ARTICLE IN PRESS

Journal of Econometrics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

On the choice of test statistic for conditional moment inequalities

Timothy B. Armstrong

Yale University, United States

ARTICLE INFO

Article history:
Received 15 January 2016
Received in revised form 12 July 2017
Accepted 31 October 2017
Available online xxxx

ABSTRACT

This paper derives asymptotic approximations to the power of Cramer–von Mises (CvM) style tests for inference on a finite dimensional parameter defined by conditional moment inequalities in the case where the parameter is set identified. Combined with power results for Kolmogorov–Smirnov (KS) tests, these results can be used to choose the optimal test statistic, weighting function and, for tests based on kernel estimates, kernel bandwidth. The results show that, in the setting considered here, KS tests are preferred to CvM tests, and that a truncated variance weighting is preferred to bounded weightings.

1. Introduction

This paper compares methods for inference on a parameter θ defined by the conditional moment inequalities

$$E(m(W_i,\theta)|X_i) \geq 0 \ \ a.s.$$

where $m : \mathbb{R}^{d_W+d_\theta} \to \mathbb{R}^{d_Y}$ is a known function of data W_i and a parameter $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$, and $\geq$ is defined elementwise. Here, W_i is a $\mathbb{R}^{d_W}$ valued random variable and X_i is a $\mathbb{R}^{d_X}$ valued random variable. We are given independent, identically distributed (iid) observations $\{(X_i', W_i')'\}_{i=1}^n$. This defines the identified set

$$\Theta_0 \equiv \{\theta \in \Theta | E(m(W_i,\theta)|X_i) \geq 0 \ \ a.s.\}$$

where $\Theta \subseteq \mathbb{R}^{d_\theta}$ is the parameter space. If Θ_0 contains more than one element, the model is said to be set identified.

Following Imbens and Manski (2004), we are interested in confidence regions $\mathcal{C}_n$ that satisfy the coverage criterion

$$\text{for all } \theta_0 \in \Theta_0,\ \liminf_{n\to\infty} P(\theta_0 \in \mathcal{C}_n) \geq 1-\alpha. \tag{1}$$

We consider confidence regions constructed by inverting a family of tests $\phi_n(\theta) = \phi_n(\theta, \{X_i, W_i\}_{i=1}^n)$, where $\phi_n(\theta)$ is a test of $H_{0,\theta} : \theta \in \Theta_0$:

$$\mathcal{C}_n = \{\theta | \phi_n(\theta) = 0\}.$$

Subject to the coverage criterion (1), we would like the confidence region $\mathcal{C}_n$ not to contain points that are far away from the identified set Θ_0. In particular, if we take a parameter θ_0 on the boundary of Θ_0 and consider a sequence $\theta_n = \theta_0 + a_n$ where $a_n \to 0$, we would like to have $\theta_n \notin \mathcal{C}_n$ with high probability for a_n converging to zero as quickly as possible (so long as θ_n approaches Θ_0 from the outside, rather than from the interior). Note that

$$P(\theta_n \notin \mathcal{C}_n) = P(\phi_n(\theta_n) = 1).$$

Thus, we can determine whether $\mathcal{C}_n$ contains points that are far away from Θ_0 by examining the behavior of $P(\phi_n(\theta_n) = 1)$, which is the power of the test $\phi_n(\theta_n)$ of H_{0,θ_n} at the alternative P.

This paper provides an asymptotic answer to this question by examining the asymptotic behavior of $P(\phi_n(\theta_n) = 1)$ as $n \to \infty$. We refer to limit of $P(\phi_n(\theta_n) = 1)$ as the local asymptotic power of the sequence of tests $\phi_n(\theta_n)$ (note that this terminology differs from definitions often used in the literature, since the null hypothesis varies with n while the alternative stays fixed). The local asymptotic power of this sequence of tests will depend on the distribution P, the parameter θ_0 on the boundary of Θ_0 to which the sequence $\theta_n = \theta_0 + a_n$ converges, and the sequence a_n.

This paper considers Cramer–von Mises (CvM) style test statistics, which integrate or add some function of the negative part of an objective function. These can be compared with existing results for Kolmogorov–Smirnov (KS) statistics, which take the minimum of an objective function. The results show that the power $P(\phi_n(\theta_n) = 1)$ will be greater asymptotically for KS statistics when the distribution P satisfies generic smoothness conditions of the form used in the nonparametric statistics literature. In particular, the results imply that KS statistics are preferred according to a "minimax within a smoothness class" criterion of the form used to formulate nonparametric relative efficiency results in papers such as Stone (1982).

As an example of the types of problems covered by this setup, consider the interval regression model of Manski and Tamer (2002). We observe (X_i, W_i^L, W_i^H) where $[W_i^L, W_i^H]$ is known to contain the latent variable W_i^*, which follows the linear regression model $E(W_i^*|X_i) = (1, X_i')\theta$. This falls into the setup of this paper

E-mail address: timothy.armstrong@yale.edu.

https://doi.org/10.1016/j.jeconom.2017.10.007
0304-4076/

with $W_i = (X_i, W_i^L, W_i^H)$ and $m(W_i, \theta) = (W_i^H - (1, X_i')\theta, (1, X_i')\theta - W_i^L)'$. The identified set is then given by

$$\Theta_0 = \{\theta | E(W_i^L|X_i) \le (1, X_i')\theta \le E(W_i^H|X_i) \ \ a.s.\}.$$

Thus, a parameter θ_0 in the identified set corresponds to a regression line $(1, x')\theta_0$ that is between the conditional means $E(W_i^L|X_i = x)$ and $E(W_i^H|X_i = x)$ for all x on the support of X_i. If θ_0 is on the boundary of the identified set, it will be equal to one of these regression lines for some value of x. For $\theta_n = \theta_0 + a_n$ approaching the boundary of the identified set from the outside, the regression line $(1, x')\theta_n$ will be above $E(W_i^H|X_i = x)$ or below $E(W_i^L|X_i = x)$ for some values of x, and we would like the test $\phi_n(\theta_n)$ to detect this so that $\theta_n \notin \mathcal{C}_n$ with high probability. We use primitive conditions to apply the general results in this paper to this setting, thereby giving asymptotic approximations to this probability. These conditions correspond to smoothness conditions used in the nonparametric statistics literature and conditions on the shape of these conditional means near points where one of them is equal to $(1, x')\theta_0$ (see Section 3.4, Appendix A.5).

The remainder of this paper is organized as follows. Section 1.1 defines the tests considered in this paper. Section 1.2 discusses related literature. Section 2 gives an intuitive description of the power results in this paper and how they are derived. Section 3 states formally the conditions used in this paper, and provides primitive conditions for the interval regression model. Section 4 derives the power results. Section 5 reports the results of a Monte Carlo study. Section 6 concludes. An appendix contains minimax power comparisons as well as primitive conditions for the results in the main text in additional settings. A supplementary appendix contains proofs and auxiliary results.

1.1. Definition of test statistics

The test statistics considered in this paper are as follows. Given a set $\mathcal{G}$ of nonnegative instruments, the null hypothesis $H_{0,\theta} : \theta \in \Theta_0$ implies that $E(m(W_i, \theta)g(X_i)) \ge 0$ for all $g \in \mathcal{G}$. Thus, under $H_{0,\theta} : \theta \in \Theta_0$, the sample analogue

$$E_n(m(W_i, \theta)g(X_i)) \equiv \frac{1}{n}\sum_{i=1}^{n} m(W_i, \theta)g(X_i) \tag{2}$$

should not be too negative for any $g \in \mathcal{G}$. The results in this paper use classes of functions given by kernels with varying bandwidths and location, given by $\mathcal{G} = \{x \mapsto k((x - \tilde{x})/h)|\tilde{x} \in \mathbb{R}^{d_X}, h \in \mathbb{R}_+\}$ for some kernel function k. With this choice of $\mathcal{G}$, $H_{0,\theta} : \theta \in \Theta_0$ holds if and only if $E(m(W_i, \theta)g(X_i)) \ge 0$ for all $g \in \mathcal{G}$, so that (2) can be used to form a consistent test (see Andrews and Shi, 2013 for a discussion of this and other choices of $\mathcal{G}$).

Alternatively, one can test $H_{0,\theta} : \theta \in \Theta_0$ by estimating $E(m(W_i, \theta)|X_i = x)$ directly using the kernel estimate

$$\hat{\bar{m}}_j(\theta, x) = \frac{\sum_{i=1}^{n} m(W_i, \theta)k((X_i - x)/h)}{\sum_{i=1}^{n} k((X_i - x)/h)} \tag{3}$$

for some sequence $h = h_n \to 0$ and kernel function k. If $H_{0,\theta}$ holds, (3) should not be too negative for any x.

Thus, a test statistic of the null that $\theta \in \Theta_0$ can be formed by taking any function that is positive and large in magnitude when (2) is negative and large in magnitude for some $g \in \mathcal{G}$, or when (3) is negative and large in magnitude for some x. One possibility is to use a CvM statistic that integrates the negative part of (2) over some measure μ on $\mathcal{G}$. This CvM statistic is given by

$$T_{n,p,\omega,\mu}(\theta) = \left[\int \sum_{j=1}^{d_Y} |E_n m_j(W_i, \theta)g(X_i)\omega_j(\theta, g)|_-^p \, d\mu(g)\right]^{1/p} \tag{4}$$

for some $p \ge 1$ and weighting ω, where $|t|_- = |\min\{t, 0\}|$. I refer to this as an instrument based CvM (IV-CvM) statistic. The CvM statistic based on the kernel estimate integrates the negative part of (3) against some weighting ω, and is given by

$$T_{n,p,\text{kern}}(\theta) = \left[\int \sum_{j=1}^{d_Y} \left|\hat{\bar{m}}_j(\theta, x)\omega_j(\theta, x)\right|_-^p dx\right]^{1/p} \tag{5}$$

for some $p \ge 1$. I refer to this as a kernel based CvM (kern-CvM) statistic.

For the instrument based CvM statistic, the scaling for the power function will depend on ω. This paper considers both a bounded weighting which, without loss of generality, can be taken to be constant (the measure μ can absorb any weighting that does not change with the sample size)

$$\omega_j(\theta, g) = 1 \text{ all } \theta, g, j \tag{6}$$

as well as the truncated variance weighting used for KS statistics by Armstrong (2014b), Armstrong and Chan (2016) and Chetverikov (2012), which is given by

$$\omega_j(\theta, g) = (\hat{\sigma}_j(\theta, g) \vee \sigma_n)^{-1} \tag{7}$$

where

$$\hat{\sigma}_j(\theta, g) = \{E_n[m_j(W_i, \theta)g(X_i)]^2 - [E_n m_j(W_i, \theta)g(X_i)]^2\}^{1/2}$$

and σ_n is a sequence converging to zero and $a \vee b$ denotes the maximum of a and b for scalars a and b.

The results for CvM statistics derived in this paper can be compared to power results for KS statistics derived in Armstrong (2015) and Armstrong (2014b). A KS statistic based on (2) simply takes the most negative value of that expression over $g \in \mathcal{G}$, and is given by

$$T_{n,\infty,\omega}(\theta) = \max_j \sup_{g \in \mathcal{G}} |E_n m_j(W_i, \theta)g(X_i)\omega_j(\theta, g)|_-. \tag{8}$$

I refer to this as an instrument based KS (IV-KS) statistic. A KS statistic based on (3) simply takes the most negative value of that expression over x, and is given by

$$T_{n,\infty,\text{kern}}(\theta) = \max_j \sup_{\theta} \left|\hat{\bar{m}}_j(\theta, x)\omega_j(\theta, x)\right|_-. \tag{9}$$

I refer to this as a kernel based KS (kern-KS) statistic. As with CvM statistics, the scaling for the local power function for the instrument based KS test depends on whether a bounded weighting or a truncated variance weighting is used.

To complete the definition of these tests, we need to define a critical value.[1] For tests that use instrument based CvM statistics with bounded weights or inverse variance weights with $p < \infty$, the test $\phi_{n,p,\omega,\mu}(\theta)$, which rejects when $\phi_{n,p,\omega,\mu}(\theta) = 1$, is defined as

$$\phi_{n,p,\omega,\mu}(\theta) = \begin{cases} 1 & \text{if } \sqrt{n}T_{n,p,\omega,\mu}(\theta) > \hat{c}_{n,p,\omega,\mu}(\theta) \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

for some critical value $\hat{c}_{n,p,\omega,\mu}(\theta)$. For kernel based CvM statistics, the test $\phi_{n,p,\text{kern}}(\theta)$, which rejects when $\phi_{n,p,\text{kern}}(\theta) = 1$, is defined as

$$\phi_{n,p,\text{kern}}(\theta) = \begin{cases} 1 & \text{if } (nh^{d_X})^{1/2}T_{n,p,\text{kern}}(\theta) > \hat{c}_{n,p,\text{kern}}(\theta) \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

While all of the new results in this paper are for CvM statistics, I refer to analogous results for KS statistics at some points for

[1] The results covered in this paper cover any critical value that is of the same order of magnitude asymptotically as a critical value based on the distribution where all moments bind. See Section 3.1 for details.