NONPARAMETRIC INFERENCE BASED
ON CONDITIONAL MOMENT INEQUALITIES

by

Donald W.K. Andrews and Xiaoxia Shi

COWLES FOUNDATION PAPER NO. 1405

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

2014

http://cowles.econ.yale.edu/
Nonparametric inference based on conditional moment inequalities

Donald W.K. Andrews \(^{a,\ast}\), Xiaoxia Shi \(^{b}\)

\(^{a}\) Cowles Foundation for Research in Economics, Yale University, United States
\(^{b}\) Department of Economics, University of Wisconsin at Madison, United States

**Article info**

Article history:
Received 20 December 2011
Received in revised form 12 August 2013
Accepted 9 October 2013
Available online 21 October 2013

JEL classification:
C12
C15

Keywords:
Asymptotic size
Kernel
Local power
Moment inequalities
Nonparametric inference
Partial identification

**Abstract**

This paper develops methods of inference for nonparametric and semiparametric parameters defined by conditional moment inequalities and/or equalities. The parameters need not be identified. Confidence sets and tests are introduced. The correct uniform asymptotic size of these procedures is established. The false coverage probabilities and power of the CS’s and tests are established for fixed alternatives and some local alternatives. Finite-sample simulation results are given for a nonparametric conditional quantile model with censoring and a nonparametric conditional treatment effect model. The recommended CS/test uses a Cramér–von-Mises-type test statistic and employs a generalized moment selection critical value.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

This paper considers inference for nonparametric and semiparametric parameters defined by conditional moment inequalities and/or equalities. The moments are conditional on \(X_i\) a.s. and \(Z_i = z_0\) for some random vectors \(X_i\) and \(Z_i\). The parameters need not be identified. Due to the conditioning on \(Z_i\) at a single point \(z_0\), the parameter considered is a nonparametric or semiparametric parameter (which varies with \(z_0\)). Due to the conditioning on \(X_i\) a.s., the moment conditions are typical conditional moments which involve an infinite number of restrictions.

Examples covered by the results of this paper include: a nonparametric conditional distribution with selection, a nonparametric conditional quantile with selection, an interval-outcome partially-linear regression, a semiparametric discrete-choice model with multiple equilibria, a nonparametric revealed preference model, tests of a variety of functional inequalities, including nonparametric average treatment effects for certain sub-populations, and nonparametric binary Roy models, as in Henry and Mourifié (2012).

As far as we are aware, the only other paper in the literature that covers the examples described above is Chernozhukov et al. (2013) (CLR). In this paper, we employ statistics that are akin to Bierens (1982)-type model specification test statistics. In contrast, CLR employ statistics that are akin to Härdle and Mammen (1993)-type model specification statistics, which are based on nonparametric regression estimators. These approaches have different strengths and weaknesses. Specifically, the tests proposed in this paper have higher power against conditional moment functions that are relatively flat (but not necessarily completely flat) as a function of \(x\), whereas the CLR tests have higher power against conditional moment functions that are more curved. This is shown by the finite-sample simulations reported here and the asymptotic local power results reported in the Appendix, see Andrews and Shi (2013a).

For example, flat conditional moment functions arise in models with moment equalities, as well as inequalities, such as entry games with complete information and pure strategy equilibrium (which is Example 4 in Section 2.2). In addition, relatively flat bounds arise in models with censoring (see Example 1 in Section 2.2) when the censoring effect is small for large values of the conditioning variable \(X_i\). In this case, for a continuum of values of \(X_i\) at the top end of its distribution, both inequalities are close to binding and the bounds are relatively flat. Such a data structure is what the “identification at infinity” strategy for censoring models relies upon, e.g., see Lewbel (2007). Lastly, relatively flat conditional moment functions arise in any model with weak instrumental/conditioning variables \(X_i\) (in the sense that \(X_i\) has low correlation with the stochastic moment functions). Weak instruments are known to arise in a variety of economic models.
We provide confidence sets (CS’s) and tests concerning the true parameter. The class of test statistics used in this paper are like those used in Andrews and Guggenberger (2009), which are extended in Andrews and Shi (2013b,c) (AS1, AS2) to handle moment conditions that are conditional on $X_i$ a.s. Here the test statistics are extended further to cover moment conditions that are conditional on $Z_i = z_0$ as well. The latter conditioning is accomplished using kernel smoothing. The critical values considered here are generalized moment selection (GMS) and plug-in asymptotic (PA) critical values, as in Andrews and Soares (2010), which are extended to cover conditional moment inequalities, as in AS1 and AS2.

The results of the paper are analogous to those in AS1 and AS2. In particular, we establish the correct uniform asymptotic size of the CS’s and tests. We also determine the asymptotic behavior of the CS’s and tests under fixed alternatives and some local alternatives.

We provide finite-sample simulation results for two models: a nonparametric conditional quantile model with selection and a nonparametric conditional treatment effect model. The conclusions from the finite-sample results are similar in many respects to those from Andrews and Soares (2010) and Andrews and Barrick (2012), AS1, and AS2. Cramer–von-Mises (CvM) versions of the CS’s and tests out-perform Kolmogorov–Smirnov (KS) versions in terms of false-coverage probabilities (FCP’s) and power and have similar size properties. Likewise, GMS critical values out-perform PA critical values according to the same criteria. The “Gaussian asymptotic” versions of the critical values perform similarly to the bootstrap versions in terms of size, FCP’s, and power. The finite-sample sizes of the CvM/GMS CS’s and tests are close to their nominal size. The CS’s and tests show some sensitivity to the nonparametric smoothing parameter employed, but not much sensitivity to other tuning parameters.

In the simulation results for these two models, the CI’s and tests proposed in this paper are found to have more robust size properties than the series and local linear CLR procedures. The CI’s and tests proposed in this paper are found to have higher power (and lower FCP’s) for flat bound functions and lower power (and higher FCP’s) for peaked bound functions compared to the CLR procedures.

We note that the results given here also apply to nonparametric models based on moments that are unconditional on $X_i$ but conditional on $Z_i = z_0$. The results also cover the case where different moment functions depend on different sub-vectors of $X_i$, e.g., as occurs in some panel models. In addition, the results can be extended to the case of an infinite number of moment functions along the lines of Andrews and Shi (2010).

The technical results in this paper differ from those in AS1 and AS2 because (i) the conditional moment inequalities (when evaluated at the true parameter) do not necessarily hold for values $Z_i$ that are in a neighborhood of $z_0$, but do not equal $z_0$, and (ii) the sample moments do not satisfy a functional CLT with $n^{1/2}$-norming due to local smoothing, and, hence, need to be normalized using their standard deviations which are $o(1)$ as $n \rightarrow \infty$.

Now, we discuss the related literature. The literature on inference based on unconditional moment inequalities for parameters that are partially identified is now quite large. For brevity, we do not give references here. See Andrews and Soares (2010) for references. The literature on inference for partially-identified models based on conditional moment inequalities includes AS1, AS2, CLR, Fan (2008), Kim (2008), Ponomareva (2010), Armstrong (2011a,b), Beresteau et al. (2011), Chetverikov (2011), Fan and Park (2011), Hsu (2011), Lee et al. (2013), Aradillas-López et al. (2013), and Armstrong and Chan (2013). Khan and Tamer (2009) considers conditional moment inequalities in a point-identified model. Galichon and Henry (2009) considers a testing problem with an infinite number of unconditional moment inequalities of a particular type. Menzel (2009) investigates tests based on a finite number of moment inequalities in which the number of inequalities increases with the sample size.

Of these papers, the only one that allows for conditioning on $Z_i = z_0$, which is the key feature of the present paper, is CLR. As noted above, the forms of the tests considered here and in CLR differ. Other differences are as follows. The assumptions given here are primitive, whereas those in CLR are high-level. The present paper provides uniform asymptotic size results, whereas CLR does not.

The remainder of the paper is organized as follows. Section 2 describes the nonparametric model and discusses six examples covered by the model. Section 3 introduces the test statistics and critical values, establishes the correct asymptotic size (in uniform sense) of the CS’s, and establishes the power of the tests against fixed alternatives. Section 4 provides Monte Carlo simulation results for two models.

An Appendix provides proofs of all of the results stated in the paper. For brevity, the Appendix is given in Andrews and Shi (2013a). The results in the Appendix allow for a much broader range of test statistics than is considered in the paper. Specifically, the results cover a wide variety of kernel functions $K$, test statistic functions $S$, instrumental functions $g \in g_i$, and weight measures $Q$. The Appendix provides two sets of results for local alternatives. The first set considers $(nb^{d_k})^{-1/2}-local alternatives, for which the bound functions are asymptotically flat near their minimum, where $d_k$ denotes a bandwidth parameter and $d_k$ denotes the dimension of $Z_i$. The tests proposed in this paper have non-trivial power against such alternatives, whereas the tests of CLR do not. The second set considers all local alternatives, for which the bound functions are asymptotically non-flat near their minimum. Here, $a_n \rightarrow 0$ as $n \rightarrow \infty$ at a rate slower than $(nb^{d_k})^{-1/2}$. For such alternatives, if the functions are sufficiently curved then the CLR tests have higher asymptotic local power than the tests considered here. On the other hand, if the functions are less curved, then the tests proposed here have higher asymptotic power than the CLR tests. The Appendix also gives some additional simulation results for the two models considered in the paper.

2. Nonparametric conditional moment inequalities and equalities

2.1. Model

The nonparametric conditional moment inequality/equality model is defined as follows. We suppose there exists a true parameter $\theta_0 \in \Theta \subset \mathbb{R}^b$ that satisfies the moment conditions:

$$E_{\theta_0} (m_i (W_i, \theta_0) | X_i, Z_i = z_0) \geq 0$$

a.s. for $j = 1, \ldots, p$ and

$$E_{\theta_0} (m_j (W_i, \theta_0) | X_i, Z_i = z_0) = 0$$

a.s. for $j = p + 1, \ldots, p + v$, (2.1)

where $\theta_0 (\cdot, \theta)$ for $j = 1, \ldots, p + v$ are (known) real-valued moment functions, $W_i = (Y_i, X_i', Z_i', \cdot)$ are observed iid random vectors with distribution $F_0$, $F_0$ is the marginal distribution of $X_i \in \mathbb{R}^{d_b}$, $Z_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}^h$, and $W_i \in \mathbb{R}^{d_h}$.

The object of interest is a CS for the true parameter $\theta_0$. We do not assume that $\theta_0$ is point identified. However, the model restricts the true parameter value to the identified set (which could be a singleton) that is defined as follows:

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

We are interested in CS’s that cover the true value $\theta_0$ with probability greater than or equal to $1 - \alpha$ for $\alpha \in (0, 1)$. As is standard,
we construct such CS’s by inverting tests of the null hypothesis that \( \theta \) is the true value for each \( \theta \in \Theta \). Let \( T_{\alpha}(\theta) \) be a test statistic and \( c_{n, 1 - \alpha}(\theta) \) be a corresponding critical value for a test with nominal significance level \( \alpha \). Then, a nominal level \( 1 - \alpha \) CS for the true value \( \theta_0 \) is
\[
\text{CS}: \{ \theta \in \Theta : T_{\alpha}(\theta) \leq c_{n, 1 - \alpha}(\theta) \}.
\]

### 2.2. Examples

In this section, we provide several examples in which the nonparametric conditional moment inequality/equality model arises. Note that Examples 2 and 6, for a conditional quantile bound and a conditional treatment effect, respectively, are used in a simulation study in Section 4.

**Example 1 (Conditional Distribution with Censoring).** The first example is a missing data example. The observations are i.i.d. Let \( Y^*_i \) be a variable that is subject to censoring: it is observed only for observations i with \( D_i = 1 \) and not for observations with \( D_i = 0 \). Let \( Z_i \) be a vector of covariates and \( X_i \) be a vector of excluded instruments that are independent of \( Y^*_i \) conditional on \( Z_i \). Then, the conditional distribution of \( Y^*_i \) given \( Z_i \), denoted \( F_{Y^*_i | Z_i} \), satisfies: for fixed \( y_0 \in \mathbb{R} \) and \( z_0 \in \text{Supp}(Z_i), \)
\[
E_{\{Y^*_i \leq y_0, D_i = 1\}}[X_i | Z_i = z_0] - E_{\{Y^*_i > y_0, D_i = 1\}}[X_i | Z_i = z_0] \geq 0. \tag{2.4}\]

This model fits into the general model (2.1) with \( \theta_0 = \theta_{Y^*_i | Z_i} \), \( m_i(D_i, \theta_0) = \int [1\{Y^*_i \leq y_i, D_i = 1\} - 1\{Y^*_i > y_i, D_i = 1\}] dF(y_i | D_i = 1) \).

A model similar to this one is used in Blundell et al. (2007) to study the distribution of female wages. In their study, \( Y^*_i \) is the potential wage of woman i, \( D_i \) is the dummy for employment status, \( Z_i \) are demographic variables, and \( X_i \) is non-wage income. Parametric and nonparametric versions of this example are discussed in CLR. Notice that the parametric version can be estimated using AS1.

**Example 2 (Conditional Quantile with Censoring).** In some cases, it is more useful to bound the conditional quantiles of \( Y^*_i \), rather than its conditional distribution. Again, suppose the observations are i.i.d. Let \( q_{\tau | y} \) denote the \( \tau \) quantile of \( Y^*_i \) given \( Z_i = z_0 \). Then under the conditional quantile independence assumption: \( q_{\tau | y} \) is the joint distribution function of \( (\varepsilon_1, \varepsilon_2) \), which is known up to the finite-dimensional parameter \( \alpha_0 \). Let \( F_{\varepsilon_1, \varepsilon_2} \) denote the marginal distributions of \( \varepsilon_1 \) and \( \varepsilon_2 \) respectively. Let \( D_i \) be the dummy that equals one if player j enters the market. Suppose that it is a simultaneous-move static game. Then, following Andrews et al. (2004) and Ciliberto and Tamer (2009), we can summarize the game by moment inequalities/equalities:
\[
E_{\{P_{00}(X, \theta_0) = 1 - D_i(1 - D_j)|X, Z = z_0\}} = 0,
\]
\[
E_{\{P_{11}(X, \theta_0) - 1\}} = 0,
\]
\[
E_{\{P_{10}(X, \theta_0) - 1\}} = 0,
\]
\[
E_{\{P_{01}(X, \theta_0) - 1\}} = 0.
\]

**Example 3 (Interval-Outcome Partially Linear-Regression).** This example is a partially-linear interval-outcome regression model. Let \( Y^*_i \) be a latent dependent variable and \( Y_i^* = X_i^\prime \beta_0 + \psi_0(Z_i) + \varepsilon \).
\[
E(\varepsilon | X_i, Z_i) = 0 \text{ a.s., where } (X_i, Z_i) \text{ are exogenous regressors some of which may be excluded from the regression. The latent variable } Y^*_i \text{ is known to lie in the observed interval } [Y_i', Y_i^*]. \text{ Then, the following moment inequalities hold for fixed } z_0 \in \text{Supp}(Z_i):}
\]
\[
E(Y_i^* - X_i^\prime \beta_0 - \psi_0(z_0) | X_i, Z_i = z_0) \geq 0 \quad \text{and}
\]
\[
E((X_i^\prime \beta_0 + \psi_0(z_0) - Y_i^* | X_i, Z_i = z_0) \geq 0. \tag{2.6}\]

This model fits into the general model (2.1) with \( \theta_0 = (\beta_0, \psi_0(z_0)) \), \( W_i = (Y_i', X_i, Z_i) \), \( m_i(W_i, \theta_0) = Y_i^* - X_i^\prime \beta_0 - \psi_0(z_0) \), and \( m_2(W_i, \theta_0) = X_i^\prime \beta_0 + \psi_0(z_0) - Y_i^* \).

**Example 4 (Semiparametric Discrete Choice Model with Multiple Equilibria).** Consider an entry game with two potential entrants, \( j = 1, 2 \), and possible multiple equilibria. For notational simplicity, we suppress the observation index for \( i = 1, \ldots, n \). The payoff from not entering the market is normalized to zero for both players. The payoff from entering is assumed to be \( \tau_j = \beta_0 + \psi(Z) - \delta_0 D_j - \delta_j \), where \( D_j \) is a dummy that equals one if the other player enters the market, \( \delta_0 > 0 \) is the competition effect, \( \delta_j \) is the rate of the payoff that is observable to both players but unobservable to the econometrician, and \( (X, Z) \) is a vector of firm or market characteristics. Let \( F(\varepsilon_1, \varepsilon_2; \theta_0) \) be the joint distribution function of \( (\varepsilon_1, \varepsilon_2) \).

**Example 5 (Revealed Preference Model).** Consider a multiple-agent discrete choice model with \( J \) players, where each player \( j \) has a choice set \( A_j \). Again, for notational simplicity, we suppress the \( i \) subscript. Let \( \pi(a_i, a_{-i}, W) \) be the payoff of agent \( j \) that depends on his own action \( a_i \), his opponents action \( a_{-i} \), and his own and opponents’ characteristics \( W \). Let \( I_j \) be the information set of player \( j \) at the time of his decision. Rationality of the agents implies the following basic rule of action:
\[
\text{sup}_{a_i \in A_i} E(\pi(a_i, a_{-i}, W) | I_j) \leq E(\pi(a_i^*, a_{-j}, W) | I_j)
\]
where $y$ depends on observed or unobserved characteristics.

The model falls into our framework if we parametrize $r$ as follows:

$$r(a^\ast, a, \cdot, W) = r(a, a, \cdot, W)$$

where $a^\ast$ is unobservable to both the agents and the econometrician, while $r(a, a, \cdot, W)$ is observable to the agents but not to the econometrician. 

Pakes (2010) proposes several assumptions on $v_1$ and $v_2$ that guarantee that (2.9) implies a moment inequality model of the following form:

$$E(r(a^\ast, a, \cdot, W) - r(a, a, \cdot, W)|W) \geq 0 \quad \forall a_j \in A_j.$$

The null hypothesis that $\theta$ is the true value for different $\theta \in \Theta$ can be used directly to evaluate the validity of certain conditional moment inequalities/equalities as described in Example 6, which follows.

**Example 6 (Functional Inequalities).** Tests constructed in this paper are suitable for testing functional inequalities of the form:

$$H_0 : u_j(x, z_0) \geq 0 \quad \text{for } z_0 \in Z \text{ and all } (x, j) \in X \times \{1, \ldots, p\},$$

$$u_j(x, z) = E(m_j(W_i)|x_1 = x, Z_i = z)$$

and the observations $W_i = (Y_i, X_i, Z_i) : i \leq n$ are from a stationary process. When the $Z_i$ variable is not present, the model reduces to that considered in Lee et al. (2013). The current model allows one to specify the inequality hypotheses for a subpopulation with characteristic $Z_i = z_0$. Each of Lee, Song, and Whang’s (2013) examples extend straightforwardly to our framework. An illustration of the extension is now given for the conditional treatment effect example.

Consider a controlled experiment, where treatment is randomly assigned to a group of subjects. Each subject is assigned the treatment with known probability $\pi(X_i, Z_i)$, where $(X_i, Z_i)$ are the observed characteristics of the subject. The researcher observes the treatment status $D_i \in \{0, 1\}$ and the outcomes $Y_i(1)$ if treated and $Y_i(0)$ if not treated. That is, the researcher observes $D_i$ and $Y_i = Y_i(1)\cdot D_i + Y_i(0)\cdot (1-D_i)$. The treatment effect for the $i$th individual is the difference between $Y_i(1)$ and $Y_i(0)$. The researcher is interested in testing if the average treatment effect given $X_i = x$ is positive for all $x \in X$ for the subpopulation with characteristic $Z_i = z_0$. Then, our test for the hypotheses in (2.13) can be applied with $p = 1$ and

$$m(W_i) = \frac{D_i Y_i}{\pi(X_i, Z_i)} - \frac{(1-D_i) Y_i}{1 - \pi(X_i, Z_i)}.$$

where $W_i = (Y_i, D_i, X_i, Z_i)$ and no parameter $\theta$ appears in the problem.

### 2.3. Parameter space

Let $(\theta, F)$ denote generic values of the parameter and distribution. Let $\mathcal{F}$ denote the parameter space for $(\theta_0, F_0)$. To specify $\mathcal{F}$ we need to introduce some notation.

Let $F_{Y_i|X_i}$ denote the conditional distribution of $Y_i$ given $X_i = x$ and $Z_i = z$ under $(\theta, F)$. Let $F_{X_i|Y_i}$ denote the conditional distribution of $X_i$ given $Z_i = z$ under $(\theta, F)$. Let $F_{Y_i}$ and $F_{X_i}$ denote the marginal distributions of $Z_i$ and $X_i$, respectively, under $(\theta, F)$.

Define

$$m_1(\theta, x, z) = F_1(x, z) = F_1(\theta, x, z) = F_{Y_i|X_i}(x|z, \theta)$$

and $\sigma_2(\theta, z)$ is defined to be the collection of $(\theta, F)$ that satisfy the following parameter space (PS) assumptions, which are necessary and moment conditions. The smoothness conditions are on $F$ and $\sigma_2(\theta, z)$ is the conditional density with respect to Lebesgue measure of $Z_i$ given $X_i = x$ and $Z_i = z$ with density $F_{Y_i}(F_{X_i}(x))$.

**Assumption PS1.** (a) $\theta \in \Theta$,

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

(c) $E(F(m_i(W_i, \theta)|X_i = x, Z_i = z) df_Z(x), \theta) = E_{X_i|Y_i}(x)\mu_{X_i}(F_{X_i}(x))$,

(d) $F_{Y_i}$ is absolutely continuous with density $f(x)$ on $X_i \subset X$, and

(e) $F_{Y_i}$ is absolutely continuous with density $f(x)$ on $X_i \subset X$, and

$$E_{X_i|Y_i}(x)\mu_{X_i}(F_{X_i}(x)).$$

Let $\{c_{ij} : l \leq 4\}$ be some finite constants and $[\delta_i : j \leq k]$ be some positive constants that do not depend on $(\theta, F)$.

**Assumption PS2.** (a) $F_2$ restricted to $z \in Z_0$ is absolutely continuous wrt $\mu_{X_0}$ with density $f(z) \in Z_0$,

(b) $F_2$ is absolutely continuous wrt $\mu_X$ with density $f(x) \in X$,

(c) $F_{Y_i|X_0}$ is absolutely continuous wrt $\mu_Y$ with density $f(y|x) \in y \times X \times Z_0$, and

$$\sigma_2(\theta, z) \in X \times \{1, \ldots, p\}$$

and $\mu_Y$ is the marginal distribution of $Y_i$ on $X_i \subset X$, and

$$E_{X_i|Y_i}(x)\mu_{X_i}(F_{X_i}(x))$$

and $\mu_{Y_0}$ is the marginal distribution of $Y_i$ on $X_i \subset X$, and

$$E_{X_i|Y_i}(x)\mu_{X_i}(F_{X_i}(x)).$$

2. Note that the model is also covered by AS1 when $Z_i$ is not present.

3. The function $p(x, z)$ can be a constant. In this case, the assignment does not depend on observed or unobserved characteristics.
expectations, not on the underlying functions themselves, which makes them relatively weak.

Let \( f(y, x, z) = f(x|y) f(z|y) f(z) \) and \( f(x, z) = f(x|z) f(z) \).

The \( k \)-vector of moment functions is denoted
\[
m(W_i, \theta) = (m_1(W_i, \theta), \ldots, m_k(W_i, \theta))'.
\] (2.16)

3. Tests and confidence sets

3.1. Test statistics

Here we define the test statistic \( T_n(\theta) \) that is used to construct a CS. We consider the conditional moment inequalities/equalities given \( X_i \) and \( Z_i = z_0 \) into equivalent conditional moment inequalities/equalities given only \( Z_i = z_0 \) by choosing appropriate weighting functions of \( X_i \), i.e., \( X_i \)-instruments. Then, we construct a test statistic based on kernel averages of the instrumented moment conditions over \( Z_i \) values that lie in a neighborhood of \( z_0 \).

The instrumented conditional moment conditions given \( Z_i = z_0 \) are of the form:
\[
E_{f_0} \left[ m_i(W_i, \theta_0, g_j(X_i)) | Z_i = z_0 \right] \geq 0 \quad \text{for} \quad j = 1, \ldots, p \quad \text{and} \quad (3.1)
\]
\[
E_{f_0} \left[ m_i(W_i, \theta_0, g_j(X_i)) | Z_i = z_0 \right] = 0 \quad \text{for} \quad j = p + 1, \ldots, k, \quad \text{for} \quad g = (g_1, \ldots, g_k) \in \mathcal{g}_c\text{-cube},
\]
where \( g = (g_1, \ldots, g_k) \) are instruments that depend on the conditioning variables \( X_i \) and \( \mathcal{g}_c\text{-cube} \) is a collection of instruments defined in (3.6) below. The collection \( \mathcal{g}_c\text{-cube} \) is chosen so that there is no loss in information.

We construct test statistics based on (3.1). The sample moment functions are
\[
m_n(\theta, g) = n^{-1} \sum_{i=1}^n m(W_i, \theta, g, b) \quad \text{for} \quad g \in \mathcal{g}_c\text{-cube},
\]
where
\[
m(W_i, \theta, g, b) = b^{-d/2} K_b(Z_i) m(W_i, \theta, g),
\]
\[
K_b(Z_i) = \max \{ 1 - |(Z_i - z_0)/b|^2, 0 \},
\]
\[
m(W_i, \theta, g) = \left( \begin{array}{c} m_1(W_i, \theta) g_1(X_i) \\ m_2(W_i, \theta) g_2(X_i) \\ \vdots \\ m_k(W_i, \theta) g_k(X_i) \end{array} \right) \quad \text{for} \quad g \in \mathcal{g}_c\text{-cube},
\] (3.2)
and \( b > 0 \) is a scalar bandwidth parameter for which \( b = b_n = o(n^{-1/4 + d/4}) \) and \( nb^{d/2} \to \infty \) as \( n \to \infty \). \( ^5 \) In the scalar \( Z_i \) case, we take \( b = b_n^{1+d/2} \) where \( b_n = 4.6865 \) and \( \hat{\sigma}_c \) is the estimated standard deviation of \( Z_i. \) \( ^6 \) The kernel employed in (3.2) is the Epanechnikov kernel. For notational simplicity, we omit the dependence of \( m_n(\theta, g) \) (and various other quantities below) on \( b. \)

Note that the normalization \( b^{-d/2} \) that appears in \( m(W_i, \theta, g, b) \) yields \( m(W_i, \theta, g, b) \) to have a variance matrix that is \( O(1) \), but not \( o(1). \) In fact, under the conditions given below, \( \text{Var}_f(m(W_i, \theta, g, b)) \to Var_f(m(W_i, \theta, g)|Z_i = z_0) f(z_0) \) as \( n \to \infty \) under \( (\theta, F) \in \mathcal{F}. \)

If the sample average \( \bar{m}_n(\theta, g) \) is divided by the scalar \( n^{-1} \sum_{i=1}^n b^{-d/2} K_b(Z_i) \), it becomes the Nadaraya–Watson nonparametric kernel estimator of \( E(m(W_i, \theta, g)|Z_i = z_0). \) We omit this divisor because doing so simplifies the statistic and has no effect on the test defined below. \( ^7 \)

The sample variance–covariance matrix of \( n^{1/2} \bar{m}_n(\theta, g) \) is
\[
\hat{\Sigma}_{n}(\theta, g) = n^{-1} \sum_{i=1}^n (m(W_i, \theta, g, b) - \bar{m}_n(\theta, g))' \times (m(W_i, \theta, g, b) - \bar{m}_n(\theta, g))' \quad \text{for} \quad g \in \mathcal{g}_c\text{-cube} \quad \text{for} \quad \epsilon = 5/100.
\] (3.3)

By design, \( \hat{\Sigma}_{n}(\theta, g) \) is a linear combination of two scale equivariant functions and hence is scale equivariant. \( ^8 \) This yields a test statistic that is invariant to rescaling of the moment functions \( m(W_i, \theta) \), which is an important property.

The quantity \( \epsilon \) in (3.4) is a tuning parameter that prevents the variance estimator from being too close to singularity. For the asymptotics considered here and in Andrews and Shi (2013b,c), \( \epsilon \) is taken to be fixed as \( n \to \infty. \) Armstrong (2011b) provides asymptotics when \( \epsilon \) goes to zero as \( n \to \infty \) (for the model with no conditioning on \( Z_i = z_0 \)). Asymptotics with \( \epsilon \) fixed are analogous to the "fixed \( b \) asymptotics" in Kiefer and Vogelsang (2002, 2005) for test statistics based on heteroskedasticity and autocorrelation consistent variance estimators. They also are analogous to the fixed bandwidth asymptotics employed in Cattaneo et al. (2010, forthcoming).

In each of these two cases, the fixed bandwidth asymptotics are found to provide better approximations to the finite-sample behavior of the statistics being considered because the asymptotic distribution depends on the tuning parameter, whereas it does not under asymptotics in which the tuning parameter converges to zero as \( n \to \infty. \)

The functions \( g \) that we consider are hypercubes on \([0, 1]^d}. \) 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 \( X_i \in \mathbb{R}^d \) denotes the untransformed IV vector and we let \( X_i \) denote the transformed IV vector. We transform \( X_i \) via a shift and rotation and then an application of the standard normal distribution function \( \Phi(x) \). Specifically, let
\[
X_i = \Phi \left( \frac{\hat{\Sigma}_{X,n}^{-1/2}(X_i - \bar{X}_n)'}{\epsilon} \right),
\]
where \( \Phi(x) \) is \( (\Phi(x_1), \ldots, \Phi(x_d))' \)
\[
\text{for} \quad x = \left( x_1, \ldots, x_d \right)' \in \mathbb{R}^d.
\]
\[
\hat{\Sigma}_{X,n} = n^{-1} \Sigma_{n,n}'(X_i - \bar{X}_n)(X_i - \bar{X}_n)' \quad \text{and} \quad \bar{X}_n = n^{-1} \sum_{i=1}^n X_i
\] (3.5)

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

---

\(^4\) The conditions on \( b \) are standard assumptions in the nonparametric density and regression literature. When these conditions are applied to a nonparametric regression or density estimator, the first condition implies that the bias of the estimator goes to zero faster than the variance (and is the weakest condition for which this holds). The second condition implies that the estimator is asymptotically normal (because it implies that it goes to zero sufficiently slowly that a Lindeberg condition holds). \(^5\) The bandwidth \( b \) is under-smoothed due to the factor \( n^{-2/3} \), which is the same as in Chernozhukov et al. (2013), rather than \( n^{-1/2} \). It is somewhat arbitrary, but seems to work well in practice. \(^6\) The definition of \( m_n(\theta, g) \) in (3.2) is the same as the definition of \( m_n(\theta, g) \) in AS1 except for the multiplicative \( b^{-d/2}K_b(Z_i) \) in \( m(W_i, \theta, g, b) \). \(^7\) This holds because division by \( n^{-1} \sum_{i=1}^n b^{-d/2}K_b(Z_i) \), rescales the test statistic and critical value identically and in consequence the rescaling cancels out. \(^8\) That is, multiplying the moment functions \( m(W_i, \theta) \) by a diagonal matrix, \( D \), changes \( \hat{\Sigma}_{n}(\theta, g) \) into \( D \hat{\Sigma}_{n}(\theta, g)D \).
for some positive integer \(r_0\). The terminology "c-cube" abbreviates countable cubes. Note that \(C_{c,r}\) is a hypercube in \([0, 1]^d\) with smallest vertex indexed by \(a\) and side lengths equal to \((2r)^{-1}\).

Note that the proofs of the asymptotic results in this paper given in the Appendix, see Andrews and Shi (2013a), cover a wide variety of classes of functions \(g\), not just the class \(g_{c,cube}\).

The test statistic \(T_{n,r_1,n}(\theta)\) is either a Cramér–von-Mises-type (CvM) or Kolmogorov–Smirnov-type (KS) statistic. The CvM statistic is

\[
T_{n,r_1,n}(\theta) = \sum_{r=1}^{r_1,n} (r^2 + 100)^{-1} \sum_{a \in \{1, \ldots, 2r\}^d} (2r)^{-d} \times S(n^{1/2} \overline{m}_n(\theta, g_{a,r}), \overline{S}_n(\theta, g_{a,r})),
\]

where \(S = S_1, S_2, S_3\), 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 \(T_{n,r_1,n}(\theta)\) 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 \(r_1,n \to \infty\).

The Kolmogorov–Smirnov-type (KS) statistic is

\[
T_{n,r_1,n}(\theta) = \sup_{g_{a,r} \in \mathcal{G}_{c,cube}: r \leq r_1,n} \times S(n^{1/2} \overline{m}_n(\theta, g_{a,r}), \overline{S}_n(\theta, g_{a,r})),
\]

where \(g_{c,cube}: r \leq r_1,n\). For brevity, the discussion in this paper focuses on CvM statistics and all results stated concern CvM statistics. Similar results hold for KS statistics.

The functions \(S_1, S_2, S_3\) are defined by

\[
S_1(m, \Sigma) = \sum_{j=1}^{p} \left[ m_j / \sigma_j \right]^2 + \sum_{j=p+1}^{p+n} \left[ m_j / \sigma_j \right]^2,
\]

\[
S_2(m, \Sigma) = \inf_{t = (t_1, \ldots, t_{d}) \in R_{d_{\infty}}} \Sigma^{-1}(m - t), \quad \text{and} \quad (3.9)
\]

\[
S_3(m, \Sigma) = \max \{|m_j/\sigma_j|^2, \ldots, |m_j/\sigma_j|^2, (m_{p+1}/\sigma_{p+1})^2, \ldots, (m_{p+n}/\sigma_{p+n})^2\},
\]

where \(m_j\) is the \(j\)th element of the vector \(m, \sigma_j^2\) is the \(j\)th diagonal element of the matrix \(\Sigma\), and \(x_j = -x\) if \(x < 0\) and \(|x_j| = 0\) if \(x \geq 0\), \(R_{\infty} = \{x \in R : x \geq 0\} \cup \{+\infty\}\), and \(R_{d_{\infty}} = R_{\infty} \times \cdots \times R_{\infty}\) with \(p\) copies. The functions \(S_1, S_2, S_3\) are referred to as the modified method of moments (M3M) or Sum function, the quasi-likelihood ratio (QLR) function, and the Max function, respectively.

### 3.2. Critical values

#### 3.2.1. GMS critical values

In this section we define two GMS critical values. The first is based on the asymptotic distribution. The second is a bootstrap version of the first. Both require simulation.

We first describe how to compute the GMS critical value that is based on the asymptotic null distribution of the test statistic.

Step 1. Compute \(\psi_n(\theta, g_{a,r})\) for \(g_{a,r} \in \mathcal{G}_{c,cube,r_1,n}\), where \(\psi_n(\theta, g_{a,r})\) is defined as follows. For \(g = g_{a,r}\), let \(\psi_{n}(\theta, g) = k_{n}^{-1}n^{1/2}\overline{D}_n(\theta, g)\overline{m}_n(\theta, g)\), where

\[
\overline{D}_n(\theta, g) = \text{Diag}(\overline{S}_n(\theta, g)), \quad k_{n} = (0.3 \ln(n))^{1/2}, \quad (3.10)
\]

and \(\overline{S}_n(\theta, g)\) is defined in (3.4). The \(j\)th element of \(\psi_{n}(\theta, g)\), denoted \(\psi_{n,j}(\theta, g)\), measures the slackness of the moment inequality \(E_p m_r(W_i, \theta, g) \geq 0\) for \(j = 1, \ldots, p\). It is shrunk towards zero via \(k_{n}^{-1}\) to ensure that one does not over-estimate the slackness.

Define \(\psi_{n,j}(\theta, g) = (\psi_{n,j}(\theta, g), \ldots, \psi_{n,j}(\theta, g), 0, \ldots, 0)' \in \mathbb{R}^p\) via, for \(j \leq p\),

\[
\psi_{n,j}(\theta, g) = \psi_{n,j}(\theta, g)B_1(1, \psi_{n,j}(\theta, g), 0, \ldots, 0)' \in \mathbb{R}^p
\]

and \(B_n = (0.4 \ln(n) / \ln(n))^{1/2}\),

\[
B_n = \sup_{\psi_{n,j}(\theta, g)} \epsilon(\psi_{n,j}(\theta, g), \psi_{n,j}(\theta, g)), (3.11)
\]

where \(\Sigma_{n,j}(\theta, g)\) and \(\Sigma_{n,j}(\theta, g)\) denote the \((j, j)\) elements of \(\overline{S}_n(\theta, g)\) and \(\overline{S}_n(\theta, g)\), respectively.

Step 2. Simulate a \((kN_2) \times \tau_{\text{reps}}\) matrix \(Z\) of standard normal random variables, where \(k\) is the dimension of \(m_r(W_i, \theta, g)\), \(N_k = \sum_{j=1}^{k} (2r)^d\) is the number of \(g\) functions employed in the test statistic, and \(\tau_{\text{reps}}\) is the number of simulation repetitions used to simulate the asymptotic distribution.

Step 3. Compute the \((kN_2) \times (kN_2)\) variance matrix \(\overline{\Sigma}_{n,m}(\theta, g)\) for \(a \in \{1, \ldots, 2r\}^d\) and \(r = 1, \ldots, r_1,n\), which are defined as follows. For \(g = g_{a,r}\) and \(g^* = g_{a,r}\), let

\[
\overline{\Sigma}_{n,j}(\theta, g, g^*) = n^{-1} \sum_{i=1}^{n} \left( m_r(W_i, \theta, g, b) - \overline{m}_n(\theta, g) \right) \times \left( m_r(W_i, \theta, g^*, b) - \overline{m}_n(\theta, g^*) \right)' .
\]

(3.12)

Note that \(\overline{\Sigma}_{n,j}(\theta, g)\), defined in (3.3), equals \(\overline{\Sigma}_{n,j}(\theta, g)\).

Step 4. Compute the \((kN_2) \times \tau_{\text{reps}}\) matrix \(\psi_n(\theta, g_{a,r})\). Let \(\psi_n(\theta, g_{a,r})\) denote the \(k\) dimensional sub-vector of \(\psi_n(\theta, g_{a,r})\) that corresponds to the \(k\) rows indexed by \(g_{a,r}\) and column \(r\) for \(r = 1, \ldots, \tau_{\text{reps}}\).

Step 5. For \(r = 1, \ldots, \tau_{\text{reps}}\), compute the simulated test statistic \(T_{n,r_1,n}(\tau)\) just as \(T_{n,cube}(\theta)\) or \(T_{n,r_1,n}(\theta)\) is computed in (3.7) or (3.8) but with \(n^{1/2}\overline{m}_n(\theta, g_{a,r})\) replaced by \(\psi_n(\theta, g_{a,r}) + \psi_n(\theta, g_{a,r})\).

Step 6. Take the critical value \(c_{\alpha,1-\alpha}(\theta)\) to be the \(1 - \alpha + n\) sample quantile of the simulated test statistics \(\{T_{n,r_1,n}(\tau) : \tau = 1, \ldots, \tau_{\text{reps}}\}\) plus \(n\), where \(n = 10^{-6}\).

---

9 When \(a = 1\), the left endpoint of the interval \((0, 1/(2r))\) is included in the interval.

10 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.
For the bootstrap version of the GMS critical value, Steps 2 and 4–6 are replaced by the following steps:

Step 2bootstrap: Generate B bootstrap samples \([W^*_\ell : \ell = 1, \ldots, n]\) for \(\tau = 1, \ldots, B\) using the standard nonparametric i.i.d. bootstrap. That is, draw \(W^*_\ell\) from the empirical distribution of \([W_\ell : \ell = 1, \ldots, n]\) independently across \(i\) and \(\tau\).

Step 4bootstrap: For each bootstrap sample, transform the regressors as in (3.5) (using the bootstrap sample in place of the original sample) and compute \(\overline{m}_{n,\tau}(\theta, g, \alpha)\) and \(\overline{\Sigma}_{n,\tau}(\theta, g, \alpha)\) just as \(\overline{m}_n(\theta, g, \alpha)\) and \(\overline{\Sigma}_n(\theta, g, \alpha)\) are computed, but with the bootstrap sample in place of the original sample.

Step 5bootstrap: For each bootstrap sample, compute the bootstrap test statistic \(T_{n,\tau,i}^{BOOT}(\theta) = T_{n,\tau,i}^{CM}(\theta)\) (or \(T_{n,\tau,i}^{KS}(\theta)\)) is computed in (3.7) (or (3.8)) but with \(n/2\) replaced by \(n/2(\overline{m}_n(\theta, g, \alpha) - \overline{m}_n(\theta, g, \alpha)) + \overline{\Sigma}_n(\theta, g, \alpha)\) and with \(\overline{\Sigma}_n(\theta, g, \alpha)\) replaced by \(\overline{\Sigma}_{n,\tau}(\theta, g, \alpha)\).

Step 6bootstrap: Take the bootstrap GMS critical value \(c_{n,1-\alpha}(\theta)\) to be the \(1-\alpha + \eta\) sample quantile of the bootstrap test statistics \((T_{n,\tau,i}^{CM}(\theta) : \tau = 1, \ldots, B)\) plus \(\eta\), where \(\eta = 10^{-6}\).

The CvM (or KS) GMS CS is defined in (2.3) with \(T_{n,\tau,i}(\theta) = T_{n,\tau,i}^{CM}(\theta)\) (or \(T_{n,\tau,i}^{KS}(\theta)\)) and \(c_{n,1-\alpha}(\theta) = c_{n,1-\alpha}(\theta)\) (or \(c_{n,1-\alpha}(\theta)\)).

The CvM GMS test of \(H_0 : \theta = \theta_0\) rejects \(H_0\) if \(T_{n,\tau,i}(\theta_0) > c_{n,1-\alpha}(\theta_0)\). The KS GMS test is defined likewise using \(T_{n,\tau,i}(\theta, g)\) and the KS critical value \(c_{n,1-\alpha}(\theta, g)\).

The choices of \(\varepsilon, \kappa, B_0\), and \(\eta\) above are based on some experimentation (in the simulation results reported AS1 and AS2).

The asymptotic results reported in the Appendix allow for other choices.

The number of cubes with side-edge length indexed by \(r\) is \((2^r)^k\), where \(d_k\) denotes the dimension of the covariance \(X_i\). The computation time is approximately in the number of cubes. Hence, it is linear in \(N_k = \sum_{r=1}^{K}(2^r)^k\). The dimension of \(Z_i\) does not affect the computation time.

In terms of computation time, the tests in this paper are not much different from those in Andrews and Shi (2013b). For details on the computation times, see Section 10.2.4 of Andrews and Shi (2013b) and Section 17.3 of Andrews and Shi (2013c). To give a general idea, to implement the test or CI for one \(\tau\) value, it takes less than one second for each of the procedures (including the CLR procedures) that we implement in the examples considered in Section 4 below.

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

3.2.2. Plug-in asymptotic critical values

Next, for comparative purposes, we define plug-in asymptotic (PA) critical values. Subsampling critical values also can be considered, see Appendix B of AS2 for details. We strongly recommend GMS critical values over PA and subsampling critical values for the same reasons as given in AS1 plus the fact that the finite-sample simulations in Section 4 show better performance by GMS critical values than PA and subsampling critical values.

PA critical values are based on the least-favorable asymptotic null distribution with an estimator of its unknown covariance kernel plugged-in. They are computed just as the GMS critical values are computed but with \(\overline{m}_n(\theta, g, \alpha) = 0_k\) (\(\in R^k\)).

The nominal \(1-\alpha\) PA CS is given by (2.3) with \(T_{n,\tau,i}(\theta) = T_{n,\tau,i}^{CM}(\theta)\) (or \(T_{n,\tau,i}^{KS}(\theta)\)) and the critical value \(c_{n,1-\alpha}(\theta)\) equal to the PA critical value. The CvM (or KS) PA test of \(H_0 : \theta = \theta_0\) rejects \(H_0\) if \(T_{n,\tau,i}^{CM}(\theta_0)\) (or \(T_{n,\tau,i}^{KS}(\theta_0)\)) exceeds the CvM (or KS) PA critical value evaluated at \(\theta = \theta_0\).

PA critical values are greater than or equal to GMS critical values for all \(n\) (because \(\overline{m}_n(\theta, g) \geq 0\) for all \(g \in \mathcal{G}_{c,-cube}\) for \(j \leq p\) and \(S_{\{m, \Sigma\}}\) is non-increasing in \(m \in R^p\), where \(m = (m_1, m_2, \ldots, m_p)\), for \(\ell = 1; 2; 3\)). Hence, the asymptotic local power of a GMS test is greater than or equal to that of a PA test for all local alternatives. Strict inequality typically occurs whenever the conditional moment inequality \(E_{\theta}(m(W_i, \theta, \alpha))\) (or \(E_{\theta}(m(W_i, \theta, \alpha))\)) is computed in (3.17). Also note that the means, \(E_{\theta}(m(W_i, \theta, g), E_{\theta}(m(W_i, \theta, g))\), and \(E_{\theta}(m(W_i, \theta))\), are not subtracted off in the definitions of \(S_{\{m, \Sigma\}}(\theta, g, \alpha)\) and \(D_{\{m, \Sigma\}}(\theta, g, \alpha)\). The reason is that the population means of the sample-size \(n\) quantities based on \(m(W_i, \theta, g, \alpha)\) are smaller than the second moments by an order of magnitude and, hence, are asymptotically negligible. See Lemmas A6 and A7 in the Appendix.

12 Note that \(D_{\{m, \Sigma\}}(\theta, z) = \text{Diag}(\sigma_{\{m, \Sigma\}}(\theta, z))\) where \(\sigma_{\{m, \Sigma\}}(\theta, z)\) is defined in (2.15). Also note that the means, \(E_{\theta}(m(W_i, \theta, g), E_{\theta}(m(W_i, \theta, g))\), and \(E_{\theta}(m(W_i, \theta))\), are not subtracted off in the definitions of \(S_{\{m, \Sigma\}}(\theta, g, \alpha)\) and \(D_{\{m, \Sigma\}}(\theta, g, \alpha)\).
"weak" or "strong" instrumental variables \( X_i \). That is, weak identification of the parameter \( \theta \) due to a low correlation between \( X_i \) and the functions \( m_j(W, \theta) \) does not affect the uniformity results.

2. **Theorem N1(b)** also holds for GMS CS's based on the QLR function \( \mathcal{S}_2 \) provided the asymptotic distribution function of the test statistic under some fixed \((\theta_i, F_i) \in \mathcal{F}\) with \( h_{2,cpt}(\theta_i) \) is continuous and strictly increasing at its 1st function \( \left( \theta_i, F_i \right) \in \mathcal{F}_{2,cpt} \) for all \( \alpha \) sufficiently small and \( \delta = 0 \). This condition likely holds in most models, but it is hard to give primitive conditions under which it holds.

3. As in AS1, an analogue of **Theorem N1(b)** holds for PA CS's if \( E_{F_{i}}(m_j(W_i, \theta_i)|X_i, Z_i = z_0) = 0 \) a.s. for \( j \leq p \) (i.e., if the conditional moment inequalities hold as equalities a.s.) under some \((\theta_i, F_i) \in \mathcal{F}\). However, the latter condition is restrictive—it fails in many applications.

4. The proofs in the Appendix cover asymptotic critical values, but not bootstrap critical values. Extending the results to cover bootstrap critical values just requires a suitable bootstrap empirical process result. For brevity, we do not give such a result. 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 non-random hypercubes that converge in probability to random hypercubes (in an \( L^2 \) sense). Again, for brevity, we do not do so. Finally, the asymptotic results cover non-data dependent bandwidths, as is typical in the nonparametric and semiparametric literature.

### 3.4. Power against fixed alternatives

We now show that the power of GMS and PA tests converges to one as \( n \to \infty \) for all fixed alternatives (for which the moment functions have 4 + \( \delta \) moments finite). Thus, both tests are consistent tests. This implies that for any fixed distribution \( F_0 \) and any parameter value \( \theta_0 \) not in the identified set \( \Theta_{\theta_0} \), the GMS and PA CS's do not include \( \Theta_{\theta_0} \) with probability approaching one. In this sense, GMS and PA CS's based on \( T_n(\theta) \) fully exploit the conditional moment inequalities and equalities. CS's based on a finite number of unconditional moment inequalities and equalities do not have this property.

The null hypothesis is

\[
H_0 : E_{F_0}(m_j(W_i, \theta_i)|X_i, Z_i = z_0) \geq 0
\]

a.s. for \( j = 1, \ldots, p \) and

\[
E_{F_0}(m_j(W_i, \theta_i)|X_i, Z_i = z_0) = 0
\]

a.s. for \( j = p + 1, \ldots, k \),

(3.16)

where \( \Theta_{\theta_0} \) denotes the null parameter value and \( F_0 \) denotes the fixed 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 \( \mathcal{F}_{\gamma} \) denote all \((\theta, F)\) that satisfy Assumptions PS1–PS3 that define \( \mathcal{F} \) except Assumption PS1(c) and (d) which impose the conditional moment inequalities and equalities. As defined, \( \mathcal{F} \subset \mathcal{F}_{\gamma} \). Note that \( \mathcal{F}_{\gamma} \) includes \((\theta, F)\) pairs for which \( \theta \) lies outside of the identified set \( \Theta_{\theta_0} \) as well as all values in the identified set.

13 This condition is Assumption GMS2(a) in Section 7.4 of the Appendix.
14 This holds because the identified set based on a finite number of moment inequalities typically is larger than the identified set based on 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 set, \( \mathcal{X}_F(\theta) \), of values \( x \) for which the moment inequalities or equalities evaluated at \( \theta \) are violated under \( F \) is defined as follows. For any \( \theta \in \Theta \) and any distribution \( F \) with \( E_F(||m(W_i, \theta)|| | Z_i = z_0) < \infty \), let

\[
\mathcal{X}_F(\theta) = \{x \in \mathbb{R}^k : E_F(m_j(W_i, \theta)|X_i = x, Z_i = z_0) < 0 \}
\]

for some \( j \leq p \) or \( E_F(m_j(W_i, \theta)|X_i = x, Z_i = z_0) \neq 0 \) for some \( j = p + 1, \ldots, k \).

(3.17)

The next assumption, Assumption NFA, states that violations of the conditional moment inequalities or equalities occur for the null parameter \( \theta_0 \) for \( X_i \) values in a set with positive conditional probability given \( Z_i = z_0 \) under \( F_0 \). Thus, under Assumption NFA, the moment conditions specified in (3.16) do not hold.

**Assumption NFA.** The null value \( \theta_0 \in \Theta \) and the true distribution \( F_0 \) satisfy: (a) \( P_{F_0}(X_i \in \mathcal{X}_{F_{\theta_0}}(\theta_0)|Z_i = z_0) > 0 \), where \( \mathcal{X}_{F_{\theta_0}}(\theta_0) \) is defined in (3.17), and (b) \((\theta_0, F_0) \in \mathcal{F}_{\gamma}\).

The following theorem shows that GMS and PA tests are consistent against all fixed alternatives that satisfy Assumption NFA.

**Theorem AN2.** Suppose Assumption NFA holds. Then,

(a) \( \lim_{n \to \infty} P_{F_0}(T_n(\theta_0) > c(\varphi_n(\theta_0), \hat{h}_{2,n}(\theta_0), 1 - \alpha)) = 1 \)

(b) \( \lim_{n \to \infty} P_{F_0}(T_n(\theta_0) > c(0_{k-cube}, \hat{h}_{2,n}(\theta_0), 1 - \alpha)) = 1 \), where \( 0_{k-cube} \) denotes the zero function on \( 0_{k-cube} \).

Comment 4 to **Theorem N1** applies also to **Theorem AN2.**

### 4. Monte Carlo simulations

This section provides simulation evidence concerning the finite-sample properties of the confidence intervals (CI's) and tests introduced in the paper. We consider two models: a quantile selection and a conditional treatment effect model. In the quantile selection model, we compare different versions of the CI's introduced in the paper. In the conditional treatment effect model, the tests are used directly (rather than to construct CI's), and we compare different versions of the tests. In both models, we provide comparisons of the proposed procedures with the series and local linear procedures in CLR.

#### 4.1. Confidence intervals and tests considered

To be specific, we compare different test statistics and critical values in terms of their coverage probabilities (CP's) for points in the identified set and their false coverage probabilities (FCP's) for points outside the identified set in the quantile selection model. We compare different test statistics and critical values in terms of their rejection probabilities under the null (NRP's) and under alternatives (ARP's) in the conditional treatment effect model. Obviously, one wants FCP's (ARP's) to be as small (large) as possible. FCP's are directly related to the power of the tests used to construct the CI and are related to the length of the CI, see Pratt (1961).

The following test statistics are considered: (i) CvM/Sum, (ii) CvM/QLR, (iii) CvM/Max, (iv) KS/Sum, (v) KS/QLR, and (vi) KS/Max, as defined in Section 3. In the conditional treatment effect model, different choices of the S function (Sum, QLR and Max) coincide because there is only one conditional moment inequality. We thus do not distinguish them in the results. Asymptotic normal, bootstrap, and subsampling critical values are computed. In particular, we consider PA/Asy, PA/Br, GMS/Asy, GMS/Br, and...
Sub critical values. The critical values are simulated using 5001 repetitions (for each original sample repetition). The base case values of \( \kappa_n, B_n, \) and \( \varepsilon \) for the GMS critical values are specified as follows and are used in both models: \( \kappa_n = \sqrt{0.3 \log(n)}, B_n = \sqrt{0.4 \log(n) / \log(\log(n))}, \) and \( \varepsilon = 5/100. \) Additional results are reported for variations of these values. The base case sample size is 250. Some additional results are reported for \( n = 100 \) and 500. The subsample size is 20 when the sample size is 250. Results are reported for nominal 0.95 CI’s and 0.05 tests. The number of simulation repetitions used to compute CP’s and FCP’s is 5000 for all cases. This yields a simulation standard error of 0.0031.

In the first model, the reported FCP’s are “CP-corrected” by employing a critical value that yields a CP equal to 0.95 at the closest point of the identified set if the CP at the closest point is less than 0.95. If the CP at the closest point is greater than 0.95, then no CP correction is carried out. The reason for this “asymmetric” CP correction is that CS’s may have CP’s greater than 0.95 for points in the identified set, even asymptotically, in the present context and one does not want to reward over-coverage of points in the identified set by CP correcting the critical values when making comparisons of FCP’s. In the second model, the ARP’s are “NRP-corrected” analogously.

We use the Epanechnikov kernel and the bandwidth \( b = b^0 n^{-2/7} \) described in the paragraph containing (3.2) for both simulation examples. For comparative purposes, some results are also reported for \( b = 0.5 b^0 n^{-2/7} \) and \( b = 2 b^0 n^{-2/7}. \)

We provide simulation comparisons of our CS’s and tests with those of CLR. To implement the CLR tests, we follow Example C of CLR. For the quantile selection model, for each \( \theta \), we use \( \beta(x, z, \theta) \) defined in (4.4) of CLR as the auxiliary bound function, use \( \beta(z, \theta) \) as the auxiliary parameter, and test \( H_{\beta_0} : \beta(z, \theta) \geq 0 \) against \( \beta(z, \theta) < 0. \) (The CLR’s are obtained by inverting the CLR tests.) For the treatment effect model, described below, we use \( \beta(x, z) = E[Y|D_t/p - Y(1 - D_t)/(1-p)](X_t, Z_t) \) as the auxiliary bound function, use \( \beta(z, x) = \min_{x \in X} \beta(x, z) \) as the auxiliary parameter, and test \( H_{\beta_0} : \beta(z) \geq 0 \) against \( \beta(z) < 0. \)

We implement both the series and local linear versions of CLR’s test. We use their GAUSS code and follow the implementation instructions in CLR whenever possible. The models considered here, however, are more complicated than in CLR’s examples because the nonparametric estimation of \( \beta(x, z, \theta) \) involves two regressors \( X_t \) and \( Z_t \). The latter poses new questions about the choices of knots in the series approximation and the choices of bandwidths in the local linear approximation. For the series versions, we use tensor product B-splines and allow different numbers of knots for \( X_t \) and \( Z_t \). The number of knots is the (integer part of the) number chosen by cross-validation multiplied by \( \sqrt{n^{-1/3} n^{-2/7}} \). The multiplicative factor is used to obtain undersmoothing. For the local linear version, we use the optimal bandwidth formula given for multivariate local linear regression by Yang and Tschernig (1999) (Eq. (A.1)), and use the same plug-in rule as CLR’s rule-of-thumb bandwidth to plug in the estimated quantities. The resulting plug-in bandwidth is then multiplied by \( \sqrt{n^{-1/3} n^{-2/7}} \) to obtain undersmoothing. The CLR CS’s and tests employ an estimated control set.

4.2. Nonparametric quantile selection

This model extends the quantile selection model in AS1. We are interested in the conditional quantile of a treatment response given the value of covariates \( X_i \) and \( Z_i \). The results also apply to other types of response variables with selection. As in AS1, \( X_i \) is assumed to satisfy the quantile monotone instrumental variable (QMIV) assumption. In this paper, we add an additional covariate \( Z_i \) that does not necessarily satisfy the QMIV assumption. The results of AS1 do not cover such a model.

The model setup is as follows. The observations are i.i.d. Let \( y_i(t) \in Y \) be individual \( i \)’s “constricted” response variable given treatment \( t \in T \). Let \( T_i \) be the realization of the treatment for individual \( i \). The observed outcome variable is \( Y_i = y_i(T_i) \). Let \( X_i \) be a covariate whose support contains an ordered set \( X \). Let \( Z_t \) be another covariate. We observe \( W_i = (Y_i, X_i, Z_i, T_i) \). The parameter of interest, \( \theta \), is the conditional quantile of \( y_i(t) \) given \( (X_i, Z_i) = (x, z) \) for some \( t \) in \( T \), some \( x_0 \in X \), and some \( z_0 \in Z \), which is denoted \( Q_{y_i(t)|X_i,Z_i}(t|x_0, z_0) \). We assume the conditional distribution of \( y_i(t) \) given \( (X_i, Z_i) = (x_0, z_0) \) is absolutely continuous at its quantile for all \( x \in X \). We assume that \( Z_t \) satisfies the QMIV assumption given \( Z_t = z_0 \), i.e., \( Q_{y_i(t)|X_i,Z_t}(t|x_0, z_0) \leq Q_{y_i(t)|X_i,Z_t}(t|x_0, z_0) \) for all \( x_1 \leq x_2 \). AS1 describes four empirical problems that fit in their quantile selection model. All of those problems fit in the nonparametric quantile selection model considered here if one or more of the covariates is not a QMIV.

The model setup above implies the following conditional moment inequalities:

\[ E \left( 1(X_i \leq x_0)(1(Y_i \leq \theta, T_i = t) \right) + I(1(T_i \neq t) - t \right) | (X_i, Z_i = z_0) \geq 0 \] a.s. and

\[ E \left( 1(X_i \geq x_0)(t - 1(Y_i \leq \theta, T_i = t)) | (X_i, Z_i = z_0) \geq 0 \right. \] a.s. \quad (4.1)

For the simulations, we consider the following data generating processes (DGP):

\[ y_i(1) = \mu(X_i, Z_i) + \sigma(X_i, Z_i) u_i, \]

where \( \partial \mu(x, z) / \partial x \geq 0 \) and \( \sigma(x, z) \geq 0 \),

\[ X_i, Z_i \sim \text{Unif}[0, 2], \quad (\varepsilon_i, u_i) \sim N(0, I_2), \]

\[ (X_i, Z_i) \perp (\varepsilon_i, u_i), \quad (X_i, Z_i), \quad Y_i = y_i(T_i), \quad \text{ and } \quad t = 1. \]

The variable \( y_i(0) \) is irrelevant (because \( Y_i \) enters the moment inequalities in (4.1) only through \( 1(Y_i \leq \theta, T_i = t) \) and, hence, is left undefined. With this DGP, \( X_i \) satisfies the QMIV assumption for any \( \tau \in (0, 1) \) and \( Z_i \) might not. We consider the median: \( \tau = 0.5 \). We focus on the conditional median of \( y_i(1) \) given \( (X_i, Z_i) = (1.5, 1.0) \), i.e., \( \theta = Q_{y_i(1)|X_i,Z_i}(0.5|x_0, z_0) \) with \( (x_0, z_0) = (1.5, 1.0) \).

Some algebra shows that the conditional moment inequalities in (4.1) imply:

\[ \theta \geq \hat{\theta}(x, z_0) := \mu(x, z_0) + \sigma(x, z_0) \times \Phi^{-1} \left( 1 - \left[ 2 \Phi \left( \frac{L(x, z_0) - 1}{L(x, z_0) + 1} \right) \right]^{-1} \right) \quad \text{for } x \leq 1.5 \]

\[ \hat{\theta} \geq \hat{\theta}(x, z_0) := \mu(x, z_0) + \sigma(x, z_0) \times \Phi^{-1} \left( 2 \Phi \left( \frac{L(x, z_0) - 1}{L(x, z_0) + 1} \right) \right) \quad \text{for } x \leq 1.5. \]

We call \( \theta(x, z_0) \) and \( \hat{\theta}(x, z_0) \) the lower and upper bound functions on \( \theta \), respectively. The identified set for the quantile selection model is \( \sup_{x \in X} \hat{\theta}(x, z_0) \), \( \inf_{x \in X} \hat{\theta}(x, z_0) \). The shape of the lower and upper bound functions depends on \( \mu, \sigma, \) and \( L = \) functions.

We consider three specifications, one that yields flat bound functions, another that yields kinked bound functions, and a third that yields peaked bound functions. For the flat bound DGP, \( \mu(x, z) = 2, \sigma(x, z) = 1, \) and \( L(x, z) = 1 \) for \( x, z \in [0, 2] \). In this case, \( \hat{\theta}(x, z) = 2 + \Phi^{-1} \left( 1 - \left[ 2 \Phi \left( \frac{L(x, z_0) - 1}{L(x, z_0) + 1} \right) \right]^{-1} \right) \quad \text{for } x \leq 1.5 \) and \( \hat{\theta}(x, z) = \)
\[ 2 + \Phi^{-1}\left(\left[2\Phi(1)\right]^{-1}\right) \] for \( x > 1.5 \). For the kinked bound DGP, 
\[ \mu(x, z) = (x \wedge 1) \Phi^{-1}\left(1 - \left[2\Phi(1)\right]^{-1}\right) / 2 \] for \( x \leq 1.5 \), and 
\[ \hat{\theta}(x, z) = (x \wedge 1) + (z \wedge 1) + (x + z) \cdot \Phi^{-1}\left(\left[2\Phi(1)\right]^{-1}\right) / 2 \] for \( x > 1.5 \). For the peaked bound function, 
\[ \mu(x, z) = (x \wedge 1) + (z \wedge 1) + (x^5 + z^5) \cdot \Phi^{-1}\left(1 - \left[2\Phi(1)\right]^{-1}\right) / 2 \] for \( x \leq 1.5 \), and 
\[ \hat{\theta}(x, z) = (x \wedge 1) + (z \wedge 1) + (x^5 + z^5) \cdot \Phi^{-1}\left(\left[2\Phi(1)\right]^{-1}\right) / 2 \] for \( x > 1.5 \).

The CP or FCP performance of a CI at a particular value \( \theta \) depends on the shape of the conditional moment functions, as functions of \( x \) and \( z \) and evaluated at \( \theta \). In the present model, the conditional moment functions are

\[
\beta(x, z, \theta) = \begin{cases} 
E(1(Y_i \leq \theta, T_i = 1) + 1|T_i \neq 1) 
- 0.5(X_i, Z_i = (x, z)) & \text{if } x < 1.5 \\
E(0.5 - 1|Y_i \leq \theta, T_i = 1)| 
(X_i, Z_i = (x, z)) & \text{if } x \geq 1.5.
\end{cases}
\]

The conditional moment functions as functions of \( x \) at \( z = z_0 \) are flat, kinked and peaked under the three specifications of \( \mu, \sigma \), and \( L \) functions, respectively. The functions as a function of \( z \) at each \( x \) also possess those three shapes at the point \( z = z_0 \) depending on the specification. See Fig. 1.

4.2.2. g functions

The \( g \) functions employed by the test statistics are indicator functions of hypercubes in \([0, 1]\), i.e., intervals, as in AS1. The regressor \( X_i \) is transformed via the method described in (3.5) to lie in \((0, 1)\). The hypercubes have side-edge lengths \((2r)^{-1}\) for \( r = r_0, \ldots, r_1 \), where \( r_0 = 1 \) and the base case value of \( r_1 \) is 3.\(^{16}\) The base case number of hypercubes is 12. We also report results for \( r_1 = 2, 4 \), which yield 6, and 20 hypercubes, respectively.

Note that we use a smaller value of \( r_1 \) as the base-case value in this paper than in AS1. This is because the test statistic for a non-parametric parameter of interest depends only on observations local to \( Z_i = z_0 \), which is a fraction of the full sample. For example, the Epanechnikov kernel gives positive weight only to observations within distance \( \theta \) to \( z_0 \). When \( n = 250 \) and \( Z \sim \text{Unif}[0, 2] \), observations that receive positive weight lie in an interval centered at \( z_0 \) of length about \( 2b = 9.36\sigma_2n^{-1/2} \approx 0.64 \), which is 32% of the support of \( Z \). This interval on average contains 80 effective observations when \( n = 250 \). Thus, the finest cube when \( r_1 = 3 \) contains 80/6 \( \approx 13 \) effective observations. On the other hand, the finest cube when \( r_1 = 7 \) contains only 80/14 \( \approx 5.7 \) effective observations. For this reason, a value of \( r_1 \) that is smaller than that used in AS1 leads to better CP and FCP performance of the CS’s in the nonparametric model.

4.2.2. Simulation results: confidence intervals proposed in this paper

Tables 1–3 report CP’s and CP-corrected FCP’s for a variety of test statistics and critical values proposed in this paper for a range of cases. The CP’s are for the lower endpoint of the identified interval in Tables 1–3 and for the flat, kinked, and peaked bound functions. FCP’s are for points below the lower endpoint.\(^{17}\)

Table 1 provides comparisons of different test statistics when each statistic is coupled with PA/Asy and GMS/Asy critical values. Table 2 provides comparisons of the PA/Asy, PA/Bt, GMS/Asy, GMS/Bt, and Sub critical values for the CvM/Max and KS/Max test statistics. Table 3 provides robustness results for the CvM/Max and KS/Max statistics coupled with GMS/Asy critical values. The results in Table 3 show the degree of sensitivity of the results to (i) the sample size, \( n \), (ii) the number of cubes employed, as indexed by \( r_1 \), (iii) the choice of \( (\kappa_n, B_n) \) for the GMS/Asy critical values, (iv) the value of \( \epsilon \), upon which the variance estimator \( \varphi_n(\theta, g) \) depends, and (v) the bandwidth choice. Table 3 also reports results for CI’s with nominal level 0.5, which yield asymptotically half-median unbiased estimates of the lower endpoint.

Table 1 shows that all of the CI’s have coverage probabilities greater than or equal to 0.95 for all three specifications of

\[ \text{Table 1, } n = 250, r_1 = 3, \text{ PA/Asy, GMS/Asy, CvM/Max, KS/Max.} \]

\[ \text{Table 2, } n = 250, r_1 = 3, \text{ PA/Asy, PA/Bt, GMS/Asy, GMS/Bt, Sub.} \]

\[ \text{Table 3, } n = 250, r_1 = 3, \text{ Pa/Asy, GMS/Asy, GMS/Bt, Sub.} \]

\[ \text{Table 3, } n = 250, r_1 = 3, \text{ Ch/Max, KS/Max.} \]

\[ \text{Note that the DGP is the same for FCP’s as for CP’s, just the value } \theta \text{ that is to be covered is different. For the lower endpoint of the identified set, FCP’s are computed for } \theta \text{ equal to sup}_{\theta \leq 1} \hat{\theta}(x, 1) - c \times (250/n)^{1/4}, \text{ where } c = 0.34, 0.78, \text{ and } 1.1 \text{ in the flat, kinked, and peaked bound cases, respectively. These points are chosen to yield similar values for the FCP's across the different cases considered.} \]

\(^{16}\) For simplicity, we let \( r_1 \), denote \( r_1, \text{ here and below.} \)
Table 1
Nonparametric quantile selection model: base-case test statistic comparisons.

<table>
<thead>
<tr>
<th>DGP</th>
<th>Statistic</th>
<th>Crit val</th>
<th>PA/Asy</th>
<th>PA/Bt</th>
<th>GMS/Asy</th>
<th>GMS/Bt</th>
<th>Sub</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat bound</td>
<td>PA/Asy</td>
<td>0.974 0.974</td>
<td>0.971</td>
<td>0.968</td>
<td>0.968</td>
<td>0.963</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GMS/Asy</td>
<td>0.953 0.953</td>
<td>0.951</td>
<td>0.955</td>
<td>0.955</td>
<td>0.953</td>
<td></td>
</tr>
<tr>
<td>Kinked bound</td>
<td>PA/Asy</td>
<td>0.998 0.998</td>
<td>0.997</td>
<td>0.995</td>
<td>0.995</td>
<td>0.995</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GMS/Asy</td>
<td>0.990 0.990</td>
<td>0.989</td>
<td>0.989</td>
<td>0.989</td>
<td>0.987</td>
<td></td>
</tr>
<tr>
<td>Peaked bound</td>
<td>PA/Asy</td>
<td>0.998 0.998</td>
<td>0.997</td>
<td>0.995</td>
<td>0.995</td>
<td>0.996</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GMS/Asy</td>
<td>0.992 0.992</td>
<td>0.991</td>
<td>0.991</td>
<td>0.991</td>
<td>0.991</td>
<td></td>
</tr>
</tbody>
</table>

(b) False coverage probabilities (coverage probability corrected)

<table>
<thead>
<tr>
<th>DGP</th>
<th>Statistic</th>
<th>Crit val</th>
<th>PA/Asy</th>
<th>PA/Bt</th>
<th>GMS/Asy</th>
<th>GMS/Bt</th>
<th>Sub</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat bound</td>
<td>PA/Asy</td>
<td>0.57 0.57</td>
<td>0.54</td>
<td>0.67</td>
<td>0.67</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GMS/Asy</td>
<td>0.45 0.45</td>
<td>0.45</td>
<td>0.61</td>
<td>0.61</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>Kinked bound</td>
<td>PA/Asy</td>
<td>0.67 0.67</td>
<td>0.65</td>
<td>0.67</td>
<td>0.67</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GMS/Asy</td>
<td>0.49 0.49</td>
<td>0.49</td>
<td>0.57</td>
<td>0.57</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>Peaked bound</td>
<td>PA/Asy</td>
<td>0.57 0.57</td>
<td>0.55</td>
<td>0.60</td>
<td>0.60</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GMS/Asy</td>
<td>0.50 0.50</td>
<td>0.49</td>
<td>0.55</td>
<td>0.55</td>
<td>0.53</td>
<td></td>
</tr>
</tbody>
</table>

Table 2
Nonparametric quantile selection model: base-case critical value comparisons.

<table>
<thead>
<tr>
<th>DGP</th>
<th>Statistic</th>
<th>Critical value</th>
<th>PA/Asy</th>
<th>PA/Bt</th>
<th>GMS/Asy</th>
<th>GMS/Bt</th>
<th>Sub</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat bound</td>
<td>CvM/Max</td>
<td>0.971 0.971 0.951</td>
<td>0.948</td>
<td>0.948</td>
<td>0.963</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>KS/Max</td>
<td>0.963 0.963 0.953</td>
<td>0.948</td>
<td>0.948</td>
<td>0.909</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GMS/Asy</td>
<td>0.997 0.997 0.989</td>
<td>0.988</td>
<td>0.986</td>
<td>0.999</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kinked bound</td>
<td>CvM/Max</td>
<td>0.997 0.997 0.991</td>
<td>0.990</td>
<td>0.990</td>
<td>0.999</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>KS/Max</td>
<td>0.996 0.996 0.991</td>
<td>0.990</td>
<td>0.990</td>
<td>0.968</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) False coverage probabilities (coverage probability corrected)

<table>
<thead>
<tr>
<th>DGP</th>
<th>Statistic</th>
<th>Critical value</th>
<th>PA/Asy</th>
<th>PA/Bt</th>
<th>GMS/Asy</th>
<th>GMS/Bt</th>
<th>Sub</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat bound</td>
<td>CvM/Max</td>
<td>0.54 0.55 0.45</td>
<td>0.44</td>
<td>0.44</td>
<td>0.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>KS/Max</td>
<td>0.64 0.66 0.60</td>
<td>0.57</td>
<td>0.66</td>
<td>0.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GMS/Asy</td>
<td>0.65 0.66 0.49</td>
<td>0.47</td>
<td>0.60</td>
<td>0.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kinked bound</td>
<td>CvM/Max</td>
<td>0.64 0.67 0.57</td>
<td>0.53</td>
<td>0.40</td>
<td>0.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>KS/Max</td>
<td>0.55 0.54 0.49</td>
<td>0.47</td>
<td>0.51</td>
<td>0.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GMS/Asy</td>
<td>0.56 0.55 0.53</td>
<td>0.49</td>
<td>0.39</td>
<td>0.39</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The bound functions. The PA/Asy CI’s have noticeably larger over- coverage than the GMS/Asy CI’s in all cases. The GMS/Asy CI’s have CP’s close to 0.95 with the flat bound DGP and larger than 0.95 with the other two DGP’s. The CP’s are not sensitive to the choice of the test statistics.

The FCP results in Table 1 show (i) a clear advantage of the GMS-based CI’s over the PA-based ones; (ii) a clear advantage of the CvM-based CI’s over the KS-based ones, and (iii) little difference between the test statistic specifications: Sum, QLR or Max. The comparison holds for all three types of DGP’s.

Table 2 compares the critical values PA/Asy, PA/Bt, GMS/Asy, GMS/Bt, and Sub. The results show little difference in CP’s and FCP’s between the Asy and Bt versions of the CI’s regardless of the DGP specification or the test statistic choice (CvM or KS).18

The GMS critical values noticeably outperform the PA counterparts in terms of FCP’s. The CvM/Max test statistic coupled with the GMS/Asy or GMS/Bt critical values outperforms all other combinations in terms of FCP’s in all cases. The KS/Max/Sub test undercovers noticeably in the flat bound case in Table 2, but not in the kinked and peaked bound cases. We believe this is due to two effects that cancel each other in the latter two cases, but not in the flat bound case because one of the effects is missing in the flat bound case.

The first effect is a tendency of this test to under-cover due to finite sample effects (because the subsampling error in coverage has order that depends on 1/n, not 1/n, where b ≪ n). The second effect is the over-coverage of subsampling tests asymptotically when there are slack moment inequalities, see Andrews and Guggenberger (2009). The second effect applies in the kinked and peaked bound cases, but not the flat bound case. Table 2 indicates that the CvM/Max/Sub test does not have the same tendency to under-cover due to finite-sample effects as the KS/Max/Sub test does in the quantile selection model.

Table 3 provides results for the CvM/Max and KS/Max statistics coupled with the GMS/Asy critical values for several variations of the base case. The table shows that the CI’s perform similarly at different sample sizes, with different choices of cells and with a smaller ε.19 There is some sensitivity to the magnitude of the GMS tuning parameters (κ, Bε) – doubling their values increases both the CP’s and the FCP’s, but halving their values does not decrease the CP’s much below 0.95. There is more sensitivity to the kernel bandwidth – a larger bandwidth reduces the FCP drastically while keeping the CP at around 0.95 and a smaller bandwidth does the opposite. This result is closely related to the flatness of the bound. The bound is completely flat on the support of Z. It is more efficient

---

18 Hall (1993) shows that undersmoothing or bias correction is necessary for consistency of the bootstrap. Undersmoothing is employed in this paper. Hall (1993) also shows that in the context of nonparametric curve estimation, the bootstrap has advantages over the Gaussian approximation in providing a uniform confidence band for the curve. This result does not shed light on the relative performance of Asy and Bt-based tests in this paper because (i) the test statistics are not asymptotically pivotal in the present context, whereas they are in the situation considered in Hall (1993), and (ii) we consider inference at just one point (Z = z0) of the curve.

19 The θ values at which the FCP’s are computed differs from the lower endpoint of the identified set by a distance that depends on (nb)−1/2. Table 3 suggests that the “local alternatives” that give equal FCP’s converge to the null hypothesis at a rate that is slightly faster than (nb)−1/2 for sample sizes n in the range 100–500.
to use more of the data information by using a larger bandwidth. This phenomenon does not occur with the kinked bound and the peaked bound as shown in Tables A1 and A2 in the Appendix, as Andrews and Shi (2013a).

The last two rows of Table 3 show that a CI based on $\alpha = 0.5$ provides a good choice for an estimator of the identified set. For example, the lower endpoint estimator based on the CvM/Max-GMS/Asy CS with $\alpha = 0.5$ is close to being median-unbiased. It is less than the lower bound with probability 0.525 and exceeds it with probability 0.475 when $n = 250$.

The FCP’s reported in Tables 1–3 are computed at different $\theta$ values (outside the identified set) with the three different bound functions. This is done to ensure that the FCP’s lie in a meaningful range. However, it is also of interest to consider the same $\theta$ value for all three bounds and, hence, to see how the shape of the bound function affects FCP’s. For the CvM/Max/GMS/Asy CI, the FCP’s computed for $\theta = 0.78$ are 0.02, 0.49, and 0.81 for the flat, kinked, and peaked bound functions, respectively. Thus, the FCP’s are best (lowest) for the flat bound and highest (worst) for the peaked bound function.

In summary, we find that the CI’s based on the CvM/Max statistic with the GMS/Asy critical value perform the best of those proposed in this paper in the quantile selection example considered. Equally good are the CI’s that use the Sum or QLR statistic in place of the Max statistic and the GMS/Bt critical value in place of the GMS/Asy critical value. The CP’s and FCP’s of the CvM/Max/GMS/Asy CI’s are quite good over a range of sample sizes. The findings echo those in AS1 in the parametric quantile selection example.

### 4.2.3. Simulation results: comparisons with CLR confidence intervals

Table 4 reports comparisons of CP’S and FCP’S of the CI’S proposed in this paper, denoted by AS, with the series and local linear versions of the CI’S proposed in CLR. The AS CI’S use the Max S function and GMS/Asy critical values. The CLR CI’S are described in Section 4.1. The data generating processes considered are the same as in Table 1.

Table 4 shows that the nominal 95% AS CI’s have good finite sample CP’S, being 0.951 or greater in all cases. In contrast, the series and local linear CLR CI’S under cover in the flat bound case with CP’S being 0.895 and 0.860, respectively. The FCP’S of the AS CI’S are noticeably less than those of the CLR CI’S in the flat bound case. The opposite is true in the peaked bound case. In the kinked bound case, the AS and CLR CI’S have similar FCP’S. This is consistent with the theoretical asymptotic power comparisons in Section 11 of the Appendix, see Andrews and Shi (2013a).

In sum, the CvM/Max-GMS/Asy CI has more robust null rejection probabilities than the CLR CI’S. Its FCP’S are better (i.e., lower) for the flat bound function and worse (i.e., higher) for the peaked bound function.

### 4.3. Conditional treatment effects

In this example, we illustrate how the proposed method can be used to test functional inequality hypotheses.

We are interested in the effect of a randomly assigned binary treatment ($D_i$) conditional on covariates $X_i$ and $Z_i$. The outcome variable of interest, $Y_i$, is a mixture of two potential outcomes $Y_i(1)$ and $Y_i(0)$: $Y_i = Y_i(1) + (1 - D_i)Y_i(0)$. The difference $Y_i(1) - Y_i(0)$ is the effect of treatment on individual $i$. The treatment effect for every individual cannot be identified (even partially) because $Y_i(1)$ and $Y_i(0)$ are never observed simultaneously. Thus, one often focuses on the average treatment effect of a chosen group of individuals with certain observed characteristics. The chosen group of individuals that we consider here is individuals with $Z_i = z_0 \in Z$ and $X_i \in X$, where $Z$ and $X$ are the supports of $Z_i$ and $X_i$, respectively.

We test the hypothesis:

$$E[y_i(1) - y_i(0)|(X_i, Z_i) = (x, z_0)] \geq 0 \quad \text{for all } x \in X.$$  \hspace{1cm} (4.5)

The framework can be extended to treatments with any finite number of treatment values. If the $X_i$ variable is not present, the problem is a trivial case of (2.1) where $X$ is a singleton. If the $Z_i$ variable is not present, the problem fits in the framework of AS1 and Lee (2013). The nonparametric method proposed in this paper allows us to focus on a particular value of $Z_i$.

Examples of the above hypothesis include: (i) whether a certain drug reduces blood pressure for people of all ages and genders ($X_i = (\text{age, gender})$) whose body mass index ($Z_i$) is at certain level ($z_0$); (ii) whether students of a certain IQ score ($Z_i = z_0$) do better in smaller classes than in bigger classes regardless of their parents’ income ($X_i$); and (iii) whether group liability discourages default better than individual liability in a micro-loan program for villages of all sizes ($X_i$) and certain average income level ($Z_i = z_0$).

The model setup is as follows. We assume that $D_i$ is randomly assigned and $\Pr(D_i = 1) = \pi \in (0, 1)$. Then,

$$E[y_i(1) - y_i(0)|(X_i, Z_i) = (x, z_0)]$$

$$= E \left[ \frac{Y_i D_i}{\pi} - \frac{Y_i(1 - D_i)}{1 - \pi} \bigg| (X_i, Z_i) = (x, z_0) \right].$$  \hspace{1cm} (4.6)

In consequence, the hypothesis (4.5) is equivalent to testing if $\pi = 0$ is in the identified set of the following moment inequality model:

$$E \left[ \frac{Y_i D_i}{\pi} - \frac{Y_i(1 - D_i)}{1 - \pi} - \theta \bigg| (X_i, Z_i) = (x, z_0) \right] \geq 0$$

for all $x \in X$.  \hspace{1cm} (4.7)

For the simulations, we consider the following data generating process (DGP):

$$y_i(0) = 0, \quad y_i(1) = \mu(X_i, Z_i) + u_i, \quad D_i = 1[\varepsilon_i \geq 0],$$

$$X_i \sim \text{Unif}[0, 2], \quad Z_i \sim \text{Unif}[-1, 1], \quad (\varepsilon_i, u_i) \sim N(0, 1),$$

$$\mu(x, z) = \begin{cases} \mu_1(x, z) = -a, & \mu_2(x, z) = x + |z| - a, \quad \mu_2(x, z) = \log(z + 1) - a, & a \text{ is a constant.} \end{cases}$$  \hspace{1cm} (4.8)

The function $\mu(x, z)$ is the conditional treatment effect function at $(X_i, Z_i) = (x, z)$. We focus on $z_0 = 0$.

Three different $\mu(x, z)$ functions are considered, which are flat, kinked, and tilted as a function of $z$, respectively. They are:

$$\mu_1(x, z) = -a, \quad \mu_2(x, z) = x + |z| - a, \quad \mu_3(x, z) = \log(z + 1) - a,$$

where $a$ is a constant. The hypothesis (4.5) holds if $a = 0$ and is violated if $a > 0$. The functions $\mu_1$ and $\mu_2$ do not change sign in a neighborhood around $z_0$, whereas the tilted function $\mu_3$ changes

---

20 In Table 4, the uncorrected FCP’s for the CLR CI’s are 0.62 and 0.46 in the first line and the same as reported in Table 4 in the other lines. Note that without correction the FCP numbers are not comparable across different CI’s.

21 It is easy to allow for “selection on observables,” i.e., to allow $D_i$ to depend on $X_i$ and $Z_i$, if $\pi(x, z) = \Pr(D_i = 1|x, Z_i = z)$ is known, e.g., see Imbens (2004).
Nonparametric conditional treatment effects model: base-case comparisons. Table 5

<table>
<thead>
<tr>
<th>DGP</th>
<th>Critical value: Statistic</th>
<th>PA/Asy</th>
<th>PA/Bt</th>
<th>GMS/Asy</th>
<th>GMS/Bt</th>
<th>Sub</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat bound</td>
<td>CvM</td>
<td>0.040</td>
<td>0.054</td>
<td>0.044</td>
<td>0.063</td>
<td>0.106</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.028</td>
<td>0.039</td>
<td>0.031</td>
<td>0.046</td>
<td>0.231</td>
</tr>
<tr>
<td>Tilted bound</td>
<td>CvM</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.066</td>
<td>0.085</td>
<td>0.072</td>
<td>0.094</td>
<td>0.148</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.044</td>
<td>0.057</td>
<td>0.047</td>
<td>0.064</td>
<td>0.280</td>
</tr>
</tbody>
</table>

(b) Rejection probabilities under $H_1$ (null rejection probability corrected)

<table>
<thead>
<tr>
<th>DGP</th>
<th>Statistic</th>
<th>PA/Asy</th>
<th>PA/Bt</th>
<th>GMS/Asy</th>
<th>GMS/Bt</th>
<th>Sub</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat bound</td>
<td>CvM</td>
<td>0.50</td>
<td>0.57</td>
<td>0.51</td>
<td>0.54</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.30</td>
<td>0.42</td>
<td>0.30</td>
<td>0.42</td>
<td>0.35</td>
</tr>
<tr>
<td>Tilted bound</td>
<td>CvM</td>
<td>0.32</td>
<td>0.24</td>
<td>0.52</td>
<td>0.59</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.37</td>
<td>0.19</td>
<td>0.49</td>
<td>0.53</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.36</td>
<td>0.46</td>
<td>0.36</td>
<td>0.44</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Fig. 2. Nonparametric treatment effects model: conditional moment functions, as functions of $x$ and $z$. These functions are for the lower end of the identified set for $	heta$. First column: flat bound function; second column: kinked bound function; third column: tilted bound function.

Table 6

Nonparametric conditional treatment effects model with flat bound: variations on the base case.

<table>
<thead>
<tr>
<th>Case</th>
<th>(a) Null rejection probabilities (nominal 5%)</th>
<th>(b) Rej. probs. under $H_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Statistic</td>
<td>Proportions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flat bound</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tilted bound</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) Null rejection probabilities (nominal 5%)

(b) Rejection probabilities under $H_1$ (null rejection probability corrected)

sign in any neighborhood of $z_0$ if $a = 0$. The conditional moment functions that correspond to $\mu_1$, $\mu_2$, and $\mu_3$ are graphed in Fig. 2.

Notice that there is only one conditional moment inequality in this model (i.e., $p = 1$ and $v = 0$). In consequence, the different $S$-functions, i.e., Sum, Max and QLR, are identical to each other and we do not distinguish them in the results reported below.

4.3.1. $g$ functions

The $g$ functions employed by the test statistics are indicator functions of hypercubes in $[0, 1]$, i.e., intervals, as in the example above. The regressor $X_i$ is transformed to lie in $(0, 1)$ by the same method as in the example above. The hypercubes have side-edge lengths $(2r)^{-1}$ for $r = r_0, \ldots, r_1$, where $r_0 = 1$ and the base case value of $r_1$ is $3$. The base case number of hypercubes is 12. We also report results for $r_1 = 2$ and 4, which yield 6 and 20 hypercubes, respectively.

4.3.2. Simulation results: tests proposed in this paper

Tables 5 and 6 report NRP’s and ARP’s, respectively, for a variety of test statistics and critical values proposed in this paper for a range of cases. The NRP’s are for $a = 0$ and the ARP’s are for $a > 0$.

Table 5 provides comparisons of the PA/Asy, PA/Bt, GMS/Asy, GMS/Bt, and Sub critical values for the CvM and KS test statistics. Table 6 provides robustness results for the CvM and KS test statistics in the flat bound case. Table 6 shows the degree of sensitivity of the results to (i) the sample size, (ii) the number of cubes employed, as indexed by $r_1$, (iii) the choice of $(\kappa_{n,b}, B_n)$ for the GMS/Asy critical values, (iv) the value of $\epsilon$, upon which the variance estimator $\Sigma^2_{\text{est}}(\theta, g)$ depends, and (v) the bandwidth $b$.

Table 5 shows that tests with the Asy versions of both the PA and GMS critical values have NRP’s less than or equal to the nominal level 0.05 with the flat bound and kinked bound DGP’s. The tilted bound DGP is a difficult case for NRP control because the conditional mean function changes sign at $z = z_0$ and the integral of

Note that, contrary to the previous simulation example, the DGP is different for the NRP’s and for the ARP’s. The null hypothesis stays the same. ARP’s are computed for $a$ equal to $c \times (250/n)^{1/4}$, where $c = 0.25, 1.05$, and 0.25 in the flat, kinked, and tilted bound cases, respectively. These points are chosen to yield similar values for the ARP’s across the different cases considered.
the mean function over any symmetric neighborhood around \( z_0 \) is negative under the DGP with \( a = 0 \). With this difficult DGP, tests with Asy critical values using the KS statistic have NRP's less than or equal to 0.05 and tests using the CvM statistic have NRP's slightly above 0.05. The tests using Bt critical values have noticeably greater over-rejection compared to their counterparts using Asy critical values. The tests using subsampling critical values with either the CvM or KS test statistic appear unreliable: their NRP's exceed 0.05 by a substantial amount with not only the tilted bound DGP but also the flat bound DGP. Note that all tests under-reject substantially in the kinked bound case in Table 5. This is because the conditional moment inequality is slack at all points \( x \) except one, but is not sufficiently slack that the moment selection criterion is able to ignore the moment conditions for many values of \( x \) when the GMS critical values are computed.

The ARP comparison in Table 5 shows (i) a clear advantage of CvM-based tests over KS-based tests, and (ii) clearly better performance of GMS-based tests compared to PA-based ones with the kinked bound DGP and similar performance of GMS and PA critical values with the flat and the tilted bound DGPs.

Table 6 provides results for the CvM and KS statistics coupled with the GMS/Asy critical values for several variations of the base case with the flat bound function. Analogous results for the kinked and tilted bound functions are given in Tables A3 and A4 in the Appendix, see Andrews and Shi (2013a). The results in Table 6 show little sensitivity to the sample size and a smaller \( \epsilon \) for the CvM-based test. TheARP performance of the KS-based test improves noticeably with the sample size, but stays much worse than that of the CvM-based test at all three sample sizes considered. There is some sensitivity to the number of cubes and the magnitude of the GMS tuning parameters \((k, B)\). Increasing the number of cubes or increasing \((k, B)\) reduces both the NRP's and the ARP's. As in the quantile selection example, there is some sensitivity to the bandwidth. A larger bandwidth leads to higher ARP's but still keeps the NRP's below 0.05. As discussed in the quantile selection example, this is closely related to the flatness of the bound and the same phenomenon does not occur with the other types of bounds, see Tables A3 and A4 in the Appendix, see Andrews and Shi (2013a).

The ARP results reported in Tables 5 and 6 are computed under DGP's with different \( a \) values \((a > 0)\) with the three different bound functions. For the CvM/Max/GMS/Asy test, the ARP's computed for the same value \( a = 0.25 \) for all three bound functions are 0.51, 0.00, and 0.53 for the flat, kinked, and peaked bound functions, respectively. Thus, the power is highest for the flat and tilted bound functions and worst for the kinked bound function.

In conclusion, the comparison between test statistics and critical values is largely consistent with the quantile selection example, with the CvM–GMS/Asy couple performing the best both in terms of NRP's and ARP's. The CvM–GMS/Asy couple has somewhat worse NRP than CvM–GMS/Asy. The performance of CvM–GMS/Asy is quite good over a range of sample sizes.

### 4.3.3. Simulation results: comparisons with CLR tests

Next, we compare NRP's and ARP's of the tests proposed in this paper with those of the series and local linear tests in CLR. The sample size \( n = 250 \). The parameter values at which the NRP's and ARP's are calculated are the same as in Table 5. The tests proposed in this paper, denoted AS, use the GMS/Asy critical values.

The results are reported in Table 7. The nominal 5% CvM AS test over-rejects somewhat in the tilted bound case with a NRP of 0.072. Its NRP in the flat and kinked bounded cases is less than 0.05. Both CLR tests over reject the null considerably in the flat and tilted bound cases. Specifically, the NRP's of the series CLR test are 0.103 and 0.104, respectively, while those of the local linear CLR test are 0.177 and 0.185, respectively. The power of the CvM AS test is substantially higher than that of the two CLR tests in the flat and tilted bound cases (being 0.51 versus 0.17 and 0.18 in the flat bound case and 0.53 versus 0.16 and 0.16 in the tilted bound case). In the kinked bound case, the power of the CvM AS test exceeds that of the series CLR test, but is lower than that of the local linear CLR test.

### Acknowledgments

Andrews gratefully acknowledges the research support of the National Science Foundation via grant numbers SES-0751517 and SES-1058376.

### References


23 In Table 7, the uncorrected ARP's of the CLR tests are 0.29 and 0.52 in the first line, the same as reported in Table 7 in the second line, and 0.29 and 0.53 in the third line. These uncorrected ARP's are not comparable across different tests.