1(e) holds because

$$E_F M^{2+\delta}(W) = 1/\sigma^{2+\delta}_{F,1}(0) = 1. \tag{F.22}$$

Next, we verify Assumption PS2. Consider the set $\mathcal{M}_n := \{(F_\beta(\mathcal{S}(\tau), \theta) - 1\{(y_{1i} - 1/2)B(\tau)'(1, x'_{1i}, y'_{2i})' \geq 0\})_{i=1}^n \in R^n : \tau = (c_1, \ldots, c_m), c_1, \ldots, c_m \in R^{d_\beta} \backslash \{0^{d_\beta}\}\}$ for an arbitrary realization $\{(y_{1,i}, y'_{2,i}, x'_{1,i})' : i \leq n\}$ of $\{(Y_{1,i}, Y'_{2,i}, X'_{1,i})' : i \leq n\}$. The set has pseudo-dimension at most M by Lemma 4.4 of Pollard (1990). Then, by Corollary 4.10 of Pollard (1990), there exist constants $c_1 \geq 1$ and $c_2 > 0$ (not depending on n, $\{(y_{1,i}, y'_{2,i}, x'_{1,i})' : i \leq n\}$, or ε) such that

$$D(\varepsilon\|\alpha\|, \alpha \odot \mathcal{M}_n) \leq c_1 \varepsilon^{-c_2} \tag{F.23}$$

for all $0 < \varepsilon \leq 1$ and every rescaling vector $\alpha \in R^n_+$. In consequence, the manageability of $\{(F_\beta(\mathcal{S}(\tau), \theta) - 1\{Y_{1,i} - 1/2)B(\tau)'(1, X'_{1,i}, Y'_{2,i})' \geq 0\})_{i=1}^n \in R^n : \tau = (c_1, \ldots, c_m), c_1, \ldots, c_m \in R^{d_\beta} \backslash \{0^{d_\beta}\}\}$ follows from the calculations in (F.5) with $A := c_1$ and $W := c_2$, which completes the proof. $\square$

Proof of Lemma 9.1 Assumptions PS1(a)-(c) and (e) hold by assumption. Next, we show

Assumption PS1(d) holds. We have

$$|h(Q_\theta(w,v),u)| = |\sup_{q\in Q_\theta(w,v)} q'u| \leq \sup_{q\in Q_\theta(w,v)} \|q\|\|u\| \leq \sup_{q\in Q_\theta(w,v)} \|q\| \leq M(w)/2, \qquad \text{(F.24)}$$

where the first inequality holds by the Cauchy-Schwarz inequality, the second inequality holds because u satisfies $\|u\| \leq 1$, and the last inequality holds by condition (iii) of the lemma. Assumption PS1(d) follows from the following calculations:

$$\begin{aligned}
\left|\int h(Q_\theta(W,v),u)dF_{V|X}(v,X;\theta) - u'q(X)\right| &\leq \int |h(Q_\theta(W,v),u)|dF_{V|X}(v,X;\theta) + |u'q(X)| \\
&\leq M(W)/2 + |u'q(X)| \\
&\leq M(W)/2 + \|u\|\|q(X)\| \\
&\leq M(W), \qquad \text{(F.25)}
\end{aligned}$$

where the second inequality holds by (F.24), the third inequality holds by the Cauchy-Schwarz inequality and the last inequality holds by $\|u\| \leq 1$ and condition (iv) of the lemma.

Now, we show that Assumption PS2 holds. Let $m(W,\theta,u) := \int h(Q_\theta(W,v),u)\, dF_{V|X}(v,X;\theta) - u'q(X)$. Consider an arbitrary sequence (θ_n, F_n) that satisfy the conditions in the lemma. Arguments similar to those for Assumption PS1(d) above show that $m(W,\theta,u)$ is Lipschitz continuous in u with Lipschitz constant $M(W)$ for all θ. Given the Lipschitz continuity, for any nonnegative n-vector $\alpha := (\alpha_1,\ldots,\alpha_n)$, any $u_1 \in R^d$, $u_2 \in R^d$ such that $\|u_1\| \leq 1$ and $\|u_2\| \leq 1$, and any n realizations $(w_1,\ldots,w_n)$ of W (under F_n), we have

$$\sum_{i=1}^n (\alpha_i m(w_i,\theta_n,u_1) - \alpha_i m(w_i,\theta_n,u_2))^2 \leq \left(\sum_{i=1}^n (\alpha_i M(w_i))^2\right) \|u_1-u_2\|^2. \qquad \text{(F.26)}$$

Let $\mathcal{F}_{n(w_1,\ldots,w_n)} = \{(m(w_i,\theta_n,u)_{i=1}^n : u \in R^d, \|u\| \leq 1\}$ and let $\vec{M}_n(w_1,\ldots,w_n) = (M(w_1),\ldots, M(w_n))'$. Then, (F.26) implies that, for all $\xi \in (0,1]$,

$$D(\xi\|\alpha \odot \vec{M}_n(w_1,\ldots,w_n)\|, \alpha \odot \mathcal{F}_{n(w_1,\ldots,w_n)}) \leq D(\xi, \{u \in R^d : \|u\| \leq 1\}) \leq C/\xi^d, \qquad \text{(F.27)}$$

for some constant $C < \infty$. Assumption PS2 holds because $\int_0^1 \sqrt{\log C - d\log\xi}\, d\xi = 2C^{1/d}d^{-1}$

$\int_0^\infty x^2 e^{-x^2/d} dx < \infty$. $\square$

Proof of Lemma 9.2. We prove part (a) first. Assumptions PS1(a)-(e) hold by assumption. Now we verify Assumption PS2.

Let (θ_n, F_n) be an arbitrary sequence that satisfies all the conditions of the lemma. Consider n arbitrary realizations $(w_1, \ldots, w_n)$ of W under F_n for arbitrary $n \geq 1$. By (9.2) and conditions (iii) and (iv) of the lemma, $\{M(w_i) : i \leq n, n \geq 1\}$ are envelopes for the triangular array of processes $\{m(w_i, \theta_n, \tau) : i \leq n, n \geq 1, \tau = 1, 2, \ldots\}$. Let

$$\mathcal{F}_{n(w_1,\ldots,w_n)} = \{(m(w_1, \theta_n, \tau), \ldots, m(w_n, \theta_n, \tau))' : \tau = 1, 2, \ldots\}. \tag{F.28}$$

Let $\vec{M}_n(w_1, \ldots, w_n) = (M(w_1), \ldots, M(w_n))'$. Then, for any $\xi \in (0, 1]$ and any nonnegative n-vector α, $D(\xi \|\alpha \odot \vec{M}_n(w_1, \ldots, w_n)\|, \alpha \odot \mathcal{F}_{n(w_1,\ldots,w_n)}) \leq \lambda_{\mathcal{T}}(\xi)$ because $\alpha \odot (m(w_1, \theta_n, \tau), \ldots,$ $m(w_n, \theta_n, \tau))'$ belongs to the $\xi\|\alpha \odot \vec{M}_n(w_1, \ldots, w_n)\|$-neighborhood of 0^n for all $\tau \geq \lambda_{\mathcal{T}}(\xi)$. The latter holds because, for all $\tau \geq \lambda_{\mathcal{T}}(\xi)$,

$$\begin{aligned}\sum_{i=1}^n (\alpha_i m(w_i, \theta, \tau))^2 &= w_{\mathcal{T}}^2(\tau) \sum_{i=1}^n (\alpha_i \widetilde{m}(w_i, \theta, \tau))^2 \\ &\leq w_{\mathcal{T}}^2(\tau) \sum_{i=1}^n (\alpha_i M(w_i))^2 \\ &\leq \xi^2 \|\alpha \odot \vec{M}_n(w_1, \ldots, w_n)\|^2,\end{aligned} \tag{F.29}$$

where the last inequality holds because $\tau \geq \lambda_{\mathcal{T}}(\xi)$ iff $w_{\mathcal{T}}(\tau) \leq \xi$ (because $\lambda_{\mathcal{T}}(\xi)$ is the inverse function of the decreasing function $w_{\mathcal{T}}(\xi)$). By assumption, $\int_0^1 \sqrt{\log(\lambda_{\mathcal{T}}(\xi))} d\xi < \infty$. Hence, Assumption PS2 holds.

The proof of part (b) is similar to that of Lemma 7.2(b) and is omitted for brevity. $\square$

G Additional Simulation Results

In this section we report additional Monte Carlo simulation results that investigate the sensitivity of the performance of the MCMI tests to different choices of tuning parameters.

Tables 3-6 report the null and alternative hypothesis rejection probabilities of the CvM/GMS and KS/GMS tests in the first-order stochastic dominance example studied in Section 7.2 for various choices of the tuning parameters. The first two tables cover null data generating processes for which first-order stochastic dominance holds. The last two tables cover alternative data generating processes for which first-order stochastic dominance does not hold. These data generating processes are the same as those considered in Tables 1 and 2.

In each table, we report the rejection probabilities of the nominal .05 level CvM/GMS and KS/GMS tests for a base case specification of the tuning parameters and 12 variants of the base case. The base case sets $r_{1,n} = 3$, $N_\tau = 25$, $\kappa_n = (0.3 \ln(n))^{1/2}$, $B_n = (0.4 \ln(n)/\ln(\ln(n)))^{1/2}$, $\varepsilon = 0.01$, and Weight Constant $= 100$, where the weight constant is the constant added to r^2 in (3.7). The larger the weight constant is, the more weight is given to smaller hypercubes relative to the larger cubes. In each variant of the base case, one and only one tuning parameter is changed in order to isolate the effect of that tuning parameter on the performance of the tests. The power results in Tables 5 and 6 are size-corrected based on the data generating process for Table 3 under which the tests have asymptotic null rejection probabilities equal to .05.

As one can see from the tables, the base case specification performs well compared to the other specifications. Overall, the sensitivity to the choice of tuning parameters is not large for the range of tuning parameter values that is considered.

Table 3: Sensitivity to Tuning Parameters for First-Order Stochastic Dominance Tests – Null 1: $(c_1, c_2, c_3, c_4) = (0, 0, 0.85, 0.6)$

	CvM/GMS	KS/GMS
Base case	.057	.064
$r_{1,n} = 2$	.056	.054
$r_{1,n} = 4$	.059	.055
$N_\tau = 20$	.061	.074
$N_\tau = 30$	.062	.068
$\kappa_n = (0.3\ln(n))^{1/2}/2$	.070	.077
$\kappa_n = 2(0.3\ln(n))^{1/2}$	.052	.055
$B_n = (0.4\ln(n)/\ln(\ln(n)))^{1/2}/2$	.057	.064
$B_n = 2(0.4\ln(n)/\ln(\ln(n)))^{1/2}$	.057	.064
$\varepsilon = 0.005$	.056	.065
$\varepsilon = 0.02$	.058	.056
Weight Constant = 50	.056	.067
Weight Constant = 200	.057	.064

Note: Sample size $n = 250$. The critical values use 1000 repetitions. Nominal size = .05. Base case specifications: $r_{1,n} = 3$, $N_\tau = 25$, $\kappa_n = (0.3\ln(n))^{1/2}$, $B_n = (0.4\ln(n)/\ln(\ln(n)))^{1/2}$, $\varepsilon = 0.01$, Weight Constant = 100.

Table 4: Sensitivity to Tuning Parameters for First-Order Stochastic Dominance Tests – Null 2: $(c_1, c_2, c_3, c_4) = (0.15, 0, 0.85, 0.6)$

	CvM/GMS	KS/GMS
Base case	.014	.019
$r_{1,n} = 2$	.016	.018
$r_{1,n} = 4$	.014	.019
$N_\tau = 20$	.015	.015
$N_\tau = 30$	.018	.019
$\kappa_n = (0.3\ln(n))^{1/2}/2$	.026	.029
$\kappa_n = 2(0.3\ln(n))^{1/2}$	.010	.011
$B_n = (0.4\ln(n)/\ln(\ln(n)))^{1/2}/2$	.013	.019
$B_n = 2(0.4\ln(n)/\ln(\ln(n)))^{1/2}$	.016	.019
$\varepsilon = 0.005$	.014	.018
$\varepsilon = 0.02$	.014	.017
Weight Constant = 50	.014	.019
Weight Constant = 200	.014	.019

Note: Sample size $n = 250$. The critical values use 1000 repetitions. Nominal size = .05. Base case specifications: $r_{1,n} = 3$, $N_\tau = 25$, $\kappa_n = (0.3\ln(n))^{1/2}$, $B_n = (0.4\ln(n)/\ln(\ln(n)))^{1/2}$, $\varepsilon = 0.01$, Weight Constant = 100.

Table 5: Sensitivity to Tuning Parameters for First-Order Stochastic Dominance Tests – Alternative 1: $(c_1, c_2, c_3, c_4) = (-0.25, 0.2, 0.85, 0.6)$

	CvM/GMS	KS/GMS
Base case	.505	.379
$r_{1,n} = 2$	.513	.411
$r_{1,n} = 4$	.509	.367
$N_\tau = 20$	.486	.359
$N_\tau = 30$	.470	.405
$\kappa_n = (0.3\ln(n))^{1/2}/2$	.493	.363
$\kappa_n = 2(0.3\ln(n))^{1/2}$	.507	.384
$B_n = (0.4\ln(n)/\ln(\ln(n)))^{1/2}/2$	.508	.379
$B_n = 2(0.4\ln(n)/\ln(\ln(n)))^{1/2}$	.506	.379
$\varepsilon = 0.005$	.505	.381
$\varepsilon = 0.02$	.504	.392
Weight Constant = 50	.505	.379
Weight Constant = 200	.506	.379

Note: Sample size $n = 250$. The critical values use 1000 repetitions. Nominal size = .05. Base case specifications: $r_{1,n} = 3$, $N_\tau = 25$, $\kappa_n = (0.3\ln(n))^{1/2}$, $B_n = (0.4\ln(n)/\ln(\ln(n)))^{1/2}$, $\varepsilon = 0.01$, Weight Constant = 100.

Table 6: Sensitivity to Tuning Parameters for First-Order Stochastic Dominance Tests – Alternative 2: $(c_1, c_2, c_3, c_4) = (0.35, 0, 0.85, 0.23)$

	CvM/GMS	KS/GMS
Base case	.581	.295
$r_{1,n} = 2$	.628	.397
$r_{1,n} = 4$	.609	.246
$N_\tau = 20$	.560	.275
$N_\tau = 30$	.539	.309
$\kappa_n = (0.3\ln(n))^{1/2}/2$	.590	.318
$\kappa_n = 2(0.3\ln(n))^{1/2}$	.463	.159
$B_n = (0.4\ln(n)/\ln(\ln(n)))^{1/2}/2$	.532	.258
$B_n = 2(0.4\ln(n)/\ln(\ln(n)))^{1/2}$	.589	.301
$\varepsilon = 0.005$	.615	.302
$\varepsilon = 0.02$	.529	.268
Weight Constant = 50	.580	.295
Weight Constant = 200	.582	.295

Note: Sample size $n = 250$. The critical values use 1000 repetitions. Nominal size = .05. Base case specifications: $r_{1,n} = 3$, $N_\tau = 25$, $\kappa_n = (0.3\ln(n))^{1/2}$, $B_n = (0.4\ln(n)/\ln(\ln(n)))^{1/2}$, $\varepsilon = 0.01$, Weight Constant = 100.

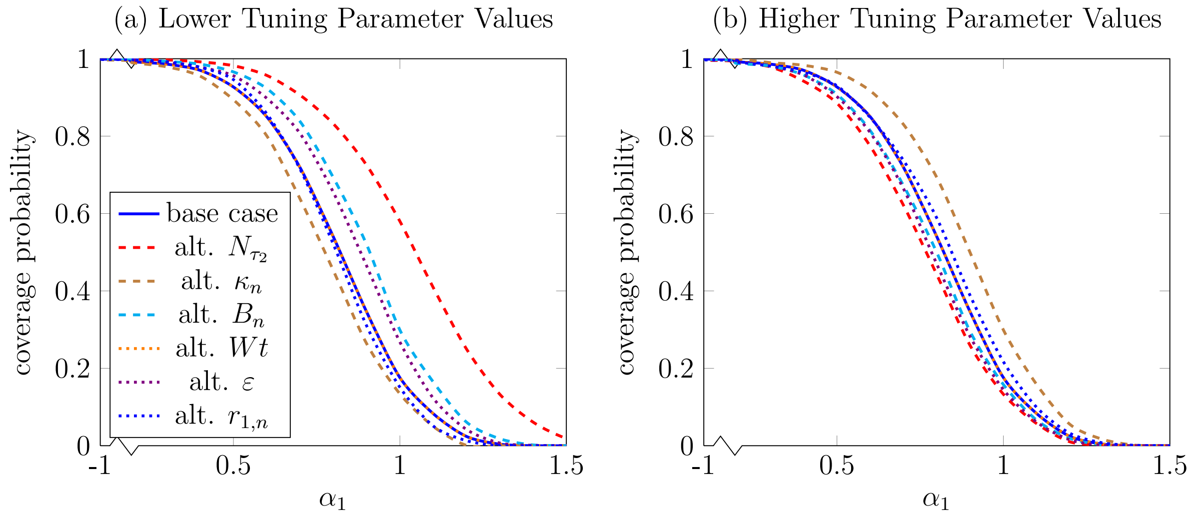

Figure 3: Sensitivity to Tuning Parameters for the KS/GMS CS Applied on the IV Random-Coefficients Binary-Outcome Model. (Nominal size = .95, $n = 250$, $(\alpha_0, \gamma_0, \gamma_1) = (0, -1, 1)$, and $(\alpha_0, \alpha_1, \gamma_0, \gamma_1)$ is in the identified set if and only if $\alpha_1 \leq -0.8274$.)

In the random-coefficient binary-outcome example studied in Section 8.2, the KS/GMS CS is the best performing CS. For this reason, we perform sensitivity analysis on this CS. Figure 3 reports the coverage probabilities of the KS/GMS CS under various choices of the tuning parameters. The base case values for $r_{1,n}$, κ_n, B_n, ε, and the Weight Constant are the same as they are in the stochastic dominance example. The base case value for N_{τ_2} is 15, which results in 120 grid points on $\mathcal{T}$. In the sensitivity analysis, we alter one and only one tuning parameter each time, and recompute the coverage probabilities.

Graph (a) of Figure 3 depicts the coverage probabilities when the tuning parameters are altered to lower values. The lower values are: $r_{1,n} = 2$, $N_{\tau_2} = 10$, $\kappa_n = (0.3\ln(n))^{1/2}/2$, $B_n = (0.4\ln(n)/\ln(\ln(n)))^{1/2}/2$, $\varepsilon = 0.005$, and Weight Constant = 50. Graph (b) depicts the coverage probabilities when they are altered to higher values. The higher values are: $r_{1,n} = 4$, $N_{\tau_2} = 20$, $\kappa_n = 2(0.3\ln(n))^{1/2}$, $B_n = 2(0.4\ln(n)/\ln(\ln(n)))^{1/2}$, $\varepsilon = 0.02$, and Weight Constant = 200.

Recall that the coverage probabilities are computed for fixed values of $(\alpha_0, \alpha_1, \gamma_0, \gamma_1)$. The fixed values considered are $(0, \alpha_1, -1, 1)$ for α_1 running from -1 (its true value), to -0.8274 (right boundary of the identified set at $(\alpha_0, \gamma_0, \gamma_1) = (0, -1, 1)$), and finally to 1.5. The coverage probabilities when $\alpha_1 \leq -0.8274$ should ideally be 0.95 or higher, and those when $\alpha_1 > -0.8274$ are false coverage probabilities, and should ideally be lower than 0.95 and get lower as α_1 moves to the right.

As one can see, the base case specification performs well relative to the other specifications. The coverage probabilities are generally insensitive to the tuning parameter changes considered, except when N_{τ_2} is changed. Coarser τ grids result in noticeably higher (worse) false coverage probabilities.

The sensitivity analysis sheds some light on the choice of the grid points for approximating $\mathcal{T}$. In both examples, we find that the size of the test or the coverate probability of the CS for points in the identified set is insensitive to $\mathcal{T}$, but the power or false coverage probability is noticeably worse when an insufficient number of grid points is used. Based on this, we recommend increasing the number of grid points until the test statistic value stabilizes, or increasing it to the maximum that the computational resources of the user allows.

References

Andrews D. W. K., Shi, X., 2013a. Inference based on conditional moment inequalities. Econometrica 81, 609–666.

Andrews D. W. K., Shi, X., 2013b. Supplemental material to 'Inference based on conditional moment inequalities.' Available at Econometrica Supplemental Material 81, http://www.econometricsociety.org/ecta/supmat/9370_Proofs.pdf.

Andrews D. W. K., Shi, X., 2014. Nonparametric inference based on conditional moment inequalities. Journal of Econometrics 179, 31–45. Appendix of proofs available as Cowles Foundation Discussion Paper No. 1840RR, Yale University, 2011.

Chernozhukov, V., D. Chetverikov, and K. Kato, 2014. Testing many moment inequalities. Department of Economics, MIT. Unpublished manuscript.

Gale, D., 1951. Convex polyhedral cones and linear inequalities. Chapter XVII in Activity Analysis of Production and Allocation, ed. by T. C. Koopmans, pp. 287–297. Wiley, New York.

Pollard, D., 1990. Empirical Process Theory and Application. NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. II. Institute of Mathematical Statistics.