ON THE CHOICE OF TEST STATISTIC FOR CONDITIONAL MOMENT INEQUALITIES

By

Timothy B. Armstrong

October 2014

COWLES FOUNDATION DISCUSSION PAPER NO. 1960

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

http://cowles.econ.yale.edu/

On the Choice of Test Statistic for Conditional Moment Inequalities

Timothy B. Armstrong*
Yale University

October 16, 2014

Abstract

This paper derives asymptotic power functions for Cramer-von Mises (CvM) style tests for conditional moment inequality models in the set identified case. 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 KS tests are preferred to CvM tests, and that a truncated variance weighting is preferred to bounded weightings under a minimax criterion, and for a class of alternatives that arises naturally in these models. The results also provide insight into how moment selection and the choice of instruments affect power. Such considerations have a large effect on power for instrument based approaches when a CvM statistic or an unweighted KS statistic is used and relatively little effect on power with optimally weighted KS tests.

1 Introduction

This paper derives power functions for tests for conditional moment inequality models. The results show that, in a broad class of models, Kolmogorov-Smirnov (KS) style statistics, which take the infimum of an objective function, are more powerful than Cramer-von Mises (CvM) style statistics, which integrate or add some function of the negative part of an

*email: timothy.armstrong@yale.edu.

objective function, for detecting local alternatives under conditions that determine the minimax rate and arise naturally in set identified models. Thus, the results also show that KS statistics are preferred to CvM statistics under a minimax criterion in these models.

Combined with results from Armstrong (2011a) and Armstrong (2014), the results in this paper give clear prescriptions for the choice of test statistic in conditional moment inequality models in the set identified case, and provide insights into the choice of critical value as well. To the author's knowledge, this paper is the first to provide a theoretical justification for the choice of test statistic (CvM vs KS) based on power results, and for user defined procedures such as moment selection procedures and bandwidths for CvM statistics in this setting.

The main points can be summarized as follows. In this setting, KS statistics are preferred to CvM statistics in terms of asymptotic power, and a truncated variance weighting for the objective function like the one proposed in Armstrong (2014) is preferred to bounded weighting functions. The power comparisons are for local alternatives that determine the minimax rate, and can be argued to arise generically in set identified models (see Section A). If one prefers CvM statistics for other reasons, but wants them to perform well in the generic set identified case considered here, the results in this paper can be used to choose optimal weightings and, for the case where the CvM statistic is based on kernel estimates, optimal bandwidths (which differ from optimal bandwidths in other settings). If a KS statistic with the truncated variance weighting is used, alleviating nonsimilarity of the test through choice of the critical value has little effect on power. If a bounded weighting is used, alleviating nonsimilarity through the choice of the critical value can have a larger effect on power.

Formally, this paper considers tests of a null hypothesis of the form

$$E(m(W_i,\theta)|X_i) \geq 0 \;\; a.s. \tag{1}$$

where $m : \mathbb{R}^{d_W+d_\theta} \to \mathbb{R}^{d_Y}$ is a known function of data W_i and a parameter $\theta \in \Theta$, and $\geq$ is defined elementwise. 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. This paper derives the asymptotic power of several tests for detecting alternatives of the form $\theta_n = \theta_0 + a_n$, where θ_0 is on the boundary of Θ_0. The results use conditions that hold generically in the set identified case for a broad class of models (see Section A.1 of this paper as well as Armstrong, 2014, which verifies a similar set of conditions for a variety of models). These conditions also determine the minimax rate within certain smoothness classes, so that the relative efficiency results derived in this paper hold in a minimax sense.

The test statistics considered in this paper are as follows. Given a set $\mathcal{G}$ of nonnegative instruments, the null hypothesis (1) implies that $E(m(W_i,\theta)g(X_i)) \geq 0$ for all $g \in \mathcal{G}$. Thus, under (1), 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.

Alternatively, one can test (1) 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 the null hypothesis holds for θ, (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 $p \geq 1$ 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 $p \geq 1$. 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 (2014), Armstrong and Chan (2012) 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.

The results for CvM statistics derived in this paper can be compared to power results for KS statistics derived in Armstrong (2011a) and Armstrong (2014). 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

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

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.

The asymptotic power results derived in this paper for the CvM statistics (4) and (5) are summarized in Table 1. For comparison, Table 2 summarizes the corresponding results for KS statistics, which are contained in Armstrong (2011a) and Armstrong (2014). These tables give the fastest rate at which a_n can approach 0 for each test to have power at $\theta_0 + a_n$

statistic	weighting function	rate
instrument based CvM	bounded weights	$n^{-\gamma/\{2[d_X+\gamma+(d_X+1)/p]\}}$
instrument based CvM	variance weights	$n^{-\gamma/\{2[d_X/2+\gamma+(d_X+1)/p]\}}$
kernel CvM	-	$\max\{(nh^{d_X})^{-1/[2(1+d_X/(p\gamma))]}, h^\gamma\}$

Table 1: Local Power for CvM Statistics

statistic	weighting function	rate
instrument based KS	bounded weights	$n^{-\gamma/\{2[d_X+\gamma]\}}$
instrument based KS	variance weights	$(n/\log n)^{-\gamma/\{2[d_X/2+\gamma]\}}$
kernel KS	-	$\max\left\{(nh^{d_X}/\log n)^{-1/2}, h^\gamma\right\}$

Table 2: Local Power for KS Statistics (Armstrong, 2011a, 2014)

for θ_0 on the boundary of the identified set. Here γ is a smoothness parameter that, roughly speaking, corresponds to the number of derivatives, up to 2, of $E(m(W_i,\theta)|X_i = x)$ with respect to x. The power results for the instrument based statistics depend on the set of functions $\mathcal{G}$, and are reported here only for the ones considered in this paper, but broader implications of the results described here (such as KS statistics being more powerful than CvM statistics in this setting) hold more generally.

These power results have several implications for how the choice of test statistic and weighting affect power. First, tests based on KS statistics are more powerful than those based on the corresponding CvM statistic in all of these cases. Second, variance weights lead to more powerful tests than bounded weights both for CvM and KS statistics.

Third, the results can be used to choose the optimal bandwidth for kernel CvM statistics. Some calculation shows that the rate in the third row of Table 1 is optimized when h_n is proportional to $n^{-1/[2(\gamma+d_X/p+d_X/2)]}$, which leads to a rate of $n^{-\gamma/[2(\gamma+d_X/p+d_X/2)]}$. The optimal bandwidth is larger than the optimal bandwidth for estimating a conditional mean at a point, or for the corresponding KS statistic.

Fourth, it is interesting to note how the choice of the class of instrument functions $\mathcal{G}$ affects power for these statistics. The main point here is that choosing a larger class of instruments by adding instruments that turn out to be irrelevant has less impact on power for KS statistics than it does for CvM statistics. This can be seen by comparing the rates for instrument based statistics to the corresponding rates for kernel based statistics with the bandwidth chosen optimally. The rates reported in these tables for instrument based statistics take $\mathcal{G}$ to be the class of functions given by $x \mapsto k((x-\tilde{x})/h)$ for all $(\tilde{x}, h)$. The kernel version of this statistic essentially uses a subset of this class of functions with $h = h_n$ restricted to a particular value for each n. For KS statistics, as long as variance weights are used, considering this larger class of functions does not lead to a decrease in the rate for local alternatives even if the optimal h_n is known. The rate in the second row of Table 2 for variance weighted instrument based KS statistics is the same as the rate for kernel based KS statistics in the third row if h is chosen optimally. In general, adding more instruments to $\mathcal{G}$ will not lead to a slower rate in the power function for variance weighted KS statistics as long as certain conditions on the complexity of $\mathcal{G}$ hold.

In contrast, considering a larger set of instruments $\mathcal{G}$ will generally decrease the rate for local alternatives if a CvM statistic is used. If a kernel CvM statistic is used instead of an instrument based CvM statistic (which corresponds to restricting $\mathcal{G}$) and prior knowledge of the data generating process is used to choose the bandwidth optimally, the kernel statistic will achieve a $n^{-\gamma/[2(\gamma+d_X/p+d_X/2)]}$ rate, which is faster than the $n^{-\gamma/[2(d_X/2+\gamma+(d_X+1)/p)]}$ rate for the instrument based CvM statistic with variance weights, where $\mathcal{G}$ includes all bandwidths. It can also be shown, using arguments similar to those in this paper, that expanding $\mathcal{G}$ to include d_X-dimensional boxes with sides of different lengths leads to slower rates for power functions with CvM statistics, but not for KS statistics. In general, CvM statistics are more sensitive to adding functions to $\mathcal{G}$ than KS statistics.

These results provide general insight into the type of objective function, weighting, and critical value one should use. However, the class of tests that are optimal for these models (tests based on KS statistics with a truncated variance weighting) still depend on certain user defined parameters. Choosing these user defined parameters for a particular sample size and data set can be done using monte carlos and criteria such as maximizing power against a particular sequence of alternatives.

Tests based on instrument based CvM and KS statistics have been considered by Andrews and Shi (2013), Kim (2008), Khan and Tamer (2009) and Armstrong (2011a) for bounded weights, and Armstrong (2014), Armstrong and Chan (2012) and Chetverikov (2012) for KS statistics with variance weights. The statistics based on instruments with bounded weights use an approach to nonparametric testing problems that goes back at least to Bierens (1982). Aradillas-Lopez, Gandhi, and Quint (2013) use a slightly different version of an instrument CvM approach. Chernozhukov, Lee, and Rosen (2013) consider kernel based KS statistics and Lee, Song, and Whang (2013) consider kernel based CvM statistics. While some of these papers derive local power results for CvM tests under conditions that appear to be common in point identified models, these results do not apply in set identified models except for in very special cases. Indeed, the results in the present paper show that, when one uses a minimax criterion requiring uniformly good power in classes of underlying distributions defined by smoothness properties, the power of CvM tests is much worse (see Section A.2). The results in this paper show that power comparisons in the set identified case considered here are much different than settings that have been studied previously. Armstrong (2011a), Armstrong (2011b), Armstrong (2014), Armstrong and Chan (2012), and Chetverikov (2012) derive power results for KS statistics under conditions similar to those used in this paper, but do not consider CvM statistics. The local alternatives considered here are also related to "small peaked" alternatives used when considering minimax power in statistical testing problems relative to the supremum norm (see, e.g., Lepski and Tsybakov, 2000).

It should be emphasized that the power results in this paper apply to tests evaluated at alternative parameter values in the conditional moment inequality model given by (1). This motivates the definition of minimax power in Section A.2, and reflects the goal of inverting these tests to form a confidence region for points in the identified set (in the sense of Imbens and Manski, 2004), where the confidence region is as tight as possible. The literature described above allows one to test a null of the form $E(Y_i|X_i) \geq 0$ a.s. (or related nonparametric hypotheses such as stochastic monotonicity), which, of course, may be applied with $Y_i = m(W_i, \theta)$ to test (1), but may also be used in other nonparametric testing problems. The present paper gives prescriptions for getting good power at alternative parameter values in set identified conditional moment inequality models. In other settings (e.g. testing stochastic dominance), one may want to have power against different types of alternatives, and the prescriptions may be different.

Inference on conditional moment inequalities can also be cast as a problem of inference with many unconditional moment inequalities, as considered by Menzel (2010). The results of the present paper can be extended to provide power results for this case by allowing $\mathcal{G}$ to depend on n. This paper also relates to the broader literature on set identified models, including models defined by unconditional moment inequalities. See Armstrong (2011a) for additional references to this literature.

This paper is organized as follows. Section 2 gives an intuitive description of the power results in this paper and how they are derived. Section 3 defines the tests considered in this paper. Section 4 derives the power results. Section 5 reports the results of a monte carlo study. Section 6 concludes. An appendix contains proofs and auxiliary results, including minimax power comparisons as well as primitive conditions for the results in the main text in the interval regression model.

2 Intuition for the Results

To get some intuition for the results, consider the case of instrument based CvM statistic with bounded weights. This paper considers the case where the class of functions $\mathcal{G}$ is given by the set of kernel functions with varying bandwidths and locations $\{x \mapsto k((x-\tilde{x})/h) | \tilde{x} \in \mathbb{R}^{d_X}, h \in \mathbb{R}_+\}$ for some kernel function k, and the measure μ has a density $f_\mu(\tilde{x}, h)$ with respect to the Lebesgue measure. For simplicity, consider the case where $d_Y = 1$.

The test statistic is given by an integral over a sample expectation. We expect that the test will have power when the integral over the corresponding population expectation is large relative to the critical value, which, as discussed below, will be of order $n^{-1/2}$. Thus, to have power at $\theta_n = \theta_0 + a_n$, we expect that

$$\left[\int\int |Em(W_i,\theta_n)k((X_i-\tilde{x})/h)|_-^p f(\tilde{x},h)\, d\tilde{x}\, dh\right]^{1/p}$$
$$= \left[\int\int \left|\int \bar{m}(\theta_n,x)k((x-\tilde{x})/h)f_X(x)\, dx\right|_-^p f(\tilde{x},h)\, d\tilde{x}\, dh\right]^{1/p} \tag{10}$$

will have to be large relative to $n^{-1/2}$, where $\bar{m}(\theta_n, x) = E(m(W_i,\theta_n)|X_i = x)$ and $f_X(x)$ is the density of X_i.

This paper considers more general classes of data generating processes, but, for simplicity, suppose that $\bar{m}(\theta_0, x) \approx \|x - x_0\|^\gamma$ near some x_0 for some γ, and is bounded from below away from zero elsewhere. This approximation and a first order approximation to $m(\theta_n, x) - m(\theta_0, x)$ suggests that (10) will be approximated well by

$$\left\{\int\int \left|\int \left[\|x-x_0\|^\gamma + \bar{m}_\theta(\theta_0,x)a_n\right] k((x-\tilde{x})/h)f_X(x)\, dx\right|_-^p f(\tilde{x},h)\, d\tilde{x}\, dh\right\}^{1/p}$$

and since the integrand will be nonzero only for x and $\tilde{x}$ close to x_0 and h close to zero, we can further approximate this by

$$\left\{\int\int \left|\int \left[\|x-x_0\|^\gamma + \bar{m}_\theta(\theta_0,x_0)a_n\right] k((x-\tilde{x})/h)f_X(x_0)\, dx\right|_-^p f(x_0,0)\, d\tilde{x}\, dh\right\}^{1/p}.$$

Let $a_n = ar_n$ for some sequence r_n to be determined later. By the change of variables

$u = (x - x_0)/r_n^{1/\gamma}$, $v = (\tilde{x} - x_0)/r_n^{1/\gamma}$, $\tilde{h} = h/r_n^{1/\gamma}$, the above display can be written as

$$\left\{ \int \int \left| \int [r_n \|u\|^\gamma + \bar{m}_\theta(\theta_0, x_0) a r_n] \, k((u - v)/\tilde{h}) f_X(x_0) r_n^{d_X/\gamma} \, du \right|_-^p f(x_0, 0) r_n^{d_X/\gamma} \, dv \, r_n^{1/\gamma} \, d\tilde{h} \right\}^{1/p}$$

$$= r_n^{[(\gamma+d_X)+(d_X+1)/p]/\gamma} \left\{ \int \int \left| \int [\|u\|^\gamma + \bar{m}_\theta(\theta_0, x_0) a] \, k((u - v)/\tilde{h}) f_X(x_0) \, du \right|_-^p f(x_0, 0) \, dv \, d\tilde{h} \right\}^{1/p}.$$

Thus, (10) is of order $r_n^{[(\gamma+d_X)+(d_X+1)/p]/\gamma}$, so we expect to get power when this is large enough relative to $n^{-1/2}$, and equating these gives

$$r_n^{[(\gamma+d_X)+(d_X+1)/p]/\gamma} = n^{-1/2} \iff r_n = n^{-\gamma/\{2[(\gamma+d_X)+(d_X+1)/p]\}}.$$

This is the rate reported in Table 1 and derived formally later in the paper.

3 Definitions of Tests

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

$$\phi_{n,p,\omega,\mu} = \begin{cases} 1 & \text{if } \sqrt{n} T_{n,p,\omega,\mu} > \hat{c}_{n,p,\omega,\mu} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

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

$$\phi_{n,p,\text{kern}} = \begin{cases} 1 & \text{if } (nh^{d_X})^{1/2} T_{n,p,\text{kern}} > \hat{c}_{n,p,\text{kern}} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

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 comparison. For KS tests with bounded weights, the critical value is defined as in (11). For KS tests based on truncated variance weights, the test $\phi_{n,\infty,(\sigma \vee \sigma_n)^{-1}}$ is defined as

$$\phi_{n,\infty,(\sigma \vee \sigma_n)^{-1}} = \begin{cases} 1 & \text{if } \sqrt{\frac{n}{\log n}} T_{n,\infty,(\sigma \vee \sigma_n)^{-1}} > \hat{c}_{n,\infty,(\sigma \vee \sigma_n)^{-1}} \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

for some critical value $\hat{c}_{n,p,\infty,(\sigma\vee\sigma_n)^{-1}}$.

The properties of these tests will depend on the choice of critical value. The only condition I impose, stated in the following assumption, is that the critical value be of the same order of magnitude as a critical value based on a least favorable asymptotic distribution where all of the moments bind. To my knowledge, this covers all CvM procedures currently available in the literature (for KS statistics, Armstrong (2011a), Chernozhukov, Lee, and Rosen (2013) and Chetverikov (2012) consider critical values that may not satisfy this condition). In particular, this covers (1) the generalized moment selection (GMS) and plug-in asymptotic (PA) critical values proposed by Andrews and Shi (2013) for ω bounded, (2) GMS and PA critical values for variance weighted L_p statistics and (3) the critical values proposed by Lee, Song, and Whang (2013) for kernel based CvM statistics using the least favorable null dgp. Case (2) has not been considered in the literature and requires some new arguments, which I consider in the appendix. The critical value for case (3) is based on results in Lee, Song, and Whang (2013) showing that the scaled statistic converges in probability to a nonzero constant, and that scaling again around this constant gives a normal limit, but all that is needed for the power results in this paper is that $(nh^{d_X})^{1/2}T_{n,p,\text{kern}}$ be compared to a critical value that is bounded away from zero or converges to a positive constant. Note that while these critical values depend on the data generating process, they will satisfy Assumption 3.1 by definition, regardless of the data.

Assumption 3.1. *The critical value $\hat{c}$ defined in (11) or (12), depending on the weighting and form of the test, is bounded from below away from zero as n increases.*

Assumption 3.1 only gives a lower bound for a critical value. This gives bounds on the power function, but to derive the exact local asymptotic power function, we need the following condition, which gives a limiting value for this critical value. Under mild conditions on the data generating process and sequence of local alternatives, this assumption will also hold for the methods of choosing critical values discussed above.

Assumption 3.2. *For the critical value $\hat{c}$ defined in (11) or (13), depending on the weighting and form of the test, and some constant $c > 0$, $\hat{c} \xrightarrow{p} c$.*

The power properties of the test will also depend on the class of functions $\mathcal{G}$ used as instruments. I derive power functions for the case where $\mathcal{G}$ consists of kernel functions with different bandwidths and locations, defined in the following assumption.

Assumption 3.3. *For some bounded, nonnegative function k with finite support and $\int k(u)\,du > 0$, $\mathcal{G} = \{x \mapsto k((x-\tilde{x})/h) | \tilde{x} \in \mathbb{R}^{d_X}, h \in \mathbb{R}_+\}$, and the covering number $N(\varepsilon, \mathcal{G}, L_1(Q))$ defined in Pollard (1984) satisfies $\sup_Q N(\varepsilon, \mathcal{G}, L_1(Q)) \le A\varepsilon^{-W}$, where the supremum is over all probability measures.*

For CvM statistics, I place the following condition on the measure μ over which the sample means are integrated.

Assumption 3.4. *The measure μ has bounded support, and has a density $f_\mu(\tilde{x}, h)$ with respect to the Lebesgue measure on $\mathbb{R}^{d_X} \times [0, \infty)$ that is bounded and continuous.*

Relaxing this assumption would lead to different power properties, although the general point that L_p statistics perform worse in these models than supremum statistics would go through.

4 Local Power Results

In this section, I derive local power results for CvM test statistics under conditions similar to those used in Armstrong (2011a), Armstrong (2014) and Armstrong and Chan (2012). The conditions hold generically in many models used in practice in the set identified case for sequences of alternative values parameter values that approach some θ_0 on the boundary of the identified set. Section A.1 of the appendix verifies these conditions for the interval regression model, and Armstrong (2014) verifies a similar set of conditions in some other settings.

While these conditions use a fixed underlying distribution and a sequence of alternative parameter values, one can also use these results to bound the minimax uniform power of CvM statistics in certain classes of underlying distributions defined by smoothness parameters, and to show that they do not achieve the optimal minimax rate. This is shown formally in Section A.2 in the appendix. In particular, Section A.2 shows that the minimax rate for CvM statistics in certain smoothness classes is worse than the minimax rate for the corresponding KS statistics (the minimax rate for KS statistics follows from results in Armstrong, 2014). I assume throughout that the data are iid.

4.1 Conditions for Local Alternatives

I place the following conditions on the data generating process when $m(W_i, \theta)$ is evaluated at θ_0 and $\theta_n = \theta_0 + a_n$. In these conditions, γ is a smoothness parameter that is generally given by the minimum of the number of derivatives of the conditional mean and 2. The truncation of the smoothness parameter at 2 comes from the fact that the test statistics here use positive kernels or instruments.

The most common cases appear to be the case of a Lipschitz continuous conditional mean, which corresponds to $\gamma = 1$, and a twice differentiable conditional mean under certain boundary conditions, which corresponds to $\gamma = 2$. In both cases, the conditions below typically hold for parameters on the boundary of the identified set, and, if they do not hold for all parameter values, will generally hold for the parameter values and data generating processes that determine minimax rates (see Section A.2 and the discussion above). Cases where γ does not take on an integer value are common in models with set identification at infinity, such as the bounds for selection models with an instrument for selection in Manski (1990) (see Appendix B.3 of Armstrong, 2014, for primitive conditions for a similar set of assumptions for this model).

Assumption 4.1. *For some version of $E(m(W_i, \theta_0)|X_i)$, the conditional mean of each element of $m(W_i, \theta_0)$ takes its minimum only on a finite set $\{x|E(m_j(W_i, \theta_0)|X = x) = 0 \text{ some } j\} = \mathcal{X}_0 = \{x_1, \ldots, x_\ell\}$. For each k from 1 to ℓ, let $J(k)$ be the set of indices j for which $E(m_j(W_i, \theta_0)|X = x_k) = 0$. Assume that there exist neighborhoods $B(x_k)$ of each $x_k \in \mathcal{X}_0$ such that, for each k from 1 to ℓ, the following assumptions hold.*

i.) $E(m_j(W_i, \theta_0)|X_i)$ is bounded away from zero outside of $\cup_{k=1}^{\ell} B(x_k)$ for all j and, for $j \notin J(k)$, $E(m_j(W_i, \theta_0)|X_i)$ is bounded away from zero on $B(x_k)$.

ii.) For $j \in J(k)$, $\bar{m}_j(\theta_0, x) = E(m_j(W_i, \theta_0)|X = x)$ is continuous on $B(x_k)$ and satisfies

$$\sup_{\|x-x_k\| \leq \delta} \left\| \frac{\bar{m}_j(\theta_0, x) - \bar{m}_j(\theta_0, x_k)}{\|x - x_k\|^{\gamma(j,k)}} - \psi_{j,k}\left(\frac{x - x_k}{\|x - x_k\|}\right) \right\| \overset{\delta \to 0}{\to} 0$$

for some $\gamma(j, k) > 0$ and some function $\psi_{j,k} : \{t \in \mathbb{R}^{d_X} | \|t\| = 1\} \to \mathbb{R}$ with $\overline{\psi} \geq \psi_{j,k}(t) \geq \underline{\psi}$ for some $\overline{\psi} < \infty$ and $\underline{\psi} > 0$. For future reference, define $\gamma = \max_{j,k} \gamma(j, k)$ and $\tilde{J}(k) = \{j \in J(k)|\gamma(j, k) = \gamma\}$.

iii.) X has a continuous density f_X on $B(x_k)$.

iv.) For each $k \in \{1, \ldots, \ell\}$ and $j \in J(k)$, $s_j^2(x, \theta) \equiv var(m_j(W_i, \theta)|X_i = x)$ is strictly positive and continuous at (x_k, θ_0).

Assumption 4.2. *For each $x_k \in \mathcal{X}_0$, $\bar{m}(\theta, x)$ has a derivative as a function of θ in a neighborhood of (θ_0, x_k), denoted $\bar{m}_\theta(\theta, x)$, that is continuous as a function of (θ, x) at (θ_0, x_k)*

and, for any neighborhood of x_k, there is a neighborhood of θ_0 such that $\bar{m}_j(\theta, x)$ is bounded away from zero for θ in the given neighborhood of θ_0 and x outside of the given neighborhood of x_k for $j \in J(k)$ and for all x for $j \notin J(k)$.

Assumption 4.3. *The data are iid and for some fixed $\overline{Y} < \infty$ and θ in a some neighborhood of θ_0, $|m(W_i, \theta)| \le \overline{Y}$ with probability one.*

The following assumption, which is used for kernel based statistics, ensures that the kernel estimators do not encounter boundary problems (cf. Assumption 1(iii) in Lee, Song, and Whang, 2013).

Assumption 4.4. *X_i has a density f_X that is bounded away from zero and infinity on its support, and the weighting function $\omega_j(\theta, x)$ is continuous for all j and, for some $\varepsilon > 0$, is equal to zero whenever $f_X(\tilde{x}) < \varepsilon$ for some $\tilde{x}$ with $\|\tilde{x} - x\| < \varepsilon$.*

4.2 Instrument Based CvM Statistics with Bounded Weights

To describe the power results, we need some additional notation. Define

$$\begin{aligned}&\lambda_{\text{bdd}}(a, j, k, p) = \lambda_{\text{bdd}}(a, \bar{m}_{\theta,j}(\theta_0, x_k), \psi_{j,k}, f_X(x_k), f_\mu(x_k, 0), p) \\ &\equiv \int\int \left| \int \left[\|x\|^\gamma \psi_{j,k}\left(\frac{x}{\|x\|}\right) + \bar{m}_{\theta,j}(\theta_0, x_k)a \right] k((x - \tilde{x})/h) f_X(x_k)\, dx \right|_-^p f_\mu(x_k, 0)\, d\tilde{x}\, dh.\end{aligned}$$

Theorem 4.1. *Let*

$$a_n = a n^{-\gamma/\{2[d_X + \gamma + (d_X + 1)/p]\}}$$

for some vector a. Under Assumptions 3.3, 3.4, 4.1, 4.2 and 4.3,

$$n^{1/2} T_{n,p,1,\mu}(\theta_0 + a_n) \xrightarrow{p} \left(\sum_{k=1}^{|\mathcal{X}_0|} \sum_{j \in \tilde{J}(k)} \lambda_{bdd}(a, j, k, p) \right)^{1/p} \equiv r_{bdd}(a)$$

where $r_{bdd}(a) \to 0$ as $a \to 0$.

Theorem 4.1 has immediate consequences for the power of tests based on CvM statistics with bounded weightings.

Theorem 4.2. *If, in addition to the conditions of Theorem 4.1, Assumption 3.1 holds, the power*

$$E\phi_{n,p,1,\mu}(\theta_0 + a_n)$$

of the CvM test with bounded weights will converge to zero for $r_{bdd}(a) < c$. *If* a *is close enough to zero,* $r_{bdd}(a)$ *will be less than c so that the power will converge to zero under* $\theta_0 + a_n$. *If, in addition, Assumption 3.2 holds, the power under* $\theta_0 + a_n$ *given by the above display will converge to* 1 *for* $r_{bdd}(a) > c$.

The $n^{-\gamma/\{2[d_X+\gamma+(d_X+1)/p]\}}$ rate for instrument based CvM statistics with bounded weights is slower than the $n^{-\gamma/\{2[d_X+\gamma]\}}$ rate derived for the corresponding KS test in Theorem 14 of Armstrong (2011a) (for $\gamma = 2$) and Theorem 5.1 of Armstrong (2014) (α from that paper plays the role of γ here). Note also that local power increases as p increases, and becomes aribrarily close to the rate for the KS test as p increases.

4.3 Instrument Based CvM Statistics with Variance Weights

Define

$$\lambda_{\text{var}}(a,j,k,p) \\ \equiv \int\int\left|\int\left[\|x\|^{\gamma}\psi\left(\frac{x}{\|x\|}\right) + \bar{m}_{\theta,j}(\theta_0,x_k)a\right] w_j(x_k)h^{-d_X/2}k((x-\tilde{x})/h)f_X(x_k)\,dx\right|_-^p \\ f_\mu(x_k,0)\,d\tilde{x}\,dh$$

where $w_j(x_k) \equiv (s_j^2(x_k,\theta_0)f_X(x_k)\int k(u)^2\,du)^{-1/2}$.

Theorem 4.3. *Let*

$$a_n = an^{-\gamma/\{2[d_X/2+\gamma+(d_X+1)/p]\}}.$$

Suppose that $\sigma_n(n/\log n)^{1/2} \to \infty$ *and Assumptions 3.3, 3.4, 4.1, 4.2 and 4.3 hold. Then*

$$n^{1/2}T_{n,p,(\hat{\sigma}\vee\sigma_n)^{-1},\mu}(\theta_0+a_n) \le \left(\sum_{k=1}^{|\mathcal{X}_0|}\sum_{j\in J(k)}\lambda_{var}(a,j,k,p)\right)^{1/p} + o_p(1) \equiv r_{var}(a) + o_p(1)$$

where $r_{var}(a) \to 0$ as $a \to 0$. If, in addition, $\sigma_n n^{d_X/\{4[d_X/2+\gamma+(d_X+1)/p]\}} \to 0$, the above display will hold with the inequality replaced by equality.

The result has immediate consequences for the power of tests based on CvM statistics with truncated variance weightings.

Theorem 4.4. *Let a_n be defined as in Theorem 4.3 and suppose that the conditions of that theorem and Assumption 3.1 hold. The power of the test based on the CvM statistic with truncated variance weights*

$$E\phi_{n,p,(\sigma \vee \sigma_n)^{-1},\mu}(\theta_0 + a_n)$$

will converge to zero for $r_{var}(a) < c$. For a close enough to 0, $r_{var}(a)$ will be less than c so that the asymptotic power under $\theta_0 + a_n$ will be 0. If, in addition, Assumption 3.2 holds and $\sigma_n n^{d_X/\{4[d_X/2+\gamma+(d_X+1)/p]\}} \to 0$, the power function under $\theta_0 + a_n$ given by the above display will converge to 1 for $r_{var}(a) > c$.

As with bounded weighting functions, the rate for detecting local alternatives with CvM statistics with variance weights is slower than the rate for the corresponding KS test. The $n^{-\gamma/\{2[d_X/2+\gamma+(d_X+1)/p]\}}$ rate for variance weighted CvM statistics derived above contrasts with the $(n/\log n)^{-\gamma/[2(d_X/2+\gamma)]}$ rate for the corresponding KS test derived in Armstrong and Chan (2012) and Armstrong (2014) (the results from the latter paper on rates of convergence of confidence regions in the Hausdorff metric imply these local power results). The rate for CvM statistics approaches the rate for KS statistics as $p \to \infty$.

4.4 Statistics Based on Kernel Estimates

To describe the local asymptotic power functions, define

$$\lambda_{\text{kern}}(a, h, j, k, p) \equiv \int \left| \int \left[\|x\|^\gamma \psi_{j,k}\left(\frac{x}{\|x\|}\right) + \bar{m}_{\theta,j}(\theta_0, x_k)a \right] h^{-d_X} k((x - \tilde{x})/h)\omega_j(\theta_0, x_k)\, dx \right|_-^p d\tilde{x}.$$

and

$$\tilde{\lambda}_{\text{kern}}(a, j, k, p) \equiv \int \left| \left[[\|\tilde{x}\|^\gamma \psi_{j,k}\left(\frac{v}{\|v\|}\right) + \bar{m}_{\theta,j}(\theta_0, x_k)a \right] \omega_j(\theta_0, x_k) \right|_-^p dv.$$

Theorem 4.5. *Suppose that Assumptions 3.4, 4.1, 4.2, 4.3 and 4.4 hold, and that the kernel function k satisfies Assumption 3.3. In addition, suppose that the bandwidth h satisfies*

$h/n^{-s} \to c_h$ for some $0 < s < 1/d_X$ and $c_h > 0$, the kernel function k satisfies $\int k(u)\,du = 1$ and that the functions $\psi_{j,k}$ in Assumption 4.1 are continuous. Let $a_n = an^{-q}$ for some $a \in \mathbb{R}^{d_\theta}$ where

$$q = \begin{cases} s\gamma & \text{if } s < 1/[2(\gamma + d_X/p + d_X/2)] \\ (1 - sd_X)/[2(1 + d_X/(p\gamma))] & \text{if } s \geq 1/[2(\gamma + d_X/p + d_X/2)] \end{cases}$$

and let $\theta_n = \theta_0 + a_n$. If $s > 1/[2(\gamma + d_X/p + d_X/2)]$, then

$$(nh^{d_X})^{1/2} T_{n,p,kern}(\theta_n) \xrightarrow{p} c_h^{d_X/2} \left(\sum_{k=1}^{|\mathcal{X}_0|} \sum_{j \in J(k)} \tilde{\lambda}_{kern}(a, j, k, p) \right)^{1/p} \equiv \tilde{r}_{kern}(a).$$

If $s = 1/[2(\gamma + d_X/p + d_X/2)]$, then

$$(nh^{d_X})^{1/2} T_{n,p,kern}(\theta_n) \xrightarrow{p} c_h^{d_X/2} \left(\sum_{k=1}^{|\mathcal{X}_0|} \sum_{j \in J(k)} \lambda_{kern}(a, c_h, j, k, p) \right)^{1/p} \equiv r_{kern}(a, c_h).$$

If $s < 1/[2(\gamma + d_X/p + d_X/2)]$, then

$$(nh^{d_X})^{1/2} T_{n,p,kern}(\theta_n)$$

will converge in probability to 0 if

$$\left(\sum_{k=1}^{|\mathcal{X}_0|} \sum_{j \in J(k)} \lambda_{kern}(a, c_h, j, k, p) \right)^{1/p}$$

is 0 in a neighborhood of (a, c_h), and will converge to ∞ if this expression is strictly positive.

The result has immediate implications for the power of tests based on kernel CvM statistics.

Theorem 4.6. *Let a_n be defined as in Theorem 4.5 and suppose that the conditions of that theorem and Assumption 3.1 hold. If $s > 1/[2(\gamma + d_X/p + d_X/2)]$, the power of the test based on the kernel CvM statistic*

$$E\phi_{n,p,kern}(\theta_0 + a_n)$$

will converge to zero for $\tilde{r}_{kern}(a) < c$. If $s = 1/[2(\gamma + d_X/p + d_X/2)]$, the power given by the above display will converge to zero for $\tilde{r}_{kern}(a, c_h) < c$. If $s < 1/[2(\gamma + d_X/p + d_X/2)]$, the power given by the above display will converge to zero if $\tilde{r}_{kern}(a, c_h) = 0$ in a neighborhood of (a, c_h). If, in addition, Assumption 3.2 holds, the power given by the above display will converge to 1 if $\tilde{r}_{kern}(a) > c$, $r_{kern}(a, c_h) > c$, or $r_{kern}(a, c_h) > 0$ in the cases where s is greater than, equal to, or less than $1/[2(\gamma + d_X/p + d_X/2)]$ respectively.

As with instrument based statistics, the rate for detecting local alternatives with the kernel CvM test is slower than the rate for the corresponding KS statistic. The rate derived in Theorem 4.5 can be written as $\max\{(nh^{d_X})^{-1/[2(1+d_X/(p\gamma))]}, h^\gamma\}$, which is slower than the $\max\left\{(nh^{d_X}/\log n)^{-1/2}, h^\gamma\right\}$ rate for kernel based KS statistics derived in Armstrong (2014). As with the instrument based statistics, the CvM test is more powerful for p larger, and the rate approaches the rate for the KS test as p goes to ∞.

Theorem 4.5 can be used to choose the optimal bandwidth in this setting. The rate $a_n = an^{-q}$ is best when $s = 1/[2(\gamma + d_X/p + d_X/2)]$, which gives an exponent in the rate of

$$q = \frac{\gamma}{2(\gamma + d_X/p + d_X/2)} = \frac{1 - sd_X}{2(1 + d_X/(p\gamma))} = s\gamma.$$

Note that this rate is faster than the $n^{-\gamma/[2(d_X/2+\gamma+(d_X+1)/p))]}$ rate that can be obtained with instrument based CvM tests with variance weights. Thus, restricting the class of instruments using prior knowledge of the data generating process leads to a faster rate with CvM statistics. In contrast, instrument based KS statistics with variance weights can achieve the same rate as kernel KS statistics that use prior knowledge of the data generating process to choose the bandwidth optimally (cf. Armstrong, 2014; Armstrong and Chan, 2012; Chetverikov, 2012).

5 Monte Carlo

This section reports the results of a monte carlo study of the finite sample properties of the statistics considered in this paper. I perform monte carlos based on a median regression model with potentially endogenously missing data. I use the same data generating processes as for the monte carlos for variance weighted KS statistics in Armstrong and Chan (2012). A description of the model and data generating processes is repeated here for convenience.

The latent variable W_i^* follows a linear median regression model given the observed covariate X_i: $q_{1/2}(W_i^*|X_i) = \theta_1 + \theta_2 X_i$ where $q_{1/2}(W_i^*|X_i)$ is the conditional median of W_i^*

given X_i. Define $W_i^H = W_i^*$ when W_i^* is observed and $W_i^H = \infty$ otherwise. This gives the conditional moment inequality $E[I(\theta_1+\theta_2 X_i \le W_i^H) - 1/2|X_i] \ge 0$ a.s. (a similar inequality can be formed with the lower bound W_i^L defined analogously, but with $W_i^L = -\infty$ when W_i^* is unobserved, but the monte carlos focus on the inequality corresponding to W_i^H for simplicity). This model allows for arbitrary correlation between the "missingness" process and (W_i^*, X_i), so that the resulting bounds can be used to assess sensitivity to missingness at random assumptions that would point identify the model.

Each design uses data from the true model $W_i^* = \theta_1^* + \theta_2^* X_i + u_i$, where (θ_1^*, θ_2^*), $u_i \sim$ unif$(-1,1)$ and $u_i \sim$ unif$(-1,1)$. The outcome variable W_i^* is then set to be missing independently of W_i^* with probability $p(X_i)$ (note that, while the data are generated according to a missingness at random assumption and a particular parameter value, the tests are robust to failure of this assumption, which leads to a lack of point identification), where $p(x)$ is varied in each of three designs:

$$\begin{aligned}
&\text{Design 1:} \quad p(x) = .1\\
&\text{Design 2:} \quad p(x) = .02 + 2 \cdot .98 \cdot |x - .5|\\
&\text{Design 3:} \quad p(x) = .02 + 4 \cdot .98 \cdot (x - .5)^2.
\end{aligned}$$

For each design, the monte carlo power of each test is reported for $\theta = (\overline{\theta}_1 + a, 0)$ where $\overline{\theta}_1 = \sup\{\theta_1 | (\theta_1, 0) \in \Theta_0\}$ and a varies over the set $\{.1, .2., .3, .4, .5\}$. This leads to local alternatives that satisfy the conditions of this paper with $\gamma = 1$ for Design 2 and $\gamma = 2$ for Design 3. Design 1 leads to a flat conditional mean for which asymptotic theory predicts the following rates (for the instrument functions used here): $n^{-1/2}$ for kernel and instrument based CvM and unweighted instrument based KS statistics, $(n/\log n)^{-1/2}$ for variance weighted instrument KS statistics and $(nh/\log n)^{-1/2}$ for kernel KS statistics (see Andrews and Shi, 2013; Armstrong, 2014; Chernozhukov, Lee, and Rosen, 2013; Lee, Song, and Whang, 2013).

For the instrument based statistics, I use the class of functions $\{x \mapsto I(s < x < s+t) | 0 \le s \le s+t \le 1\}$ and the the Lebesgue measure on $\{(s,t) | 0 \le s \le s+t \le 1\}$ for μ for the instrument based CvM statistics. This corresponds to the multiscale kernel instruments in Assumption 3.3 with the uniform kernel. For the kernel based statistics, the uniform kernel is used, and the supremum or integral is taken over the set $[h/2, 1-h/2]$, so that the support of the kernel function is always contained in the support of X_i. For the CvM statistics, the simulations use the test with L_p exponent $p = 1$. For each test statistic, the critical value is taken from the least favorable null distribution, calculated exactly (up to monte carlo error) using the distribution under $(\overline{\theta}_1, 0)$ under Design 1. For the kernel estimators, the bandwidths $n^{-1/5}$, $n^{-1/3}$ and $n^{-1/2}$ are used, and, for the truncated variance weighted CvM statistics, the values $n^{-1/5}/4$, $n^{-1/3}/4$ and $n^{-1/2}/4$ are used for the truncation parameter σ_n^2 (this corresponds to truncating the variance of functions $I(s < x < s + t)$ with t less than $n^{-1/5}$, $n^{-1/3}$ and $n^{-1/2}$). For comparison, results for the variance weighted instrument KS statistic, which corresponds to the multiscale statistic of Armstrong and Chan (2012), are reported as well (taken directly from that paper).

Overall, the monte carlo results support the claim that, for the data generating processes and classes of instrument functions considered in the theoretical results in this paper, KS statistics perform better than CvM statistics. For Design 2 and Design 3, which follow the conditions of this paper with $\gamma = 1$ and $\gamma = 2$ respectively, the instrument based KS statistic has more power than the instrument based CvM statistic in basically all cases. For the kernel statistics, the KS test performs better unless the bandwidth is chosen to be much too small. For example, for Design 3, the optimal bandwidth for the kernel statistic is of order $n^{-1/5}$, and the kernel KS statistic performs better than the kernel CvM statistic with this bandwidth. However, the kernel statistic performs worse for smaller bandwidths when the sample size is not too large (although the KS statistic does almost as well or better with 1000 observations, suggesting that the asymptotics of Theorem 4.5 have started to kick in at this point).

For Design 1, asymptotic results from elsewhere in the literature predict that the instrument based statistics with the instruments used here perform about the same (in terms of the rate for detecting local alternatives) for KS and CvM statistics, although the variance weighted KS statistic performs slightly worse (by a $\log n$ factor). For kernel statistics, asymptotic theory predicts that KS statistics will perform worse than CvM statistics in this case (the latter can achieve a $n^{-1/2}$ rate, while the former cannot if the bandwidth goes to zero). All of these predictions are borne out in the monte carlos: instrument based statistics all perform well with the weighted KS statistics performing slightly worse, while CvM version is better for kernel statistics.

The monte carlo results also fit well with the prescription of the weighted instrument KS or "multiscale" statistic of Armstrong (2011b), Armstrong (2014), Armstrong and Chan (2012) and Chetverikov (2012) as the only test among the ones considered here that comes close to having the best power among these test statistics for all three monte carlo designs (according to asymptotic approximations, the weighted instrument KS test achieves the best rate to at least within a $\log n$ factor in all three cases, while each of the other statistics considered here performs worse by a polynomial factor in at least one case). While other statistics perform slightly better in certain cases, they perform much worse in others (e.g. the kernel KS statistic performs slightly better in Design 3 with the optimal bandwidth, $n^{-1/5}$, but performs much worse when other bandwidths are chosen, or with any bandwidth choice in Design 1).

6 Conclusion

This paper derives local power results for tests for conditional moment inequality models based on several forms of CvM statistics in the set identified case. The power comparisons hold under conditions that arise naturally in the set identified case, and determine the minimax rate. Combined with results for KS statistics, these results can be used to decide on the test statistic, weighting function, class of instruments and critical value to maximize power in these models. The results show that KS tests are preferred to CvM statistics and that variance weightings are preferred to bounded weightings, and allow the researcher to choose the bandwidth optimally when a kernel based approach is used. In addition, these results show that, while choosing the critical value based on moment selection procedures or restricting the class of instrument functions has relatively little effect on power with variance weighted KS statistics, these choices can have a large effect on power with CvM statistics or unweighted KS statistics.

A Primitive Conditions and Minimax Bounds

This appendix gives primitive condtions for the assumptions used in this paper, and shows how the (pointwise in the underlying distribution) results for local alternatives considered in the paper can be used to bound the minimax power of CvM tests in classes of underlying distributions where the conditional mean is constrained only by smoothness assumptions. Since the corresponding KS statistic has a faster rate in these classes, this justifies the claim that the CvM tests considered here perform worse in these models under a minimax criterion. Section A.1 provides primitive conditions for the interval regression model. Section A.2 uses the results in the body of this paper to give conditions under which the CvM statistics considered in this paper do not achieve the optimal rate minimax rate, and verifies these conditions for the interval regression model.

A.1 Interval Regression

For the interval regression model, 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 with $W_i = (X_i, W_i^L, W_i^H)$ and $m(W_i, \theta) = (W_i^H - (1, X_i')\theta, (1, X_i')\theta - W_i^L)'$. Consider a data generating process and a parameter value θ_0 on the boundary of the identified set under this data generating process that satisfy the following assumptions.

Assumption A.1. *i.) The conditional means $E(W_i^H|X_i = x)$ and $E(W_i^L|X_i = x)$ are twice differentiable with continuous second derivatives, X_i has a continuous density and compact support, and W_i^H and W_i^L are bounded from above and below by finite constants.*

ii.) The set $\mathcal{X}_0 \equiv \{x|E(W_i^H|X_i = x) = (1, x')\theta_0\}$ is finite, and, for any point $\tilde{x} \in \mathcal{X}_0$, $\tilde{x}$ is in the interior of the support of X_i, $var(W_i^H|X_i = x)$ is positive and continuous at $\tilde{x}$ and $E(W_i^H|X_i = x)$ has a positive definite second derivative matrix at $\tilde{x}$. The same holds for $E(W_i^L|X_i = x)$ with "positive definite" replaced by "negative definite."

Theorem A.1. *Under Assumption A.1, Assumptions 4.1, 4.2 and 4.3 hold, with $\gamma = 2$ in Assumption 4.1.*

Proof. Part (ii) of Assumption 4.1 follows from a second order taylor expansion, and part (i) follows by compactness of the support of X_i and continuity of the first two derivatives of the conditional means. Part (iv) is immediate from part (ii) of Assumption A.1 and the fact that the conditional variance is constant in θ for this model. For Assumption 4.2, note that $\frac{d}{d\theta}\bar{m}_1(\theta, x) = -\frac{d}{d\theta}\bar{m}_2(\theta, x) = (1, x')$, which is clearly continuous in (θ, x). The second part of that assumption can be verified using this and the second order taylor expansion used to verify part (ii) of Assumption 4.1. Assumption 4.3 is immediate from the bounds on W_i^H and W_i^L. $\square$

For the Lipschitz case ($\gamma = 1$), we can replace the assumption of two derivatives with a condition on the directional first derivatives. In the following, $\mathbb{S}^{d_X-1}$ denotes the unit sphere $\{u \in \mathbb{R}^{d_X} | \|u\| = 1\}$.

Assumption A.2. *i.) The conditional means $E(W_i^H|X_i = x)$ and $E(W_i^L|X_i = x)$ are Lipschitz continuous, X_i has a continuous density and compact support, and W_i^H and W_i^L are bounded from above and below by finite constants.*

ii.) The set $\mathcal{X}_0 \equiv \{x|E(W_i^H|X_i = x) = (1, x')\theta_0\}$ *is finite, and, for any point* $\tilde{x} \in \mathcal{X}_0$, $\tilde{x}$ *is in the interior of the support of* X_i, $var(W_i^H|X_i = x)$ *is positive and continuous at* $\tilde{x}$ *and* $E(W_i^H|X_i = x)$ *has a directional derivative at* $\tilde{x}$ *in each direction* $u \in \mathbb{S}^{d_X-1}$ *such that* $\frac{d}{dt}[E(W_i^H|X_i = \tilde{x} + tu) - (1, (\tilde{x} + tu)')\theta_0]$ *is strictly positive and continuous at* $t = 0$ *uniformly over* $u \in \mathbb{S}^{d_X-1}$*. The same holds for* $E(W_i^L|X_i = x)$ *with "positive" replaced by "negative" in the last statement.*

Theorem A.2. *Under Assumption A.2, Assumptions 4.1, 4.2, and 4.3 hold, with* $\gamma = 1$ *in Assumption 4.1.*

Proof. Part (ii) of Assumption 4.1 follows from a first order taylor expansion, and part (i) follows by compactness of the support of X_i and the continuity and lower bound on the directional derivatives. The second part of Assumption 4.2 follows from the same reasoning used to verify part (ii) of Assumption 4.1. The verification of the remaining conditions is the same as in the twice differentiable case. □

A.2 Minimax Rates

The power results in this paper hold under conditions that are arguably common in practice in the set identified case. However, there are certainly cases (data generating processes, points on the boundary of the identified set and directions for the local alternative) for which other conditions will be appropriate. The purpose of this section is to show that, if the underlying distribution is constrained only by smoothness conditions and other regularity conditions, there will always exist a possible underlying distribution and sequence of local alternatives that satisfy these properties, with γ governed by the smoothness conditions imposed. Thus, any test that achieves good uniform power in these classes against alternatives that are closer than the pointwise rates derived here for CvM statistics will be preferred under a minimax criterion. By results in Armstrong (2014), it follows that, for certain classes of alternatives defined by smoothness conditions, the variance weighted KS statistic of Armstrong (2014), Armstrong and Chan (2012) and Chetverikov (2012) is preferred to the CvM statistics considered in this paper under a minimax criterion.

To formalize these ideas, the rest of this section considers classes $\mathcal{P}$ of underlying distributions and uses the notation E_P and $\Theta_0(P)$ to denote expectations and the identified set under a distribution P. In the results below, $d(\theta, \tilde{\theta})$ denotes the Euclidean distance $\|\theta - \tilde{\theta}\|$.

Theorem A.3. *Let* $\phi_{CvM}(\theta)$ *be one of the CvM tests defined in (11) or (12) with the critical value satisfying Assumption 3.1, the class* $\mathcal{G}$ *or kernel function* k *satisfying Assumption 3.3,* *and the measure μ satisfying Assumption 3.4 for the instrument case and the weighting satisfying Assumption 4.4 for the kernel case. Let $\mathcal{P}$ be any class of distributions such that, for some $P^* \in \mathcal{P}$ and θ_0^* on the boundary of $\Theta_0(P^*)$, Assumptions 4.1, 4.2 and 4.3 hold, and either (a) θ_0^* is on the boundary of the convex hull of $\Theta_0(P^*)$ or (b) for some $a \in \mathbb{R}^{d_\theta}$ and a constant K, $d(\theta_0^*, \theta_0^* + ar) \leq K \cdot d(\theta_0, \theta_0^* + ar)$ for all $\theta_0 \in \Theta_0(P^*)$ and r small enough. Then, for a small enough constant $C_* > 0$,*

$$\limsup_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta \text{ s.t. } d(\theta,\theta_0)\geq C_* r_n \text{ all } \theta_0\in\Theta_0(P)} E_P \phi_{CvM}(\theta) = 0,$$

where r_n depends on the test and is given in Table 1 with γ given in Assumption 4.1.

Proof. Under condition (b), the result is immediate from the results in the main text, since the quantity in the display in the theorem is less than $\limsup_{n\to\infty} E_{P^*}\phi_{CvM}(\theta_0^* + aC_* r_n K/\|a\|)$ for P^*, θ_0^* and a given in the theorem. The result follows since condition (a) implies condition (b) with $K = 1$. To see this, note that, by the supporting hyperplane theorem, there exists a vector a with $\|a\| = 1$ such that $a'\tilde{\theta}_0 \leq a'\theta_0^*$ for all $\tilde{\theta}_0$ in the convex hull of $\Theta_0(P^*)$. For this a and any scalar $r > 0$ and $\tilde{\theta}_0 \in \Theta_0(P^*)$, $d(\theta_0^*+ar, \tilde{\theta}_0)^2 - d(\theta_0^*+ar, \theta_0)^2 = \|\theta_0^*+ar-\tilde{\theta}_0\|^2 - r^2a'a = \|\theta_0^* - \tilde{\theta}_0\|^2 + 2ra'(\theta_0^* - \tilde{\theta}_0) + r^2a'a - r^2a'a \geq \|\theta_0^* - \tilde{\theta}_0\|^2 \geq 0$. □

A class $\mathcal{P}$ of underlying distributions will typically contain a P^* satisfying these conditions so long as it is sufficiently unrestricted (e.g. if the only restrictions are smoothness conditions, etc.). Theorems A.5 and A.6 below give primitive conditions for this in the interval regression model.

Under additional regularity conditions on $\mathcal{P}$, the inverse variance weighted KS statistic of Armstrong (2014), Armstrong and Chan (2012) and Chetverikov (2012) achieves a strictly better minimax rate than the upper bounds for CvM statistics given in Theorem A.3. This is stated in the next theorem, which follows immediately from results in Armstrong (2014) (the results in Armstrong, 2014 consider a stronger notion of coverage and power).

For concreteness, let us consider a specific version of the inverse variance weighted KS statistic considered in Armstrong (2014). Let $T_{n,\infty,(\sigma\vee\sigma_n)^{-1}}(\theta)$ be given by (8) with $\mathcal{G} = \{x \mapsto I(\|x - \tilde{x}\| \leq h) | \tilde{x} \in \mathbb{R}^{d_X}, h \in [0,\infty)\}$ and $\omega_j(\theta, g) = \{\hat{\sigma}_j(\theta, g) \vee [(\log n)^2/n]\}^{-1}$. Let $\phi_{n,\infty,(\sigma\vee\sigma_n)^{-1}}(\theta)$ be given by (13) with this definition of $T_{n,\infty,(\sigma\vee\sigma_n)^{-1}}(\theta)$ and with $\hat{c}_{n,\infty,(\sigma\vee\sigma_n)^{-1}}$ given by the constant K in Theorem 3.1 in Armstrong (2014). In the interest of concreteness, the above formulation uses certain conservative constants and tuning parameters in defining the test $\phi_{n,\infty,(\sigma\vee\sigma_n)^{-1}}(\theta)$. Less conservative and data driven methods for choosing these constants have been considered by Armstrong and Chan (2012) and Chetverikov (2012).

Theorem A.4. *Suppose that $\mathcal{P}$ satisfies Assumptions 4.1, 4.3, 4.4 and 4.5 in Armstrong (2014), with γ taking the place of α in that paper. Then* $\limsup_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{\theta_0\in\Theta_0(P)} E_P\phi_{n,\infty,(\sigma\vee\sigma_n)^{-1}}(\theta_0) = 0$ *and, for a large enough constant C^*,*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta \text{ s.t. } d(\theta,\theta_0)\geq C^*[(\log n)/n]^{\gamma/(d_X+2\gamma)} \text{ all } \theta_0\in\Theta_0(P)} E_P\phi_{n,\infty,(\sigma\vee\sigma_n)^{-1}}(\theta) = 1.$$

Proof. Since Assumptions 3.1-3.3 in Armstrong (2014) follow by definition of the statistic, the result follows from Theorem 4.2 in that paper, with Assumption 4.2(i) in Armstrong (2014) following from Theorem 4.3 in that paper (since Assumption 4.6 and 4.2(ii) in that paper hold by construction). For $\mathcal{C}_n$ the setwise confidence set constructed from $\phi_{n,\infty,(\sigma\vee\sigma_n)^{-1}}(\theta)$ in Armstrong (2014),

$$\begin{aligned}
&\inf_{P\in\mathcal{P}} \inf_{\theta \text{ s.t. } d(\theta,\theta_0)\geq C^*[(\log n)/n]^{\gamma/(d_X+2\gamma)} \text{ all } \theta_0\in\Theta_0(P)} E_P\phi_{n,\infty,(\sigma\vee\sigma_n)^{-1}}(\theta) \\
&= \inf_{P\in\mathcal{P}} \inf_{\theta \text{ s.t. } d(\theta,\theta_0)\geq C^*[(\log n)/n]^{\gamma/(d_X+2\gamma)} \text{ all } \theta_0\in\Theta_0(P)} P(\theta\notin\mathcal{C}_n) \\
&\geq \inf_{P\in\mathcal{P}} P(\theta\notin\mathcal{C}_n \text{ all } \theta \text{ s.t. } d(\theta,\theta_0)\geq C^*[(\log n)/n]^{\gamma/(d_X+2\gamma)} \text{ all } \theta_0\in\Theta_0(P)) \\
&\geq \inf_{P\in\mathcal{P}} P(d_H(\Theta_0(P),\mathcal{C}_n) < C^*[(\log n)/n]^{\gamma/(d_X+2\gamma)})
\end{aligned}$$

where $d_H(A,B) = \max\{\sup_{a\in A}\inf_{b\in B} d(a,b), \sup_{b\in B}\inf_{a\in A} d(a,b)\}$ is the Hausdorff distance. This converges to 1 for large enough C^* by Theorem 4.2 in Armstrong (2014). □

The classes $\mathcal{P}$ used in Theorem A.4 impose smoothness conditions on the conditional mean along with a condition on the derivative of the conditional mean with respect to θ (cases where the latter condition fails appear to favor KS statistics over CvM statistics as well; see Section A.4 of Armstrong, 2014). Note that the rate given above for the weighted KS statistic $\phi_{n,\infty,(\sigma\vee\sigma_n)^{-1}}$ corresponds to the minimax L_∞ rate for nonparametric testing problems (Lepski and Tsybakov, 2000) and to the minimax rate for estimating a conditional mean (Stone, 1982; see Menzel, 2010 for related results for estimating the identified set in a setting similar to the one considered here). The results here show that the CvM statistics considered here do not achieve this rate, and in fact have a minimax rate that is worse by at least a polynomial amount.

I now turn to the interval regression model and consider primitive conditions. The next two theorems show that certain classes of underlying distributions for the interval regression model will always contain a distribution with a sequence of local alternatives that satisfy the conditions of this paper. The conclusion of Theorem A.3 then follows immediately, since the identified set is convex in the interval regression model. Theorem A.5 considers the case where the constraints on the conditional mean embodied in $\mathcal{P}$ essentially only restrict the conditional means of W_i^H and W_i^L to a Lipschitz smoothness class. Theorem A.6 considers the smoother case where a bound is placed on the second derivative. For primitive conditions for the conditions of Theorem A.4 in the interval regression model for the case where $d_X = 1$ and $\gamma = 1$ or 2, see Armstrong (2014), Section 6.2.

Theorem A.5. *Let $\mathcal{P}$ be any class of underlying distributions for (X_i, W_i^H, W_i^L) in the interval regression model such that, for all $P \in \mathcal{P}$, W_i^H and W_i^L are bounded and X_i has a continuous density on its support $\mathcal{X}_P$. Suppose that, for some set $\mathcal{X} \subseteq \mathbb{R}^{d_X}$ and some interval $[a, b]$, the following holds: for any function $f : \mathcal{X} \to [a, b]$ such that*

$$|f(x) - f(\tilde{x})| \leq K\|x - \tilde{x}\|,$$

there exists a $P \in \mathcal{P}$ such that $E_P(W_i^H|X_i) = f(X_i)$ and $E_P(W_i^L|X_i) \leq a$ almost surely, and $\mathcal{X}_P = \mathcal{X}$. Then there exists a $P^* \in \mathcal{P}$ and $\theta_0^* \in \Theta_0(P^*)$ that satisfies the conditions of Theorem A.3, with $\gamma = 1$ and $\psi_{j,k}(u) = K$ in Assumption 4.1.

Proof. Under these assumptions, there exists a distribution $P \in \mathcal{P}$ such that $E_P(W_i^H|X_i = x) = b - K[(\varepsilon - \|x - x_0\|) \vee 0]$ for some $\varepsilon > 0$ and x_0 on the interior of the support of X_i, and $E_P(W_i^L|X_i = x)$ is bounded from above away from $b - 2\varepsilon$. For $\theta = (b - K\varepsilon, 0)$, this satisfies the conditions of Theorem A.2. □

Theorem A.6. *Let $\mathcal{P}$ be any class of underlying distributions for (X_i, W_i^H, W_i^L) in the interval regression model such that, for all $P \in \mathcal{P}$, W_i^H and W_i^L are bounded and X_i has a continuous density on its support $\mathcal{X}_P$. Suppose that, for some set $\mathcal{X} \subseteq \mathbb{R}^{d_X}$ and some interval $[a, b]$, for any function $f : \mathcal{X} \to [a, b]$ such that*

$$\left|\frac{d^2}{dt^2} f(x + tu)\right| \leq K$$

for all $u \in \mathbb{R}^{d_X}$ with $\|u\| = 1$, there exists a $P \in \mathcal{P}$ such that $E_P(W_i^H|X_i) = f(X_i)$ and $E_P(W_i^L|X_i) \leq a$ almost surely, and $\mathcal{X}_P = \mathcal{X}$. Then there exists a $P^* \in \mathcal{P}$ and $\theta_0^* \in \Theta_0(P^*)$ that satisfies the conditions of Theorem A.3, with $\gamma = 2$ and $\psi_{j,k}(u) = K/2$ in Assumption 4.1.

Proof. The result follows by similar arguments to Theorem A.5 since a function can be constructed for $E_P(W_i^H|X_i = x)$ that has a unique interior minimum with second derivative matrix KI at its minimum and takes values between, say, $(a + b)/2$ and b. □

B Proofs and Auxiliary Results

Section B.1 contains auxiliary results used in the rest of this appendix. Section B.2 of this appendix derives critical values for CvM statistics with variance weights. Section B.3 contains proofs of the results in the body of the paper.

B.1 Auxiliary Results

We first state some results that extend or restate results on uniform convergence from Pollard (1984) (see also Armstrong, 2014). Throughout this section, we consider iid observations $Z_1, \ldots, Z_n$ and a sequence of classes of functions $\mathcal{F}_n$ on the sample space. Let $\sigma(f)^2 = Ef(Z_i)^2 - (Ef(Z_i))^2$ and let $\hat{\sigma}(f)^2 = E_n f(Z_i)^2 - (E_n f(Z_i))^2$.

Lemma B.1. *Suppose that* $|f(Z_i)| \leq \overline{f}$ *a.s. and that*

$$\sup_{n\in\mathbb{N}} \sup_{Q} N(\varepsilon, \mathcal{F}_n, L_1(Q)) \leq A\varepsilon^{-W}$$

for some A *and* W*, where* N *is the covering number defined in Pollard (1984) and the supremum over* Q *is over all probability measures. Let* σ_n *be a sequence of constants with* $\sigma_n\sqrt{n/\log n} \to \infty$*. Then, for some constant* C*,*

$$\frac{\sqrt{n}}{\sqrt{\log n}} \sup_{f\in\mathcal{F}_n} \left| \frac{(E_n - E)f(Z_i)}{\sigma(f) \vee \sigma_n} \right| \leq C$$

with probability approaching one and

$$\sup_{f\in\mathcal{F}_n} \left| \frac{(E_n - E)f(Z_i)}{\sigma(f)^2 \vee \sigma_n^2} \right| \overset{p}{\to} 0.$$

Proof. The first display follows by applying Lemma A.1 in Armstrong (2014) to the sequence of classes of functions $\{f - E_P f(Z_i) | f \in \mathcal{F}_n\}$, which satisfies the conditions of that lemma by Lemma A.5 in Armstrong (2014). The second display follows from the first display since

$$\sup_{f\in\mathcal{F}_n} \left| \frac{(E_n - E)f(Z_i)}{\sigma(f)^2 \vee \sigma_n^2} \right| \leq \frac{1}{\sigma_n} \sup_{f\in\mathcal{F}_n} \left| \frac{(E_n - E)f(Z_i)}{\sigma(f) \vee \sigma_n} \right| = \frac{\sqrt{\log n}}{\sigma_n\sqrt{n}} \frac{\sqrt{n}}{\sqrt{\log n}} \sup_{f\in\mathcal{F}_n} \left| \frac{(E_n - E)f(Z_i)}{\sigma(f) \vee \sigma_n} \right|$$

and $\sqrt{\log n}/(\sigma_n\sqrt{n}) \to 0$. □

Lemma B.2. *Under the conditions of Lemma B.1,*

$$\sup_{f\in\mathcal{F}_n}\left|\frac{\hat\sigma(f)\vee\sigma_n}{\sigma(f)\vee\sigma_n}-1\right|\xrightarrow{p}0.$$

Proof. By continuity of $t\mapsto\sqrt{t}$ at 1, it suffices to prove that $\sup_{f\in\mathcal{F}_n}\left|\frac{\hat\sigma(f)^2\vee\sigma_n^2}{\sigma(f)^2\vee\sigma_n^2}-1\right|\xrightarrow{p}0$. We have

$$\sup_{f\in\mathcal{F}_n}\left|\frac{\hat\sigma(f)^2\vee\sigma_n^2}{\sigma(f)^2\vee\sigma_n^2}-1\right|=\sup_{f\in\mathcal{F}_n}\left|\frac{\hat\sigma(f)^2\vee\sigma_n^2-\sigma(f)^2\vee\sigma_n^2}{\sigma(f)^2\vee\sigma_n^2}\right|\le\sup_{f\in\mathcal{F}_n}\left|\frac{\hat\sigma(f)^2-\sigma(f)^2}{\sigma(f)^2\vee\sigma_n^2}\right|.$$

Note that

$$\hat\sigma(f)^2-\sigma(f)^2=(E_n-E)[f(Z_i)-Ef(Z_i)]^2-[(E_n-E)f(Z_i)]^2. \tag{14}$$

Since $\sigma[(f-Ef(Z_i))^2]^2\le E[f(Z_i)-Ef(Z_i)]^4\le 4\overline{f}^2\sigma(f)^2$, we have

$$\sup_{f\in\mathcal{F}_n}\frac{|(E_n-E)[f(Z_i)-Ef(Z_i)]^2|}{\sigma(f)^2\vee\sigma_n^2}\le\sup_{f\in\mathcal{F}_n}\frac{|(E_n-E)[f(Z_i)-Ef(Z_i)]^2|}{\sigma[(f-Ef(Z_i))^2]^2\vee\sigma_n^2}\cdot(4\overline{f}^2)\vee 1$$

which converges in probability to zero by Lemma B.1 (using Lemma A.5 in Armstrong, 2014 to verify that the sequence of classes of functions $\{[f-Ef(Z_i)]^2|f\in\mathcal{F}_n\}$ satisfies the conditions of the lemma). Since

$$\frac{[(E_n-E)f(Z_i)]^2}{\sigma(f)^2\vee\sigma_n^2}\xrightarrow{p}0$$

by Lemma B.1, the result now follows from this and the triangle inequality applied to (14). □

Lemma B.3. *Suppose that $|f(Z_i)|\le\overline{f}$ and that $\sigma_n\sqrt{n}\ge 1$. Then*

$$E\left|\frac{\sqrt{n}(E_n-E)f(Z_i)}{\sigma(f)\vee\sigma_n}\right|^p\le C_{p,\overline{f}}$$

for a constant $C_{p,\overline{f}}$ that depends only on p and $\overline{f}$.

Proof. By Bernstein's inequality,

$$P\left(\left|\frac{\sqrt{n}(E_n - E)f(Z_i)}{\sigma(f)\vee\sigma_n}\right| > t\right) \le \exp\left(-\frac{1}{2}\frac{n[\sigma(f)\vee\sigma_n]^2t^2}{n\sigma^2(f)+\frac{1}{3}\cdot 2\overline{f}\cdot\sqrt{n}[\sigma(f)\vee\sigma_n]t}\right)$$
$$\le \exp\left(-\frac{1}{2}\frac{t^2}{1+\frac{1}{3}\cdot 2\overline{f}\cdot\frac{t}{\sqrt{n}[\sigma(f)\vee\sigma_n]}}\right) \le \exp\left(-\frac{1}{2}\frac{t^2}{1+\frac{1}{3}\cdot 2\overline{f}\cdot t}\right) \le \exp\left(-\frac{1}{2}\frac{t^2}{1+\frac{1}{3}\cdot 2\overline{f}\cdot t}\right).$$

For $t \geq 1$, this is bounded by $\exp\left(-\frac{t}{2+\frac{2}{3}\cdot 2\overline{f}}\right)$. Thus,

$$E\left|\frac{\sqrt{n}(E_n - E)f(Z_i)}{\sigma(f)\vee\sigma_n}\right|^p = \int_{t=0}^{\infty} P\left(\left|\frac{\sqrt{n}(E_n - E)f(Z_i)}{\sigma(f)\vee\sigma_n}\right|^p > t\right)\,dt$$
$$\le 1 + \int_{t=1}^{\infty}\exp\left(-\frac{t^{1/p}}{2+\frac{2}{3}\cdot 2\overline{f}}\right)\,dt$$

which is finite and depends only on p and $\overline{f}$ as claimed. □

B.2 Critical Values for CvM Statistics with Variance Weights

For bounded choices of ω (which corresponds to σ_n bounded away from zero when a truncated variance weighting is used), Kim (2008) and Andrews and Shi (2013) derive a $\sqrt{n}$ rate of convergence to an asymptotic distribution that may be degenerate. Armstrong (2014) shows that letting σ_n go to zero generally decreases the rate of convergence to $\sqrt{n/\log n}$ for the KS statistic $T_{n,\infty,\omega}$. In contrast to the KS case, CvM statistics do not behave much differently if the variance is allowed to go to zero, although some additional arguments are needed to show this.

To deal with the behavior of the CvM statistic for small variances, I place the following condition on the measure over which the sample means are integrated.

Assumption B.1. *$\mu(\{g|\sigma_j(\theta, g) \le \delta\}) \to 0$ as $\delta \to 0$ for all j.*

This condition will hold for the choices of $\mathcal{G}$ and μ used in the body of the paper, and also allow for more general choices of $\mathcal{G}$ and μ. I also make the following assumption on the complexity of the class of functions $\mathcal{G}$, which is also satisfied by the class used in the paper.

Assumption B.2. *For some constants A and ε, the covering number $N(\varepsilon, \mathcal{G}, L_1(Q))$ defined* *in Pollard (1984) satisfies*

$$\sup_Q N(\varepsilon, \mathcal{G}, L_1(Q)) \leq A\varepsilon^{-W},$$

whre the supremum is over all probability measures.

The following condition imposes a bounded distribution of the function m.

Assumption B.3. *For some nonrandom constant $\overline{Y}$, $|m_j(W_i, \theta)| \leq \overline{Y}$ for each j with probability one.*

Theorem B.1. *Suppose that $\sigma_n \sqrt{n/\log n} \to \infty$ and that Assumptions B.1, B.2 and B.3 hold. Then, for $\theta \in \Theta_0$,*

$$n^{1/2} T_{n,p,(\hat{\sigma} \vee \sigma_n)^{-1},\mu}(\theta) \leq \left[\int \sum_{j=1}^{d_Y} \left| \frac{\sqrt{n}(E_n - E) m_j(W_i, \theta) g(X_i)}{\hat{\sigma}_j(\theta, g) \vee \sigma_n} \right|_-^p \, d\mu(g) \right]^{1/p}$$

$$\xrightarrow{d} \left[\int \sum_{j=1}^{d_Y} |\mathbb{G}_j(g, \theta)/\sigma_j(\theta, g)|_-^p \, d\mu(g) \right]^{1/p}$$

where $\mathbb{G}(g, \theta)$ is a vector of Gaussian processes with covariance function

$$\rho(g, \tilde{g}) = E[m(W_i, \theta) g(X_i) - Em(W_i, \theta) g(X_i)][m(W_i, \theta)\tilde{g}(X_i) - Em(W_i, \theta)\tilde{g}(X_i)]'.$$

Proof. The result with the integral truncated over $\{\sigma_j(\theta, g) \leq \delta | \text{all } j\}$ follows immediately from standard arguments using functional central limit theorems. This, along with Lemma B.4 below gives, letting $Z_n(\delta)$ be the integral truncated at $\{\sigma_j(\theta, g) \leq \delta | \text{all } j\}$ and $Z(\delta)$ be the limiting variable with this truncation,

$$P(Z_n(\delta) - \varepsilon \leq t) - \varepsilon \leq P(n^{1/2} T_{n,p,\omega,\mu}(\theta) \leq t) \leq P(Z_n(\delta) \leq t)$$

for large enough n for any $\varepsilon > 0$. The $\liminf$ of the left hand size is greater than $P(Z(\delta) \leq t - 2\varepsilon) - 2\varepsilon$, and the $\limsup$ of the right hand side is less than $P(Z(\delta) \leq t + \varepsilon) + \varepsilon$. We can bound $P(Z(\delta) \leq t - 2\varepsilon) - 2\varepsilon$ from below by $P(Z \leq t - 2\varepsilon) - 2\varepsilon$, and we can bound $P(Z(\delta) \leq t + \varepsilon) + \varepsilon$ from above by $P(Z \leq t + 2\varepsilon) + 2\varepsilon$ by making δ small enough by a version of Lemma B.4 for the limiting process. Since ε was arbitrary, this gives the result. □

The proof of the theorem above uses the following auxiliary lemma, which shows that functions g with low enough variance have little effect on the integral asymptotically.

Lemma B.4. *Fix j and suppose that Assumptions B.1, B.2 and B.3 hold, and that the null hypothesis holds under θ. Then, for every $\varepsilon > 0$, there exists a $\delta > 0$ such that*

$$P\left(\sqrt{n}\left[\int_{\sigma_j(\theta,g)\le\delta} |E_n m_j(W_i,\theta)g(X_i)/(\hat{\sigma}_j(\theta,g)\vee\sigma_n)|_-^p \, d\mu(g)\right]^{1/p} > \varepsilon\right) \le \varepsilon.$$

Proof. We have

$$\begin{aligned}
&E\int_{\sigma_j(\theta,g)\le\delta} |\sqrt{n}E_n m_j(W_i,\theta)g(X_i)/(\sigma_j(\theta,g)\vee\sigma_n)|_-^p \, d\mu(g) \\
&= \int_{\sigma_j(\theta,g)\le\delta} E|\sqrt{n}E_n m_j(W_i,\theta)g(X_i)/(\sigma_j(\theta,g)\vee\sigma_n)|_-^p \, d\mu(g) \\
&\le \int_{\sigma_j(\theta,g)\le\delta} E|\sqrt{n}(E_n - E) m_j(W_i,\theta)g(X_i)/(\sigma_j(\theta,g)\vee\sigma_n)|^p \, d\mu(g) \le \mu\left(\{g|\sigma_j(\theta,g)\le\delta\}\right)\cdot C_{p,\overline{Y}}
\end{aligned}$$

for $C_{p,\overline{Y}}$ given in Lemma B.3. Applying Markov's inequality and using Assumption B.1, it follows that, for any $\varepsilon > 0$, there exists a δ such that

$$P\left(\sqrt{n}\left[\int_{\sigma_j(\theta,g)\le\delta} |E_n m_j(W_i,\theta)g(X_i)/(\sigma_j(\theta,g)\vee\sigma_n)|_-^p \, d\mu(g)\right]^{1/p} > \varepsilon/2\right) \le \varepsilon/2.$$

The result follows since

$$\begin{aligned}
&\sqrt{n}\left[\int_{\sigma_j(\theta,g)\le\delta} |E_n m_j(W_i,\theta)g(X_i)/(\hat{\sigma}_j(\theta,g)\vee\sigma_n)|_-^p \, d\mu(g)\right]^{1/p} \\
&\le \sqrt{n}\left[\int_{\sigma_j(\theta,g)\le\delta} |E_n m_j(W_i,\theta)g(X_i)/(\sigma_j(\theta,g)\vee\sigma_n)|_-^p \, d\mu(g)\right]^{1/p} \cdot \sup_g(\sigma_j(\theta,g)\vee\sigma_n)/(\hat{\sigma}_j(\theta,g)\vee\sigma_n)
\end{aligned}$$

and $\sup_g(\sigma_j(\theta,g)\vee\sigma_n)/(\hat{\sigma}_j(\theta,g)\vee\sigma_n) \le 2$ with probability approaching one by Lemma B.2. □

B.3 Proofs

This section contains proofs of the results in the body of the paper. The proofs use a number of auxiliary lemmas, which are stated and proved first. In the following, θ_n is always assumed to be a sequence converging to θ_0.

Lemma B.5. *Under the assumptions of Theorem 4.5, there exists a constant C such that*

$$\sup_{x\in\mathbb{R}^{d_X}} \frac{\sqrt{n}}{\sqrt{h^{d_X}\log n}} |(E_n - E)m(W_i,\theta_n)k((X_i - x)/h)| \leq C$$

and

$$\sup_{x\in\mathbb{R}^{d_X}} \frac{\sqrt{n}}{\sqrt{h^{d_X}\log n}} |(E_n - E)k((X_i - x)/h)| \leq C$$

with probability approaching one. In addition,

$$\sup_{\{x|\omega_j(\theta_n,x)>0 \text{ some } j\}} \left| \frac{E_n k((X_i - h)/h)}{Ek((X_i - h)/h)} - 1 \right| \xrightarrow{p} 0.$$

Proof. The first two displays follow from Lemma B.1 after noting that

$$var(m(W_i,\theta_n)k((X_i - x)/h)) \leq \overline{Y}^2\overline{k}^2\overline{f}_X B^{d_X} h^{d_X}$$

where $\overline{k}$ and $\overline{f}_X$ are bounds for k and f_X, and B is such that $k(u) = 0$ whenever $\max_{1\leq j\leq d_X} |u_j| > B/2$, and similarly for $var(k((X_i - x)/h))$, and that $\sqrt{h^{d_X}}\sqrt{n}/\sqrt{\log n} \to \infty$ under these assumptions.

For the last display, note that, for x such that $\omega_j(\theta_n, x) > 0$ for some j, $Ek((X_i-x)/h) \geq \underline{f}_X h^{d_X} \int k(u)\, du$ for large enough n, where $\underline{f}_X$ is a lower bound for the density of X_i on its support. Thus,

$$\sup_{\{x|\omega_j(\theta_n,x)>0 \text{ some } j\}} \left| \frac{E_n k((X_i - h)/h)}{Ek((X_i - h)/h)} - 1 \right| \leq \sup_{x\in\mathbb{R}^{d_X}} \left| \frac{(E_n - E)k((X_i - h)/h)}{\underline{f}_X h^{d_X} \int k(u)\, du} \right|$$
$$= \sup_{x\in\mathbb{R}^{d_X}} \frac{\sqrt{n}}{\sqrt{h^{d_X}\log n}} |(E_n - E)k((X_i - h)/h)| \cdot \frac{\sqrt{h^{d_X}\log n}}{\sqrt{n}\underline{f}_X h^{d_X} \int k(u)\, du}.$$

The result then follows from the second display, since $\frac{\sqrt{\log n}}{\sqrt{nh^{d_X}}} \to 0$. □

Let

$$\tilde{T}_{n,p,(\hat{\sigma}\vee\sigma_n)^{-1},\mu}(\theta) = \left[\int_{h>0}\int_x \sum_{j=1}^{d_Y} \left| \frac{E_n m(W_i,\theta)k((X_i - x)/h)}{\sigma_j(\theta,x,h)\vee\sigma_n} \right|_-^p f_\mu(x,h)\, dx\, dh\right]^{1/p}$$

and let

$$\tilde{T}_{n,p,\text{kern}}(\theta) = \left[\int_x \sum_{j=1}^{d_Y} \left| \frac{E_n m(W_i,\theta)k((X_i - x)/h)}{Ek((X_i - x)/h)} \right|_-^p \omega_j(\theta,x)\,dx\,dh\right]^{1/p}.$$

The notation $\sigma_j(\theta, \tilde{x}, h)$ is used to denote $\sigma_j(\theta, g)$ where $g(x) = k((x - \tilde{x})/h)$.

Lemma B.6. *Under Assumptions 3.3, 3.4, 4.1, 4.2 and 4.3,*

$$\sqrt{n}T_{n,p,(\hat{\sigma}\vee\sigma_n)^{-1},\mu}(\theta_n) = \sqrt{n}\tilde{T}_{n,p,(\hat{\sigma}\vee\sigma_n)^{-1},\mu}(\theta_n)(1 + o_P(1))$$

for any sequence $\theta_n \to \theta_0$. If Assumption 4.4 holds as well, then

$$(nh^{d_X})^{1/2}T_{n,p,kern}(\theta_n) = (nh^{d_X})^{1/2}\tilde{T}_{n,p,kern}(\theta_n)(1 + o_P(1))$$

for any sequence $\theta_n \to \theta_0$.

Proof. We have

$$|\sqrt{n}T_{n,p,(\hat{\sigma}\vee\sigma_n)^{-1},\mu}(\theta_n) - \sqrt{n}\tilde{T}_{n,p,(\hat{\sigma}\vee\sigma_n)^{-1},\mu}(\theta_n)| \le \sqrt{n}\tilde{T}_{n,p,(\hat{\sigma}\vee\sigma_n)^{-1},\mu}(\theta) \cdot \sup_{x,j} \left| \frac{\sigma_j(\theta_n,x,h)\vee\sigma_n}{\hat{\sigma}_j(\theta_n,x,h)\vee\sigma_n} - 1 \right|.$$

Thus, the first display follows from Lemma B.2.

Similarly, for the second display,

$$\begin{aligned}
&|(nh^{d_X})^{1/2}T_{n,p,\text{kern}}(\theta_n) - (nh^{d_X})^{1/2}\tilde{T}_{n,p,\text{kern}}(\theta_n)| \\
&\le (nh^{d_X})^{1/2}\tilde{T}_{n,p,\text{kern}}(\theta_n) \cdot \sup_{\{x|\omega_j(\theta,x)>0 \text{ some } j\}} \left| \frac{Ek((X_i - x)/h)}{E_n k((X_i - x)/h)} - 1 \right|,
\end{aligned}$$

and the result follows from Lemma B.5. □

Let

$$\tilde{\tilde{T}}_{n,p,(\hat{\sigma}\vee\sigma_n)^{-1},\mu}(\theta) = \left[\int_{h>0}\int_x \sum_{j=1}^{d_Y} \left| \frac{Em(W_i,\theta)k((X_i - x)/h)}{\sigma_j(\theta,x,h)\vee\sigma_n} \right|_-^p f_\mu(x,h)\,dx\,dh\right]^{1/p}$$

and let

$$\tilde{\tilde{T}}_{n,p,\text{kern}}(\theta) = \left[\int_x \sum_{j=1}^{d_Y} \left|\frac{Em(W_i,\theta)k((X_i-x)/h)}{Ek((X_i-x)/h)}\right|_-^p \omega_j(\theta,x)\,dx\,dh\right]^{1/p}.$$

Also define

$$\tilde{\tilde{T}}_{n,p,1,\mu}(\theta) = \left[\int_{h>0}\int_x \sum_{j=1}^{d_Y} |Em(W_i,\theta)k((X_i-x)/h)|_-^p\, f_\mu(x,h)\,dx\,dh\right]^{1/p}.$$

Lemma B.7. *Under Assumptions 3.3, 3.4, 4.1, 4.2 and 4.3,*

$$\sqrt{n}\tilde{T}_{n,p,(\hat{\sigma}\vee\sigma_n)^{-1},\mu}(\theta_n) = \sqrt{n}\tilde{\tilde{T}}_{n,p,(\hat{\sigma}\vee\sigma_n)^{-1},\mu}(\theta_n) + o_P(1).$$

and

$$\sqrt{n}T_{n,p,1,\mu}(\theta_n) = \sqrt{n}\tilde{\tilde{T}}_{n,p,1,\mu}(\theta_n) + o_P(1).$$

Proof. Let $\tilde{\sigma}_n \to 0$ be such that $\tilde{\sigma}_n\sqrt{n/\log n} \to \infty$ and $\tilde{\sigma}_n/\sigma_n \to 0$ (i.e. $\tilde{\sigma}_n$ is chosen to be much smaller than σ_n, but such that the assumptions still hold for $\tilde{\sigma}_n$). Note that

$$\begin{aligned}&\sqrt{n}|\tilde{\tilde{T}}_{n,p,(\hat{\sigma}\vee\sigma_n)^{-1},\mu}(\theta_n) - \tilde{T}_{n,p,(\hat{\sigma}\vee\sigma_n)^{-1},\mu}(\theta_n)|\\ &\le \left[\int\int_{(x,h)\in\hat{\mathcal{G}}} \sum_{j=1}^{d_Y} \left|\sqrt{n}\frac{(E_n-E)m(W_i,\theta_n)k((X_i-x)/h)}{\sigma_j(\theta,x,h)\vee\sigma_n}\right|^p f_\mu(x,h)\,dx\,dh\right]^{1/p}\end{aligned}$$

where $\hat{\mathcal{G}} = \{(x,h)|Em(W_i,\theta_n)k((X_i-x)/h) < 0 \text{ or } E_n(W_i,\theta_n)k((X_i-x)/h) < 0\}$.

For any $\varepsilon > 0$, there exists an $\eta > 0$ such that, for $h > \varepsilon$ and large enough n,

$$Em_j(W_i,\theta_n)k((X_i-x)/h) \ge \eta Ek((X_i-x)/h) \ge \eta\cdot var[m_j(W_i,\theta_n)k((X_i-x)/h)]\cdot\frac{1}{\overline{kY}^2}$$

where the second inequality follows since

$$var[m_j(W_i,\theta_n)k((X_i-x)/h)] \le \overline{Y}^2 E[k((X_i-x)/h)^2] \le \overline{Y}^2\overline{k}Ek((X_i-x)/h).$$

Thus, for large enough n we will have

$$
\begin{aligned}
&E_n m_j(W_i,\theta_n)k((X_i-x)/h) \\
&\geq (E_n-E)m_j(W_i,\theta_n)k((X_i-x)/h) + var[m_j(W_i,\theta_n)k((X_i-x)/h)]\cdot \frac{\eta}{\overline{kY}^2},
\end{aligned}
$$

and the last line is positive for all (x,h) with $\sigma_j(\theta_n,x,h)\geq\tilde{\sigma}_n$ with probability approaching one by Lemma B.1.

From this and the fact that $Em(W_i,\theta_n)k((X_i-x)/h)\geq 0$ for all $h>\varepsilon$ for large enough n, it follows that $\hat{\mathcal{G}}\subseteq\{(x,h)|h\leq\varepsilon \text{ or } \sigma_j(\theta_n,x,h)<\tilde{\sigma}_n\}$ with probability approaching one. Note that

$$
\begin{aligned}
&E\int\int_{\{(x,h)|h\leq\varepsilon\}}\sum_{j=1}^{d_Y}\left|\frac{\sqrt{n}(E_n-E)m(W_i,\theta_n)k((X_i-x)/h)}{\sigma_j(\theta,x,h)\vee\sigma_n}\right|^p f_\mu(x,h)\,dx\,dh \\
&=\int\int_{\{(x,h)|h\leq\varepsilon\}}\sum_{j=1}^{d_Y}E\left|\frac{\sqrt{n}(E_n-E)m(W_i,\theta_n)k((X_i-x)/h)}{\sigma_j(\theta,x,h)\vee\sigma_n}\right|^p f_\mu(x,h)\,dx\,dh
\end{aligned}
$$

by Fubini's theorem, and this can be made arbitrarily small by making ε small by Lemma B.3 and Assumption 3.4. Similarly,

$$
\begin{aligned}
&E\int\int_{\{(x,h)|\sigma_j(\theta_n,x,h)<\tilde{\sigma}_n \text{ some } j\}}\sum_{j=1}^{d_Y}\left|\frac{\sqrt{n}(E_n-E)m(W_i,\theta_n)k((X_i-x)/h)}{\sigma_j(\theta,x,h)\vee\sigma_n}\right|^p f_\mu(x,h)\,dx\,dh \\
&\leq \mu(\mathbb{R}^{d_X}\times[0,\infty))\cdot\sup_{\{(x,h,j)|\sigma_j(\theta_n,x,h)<\tilde{\sigma}_n\}}E\left|\frac{\sqrt{n}(E_n-E)m(W_i,\theta_n)k((X_i-x)/h)}{\sigma_j(\theta,x,h)\vee\sigma_n}\right|^p \\
&= \mu(\mathbb{R}^{d_X}\times[0,\infty))\cdot\sup_{\{(x,h,j)|\sigma_j(\theta_n,x,h)<\tilde{\sigma}_n\}}E\left|\frac{\sqrt{n}(E_n-E)m(W_i,\theta_n)k((X_i-x)/h)}{\sigma_j(\theta,x,h)\vee\tilde{\sigma}_n}\right|^p \frac{\tilde{\sigma}_n}{\sigma_n},
\end{aligned}
$$

which converges to zero by Lemma B.3. Using this and Markov's inequality, it follows that $\sqrt{n}|\tilde{\tilde{T}}_{n,p,(\hat{\sigma}\vee\sigma_n)^{-1},\mu}(\theta)-\tilde{T}_{n,p,(\hat{\sigma}\vee\sigma_n)^{-1},\mu}(\theta)|$ can be made arbitrarily small with probability approaching one by making ε small. This gives the first display of the lemma.

The second display follows by the same argument with σ_n set to the supremum of $\sigma_j(\theta,x,h)$ over x, h on the support of μ, θ in a neighborhood of θ_0 and all j. □

Lemma B.8. *Under Assumptions 3.3, 3.4, 4.1, 4.2, 4.3 and 4.4,*

$$(nh^{d_X})^{1/2}\tilde{T}_{n,p,kern}(\theta_n)=(nh^{d_X})^{1/2}\tilde{\tilde{T}}_{n,p,kern}(\theta_n)+o_P(1).$$

Proof. For any $\varepsilon > 0$, there is an $\eta > 0$ such that $Em_j(W_i, \theta_n)k((X_i - x)/h) > \eta Ek((X_i - x)/h)$ for all $x \in \bar{\mathcal{X}}(\varepsilon)$ where $\bar{\mathcal{X}}(\varepsilon)$ is the set of x with $\|x - x_k\| \geq \varepsilon$ for all $k = 1, \ldots, \ell$ and $\omega_j(\theta_n, x) > 0$ for some j. Thus, arguing as in Lemma B.7 and using Lemma B.5, it follows that, with probability approaching one,

$$
\begin{aligned}
&(nh^{d_X})^{1/2}|\tilde{T}_{n,p,\text{kern}}(\theta_n) - \tilde{\tilde{T}}_{n,p,\text{kern}}(\theta_n)| \\
&\leq \left[\int_{x \notin \bar{\mathcal{X}}(\varepsilon)} \sum_{j=1}^{d_Y} \left| \frac{\sqrt{nh^{d_X}}(E_n - E)m_j(W_i, \theta_n)k((X_i - x)/h)}{Ek((X_i - x)/h)} \right|^p \omega_j(\theta_n, x)\, dx \right]^{1/p}.
\end{aligned}
$$

Using Markov's inequality and Fubini's theorem along with the fact that $\int_{x \notin \bar{\mathcal{X}}(\varepsilon)} w_j(\theta_n x)\, dx$ can be made arbitrarily small by making ε small, the result follows so long as

$$
E \left| \frac{\sqrt{nh^{d_X}}(E_n - E)m_j(W_i, \theta_n)k((X_i - x)/h)}{Ek((X_i - x)/h)} \right|^p
$$

can be bounded uniformly over x such that $\omega_j(\theta_n, x) > 0$. But this follows from Lemma B.3, since, by Assumptions 3.3 and 4.4, for some $\delta > 0$, $Ek((X_i - x)/h) \geq \delta h^{d_X}$ for all x with $\omega_j(\theta_n, x) > 0$. □

For the following lemma, recall that $w_j(x_k) = (s_j^2(x_k, \theta_0) f_X(x_k) \int k(u)^2\, du)^{-1/2}$ and $s_j^2(x, \theta) = var(m(W_i, \theta)|X_i = x)$.

Lemma B.9. *Under Assumptions 3.3, 3.4, 4.1, 4.2 and 4.3, for $k = 1, \ldots, \ell$*

$$
\sup_{\|(x,h)-(x_k,0)\| \leq \varepsilon_n} \left| h^{-d_X/2} \sigma_j(\theta_n, x, h) - w_j(x_k)^{-1} \right| \to 0.
$$

for any sequences $\varepsilon_n \to 0$ and $\theta_n \to \theta_0$.

Proof. By differentiability of the square root function at $w_j^{-2}(x_k)$, it suffices to show that $\sup_{\|(x,h)-(x_k,0)\| \leq \varepsilon_n} \left| h^{-d_X} \sigma_j^2(\theta_n, x, h) - w_j^{-2}(x_k) \right| \to 0$. Note that

$$
\begin{aligned}
&h^{-d_X} \sigma_j^2(\theta_n, x, h) = h^{-d_X} E[m(W_i, \theta_n)^2 k((X_i - x)/h)^2] - h^{-d_X} \{E[m(W_i, \theta_n) k((X_i - x)/h)]\}^2 \\
&= h^{-d_X} \int s_j^2(\tilde{x}, \theta_n) k((\tilde{x} - x)/h)^2 f_X(\tilde{x})\, d\tilde{x} \\
&+ h^{-d_X} \int E[m(W_i, \theta_n)|X_i = \tilde{x}]^2 k((\tilde{x} - x)/h)^2 f_X(\tilde{x})\, d\tilde{x} \\
&- h^{-d_X} \left\{ \int E[m(W_i, \theta_n)|X_i = \tilde{x}] k((\tilde{x} - x)/h) f_X(\tilde{x})\, d\tilde{x} \right\}^2.
\end{aligned}
$$

By Assumption 3.3 and part (iii) of Assumption 4.1, the second term is bounded by a constant times $\sup_{\|(x,h)-(x_k,0)\|\le\varepsilon_n} E[m(W_i,\theta_n)|X_i = x]^2$, which converges to zero by continuity of $E[m(W_i,\theta)|X_i = x]$ at (θ_0, x_k). By Assumptions 3.3 and 4.1, the third term is bounded by a constant times $h^{-d_X} \cdot h^{2d_X} \le \varepsilon_n^{d_X}$ uniformly over (x,h) with $\|(x,h)-(x_k,0)\| \le \varepsilon_n$. Using a change of variables, the first term can be written as $\int s_j^2(x+uh,\theta_n)k(u)^2 f_X(x+uh)\,du$, which converges to $w_j^{-2}(x_k)$ uniformly over $\|(x,h)-(x_k,0)\| \le \varepsilon_n$ by continuity of s_j and f_X, and by Assumption 3.3. □

Lemma B.10. *Suppose that Assumptions 3.3, 3.4, 4.1, 4.2, 4.3 and 4.4 hold, and that* $\int k(u)\,du = 1$. *Then*

$$\sup_{\|x-x_k\|\le\varepsilon} |h^{-d_X} Ek((X_i-x)/h) - f_X(x_k)|$$

as $h \to 0$ *and* $\varepsilon \to 0$ *for* $k = 1, \ldots, \ell$.

Proof. We have

$$h^{-d_X} Ek((X_i-x)/h) = h^{-d_X} \int k((\tilde{x}-x)/h) f_X(\tilde{x})\,d\tilde{x} = \int k(u) f_X(x+uh)\,du,$$

and $\int k(u)\,du = 1$ and $f_X(x+uh)$ converges to $f_X(x_k)$ uniformly over $\|x-x_k\| \le \varepsilon$ and u in the support of k as $\varepsilon \to 0$ and $h \to 0$. □

For notational convenience in the following lemmas, define, for (j,k) with $j \in J(k)$,

$$\tilde{\psi}_{j,k}(x-x_k) = \frac{\bar{m}_j(\theta_0,x) - \bar{m}_j(\theta_0,x_k)}{\|x-x_k\|^{\gamma(j,k)}}$$

so that

$$\sup_{\|x-x_k\|<\delta} \left|\tilde{\psi}_{j,k}(x-x_k) - \psi_{j,k}\left(\frac{x-x_k}{\|x-x_k\|}\right)\right| \to 0$$

under Assumption 4.1.

Lemma B.11. *Under Assumptions 3.3, 3.4, 4.1, 4.2 and 4.3, for any $a \in \mathbb{R}^{d_\theta}$,*

$$r^{-[d_X+p(d_X+\gamma)+1]/\gamma} \int\int \sum_{j=1}^{d_Y} |Em_j(W_i,\theta_0+ra)k((X_i-\tilde{x})/h)|_-^p f_\mu(\tilde{x},h)\, d\tilde{x}\, dh$$

$$\overset{r\to 0}{\to} \sum_{k=1}^{\mathcal{X}_0} \sum_{j\in\tilde{J}(k)} \lambda_{bdd}(a,j,k,p).$$

Proof. For simplicity, assume that $\gamma(j,k)=\gamma$ for all j,k. The general result follows from applying the same arguments to show that areas of (x,h) near (j,k) with $\gamma(j,k)<\gamma$ do not matter asymptotically.

For C large enough, the integrand will be zero unless $\max\{\|x-x_k\|,h\}<Cr^{1/\gamma}$ for some k with $j\in J(k)$. Thus, it suffices to prove the lemma for, fixing (j,k) with $j\in J(k)$,

$$\int\int |Em_j(W_i,\theta_0+ra)k((X_i-\tilde{x})/h)|_-^p f_\mu(\tilde{x},h)\, d\tilde{x}\, dh$$
$$=\int\int \left|\int \bar{m}_j(\theta_0+ra,x)k((x-\tilde{x})/h)f_X(x)\,dx\right|_-^p f_\mu(\tilde{x},h)\,d\tilde{x}\,dh$$
$$=\int\int \left|\int [\|x-x_k\|^\gamma\tilde{\psi}_{j,k}(x-x_k)+\bar{m}_{\theta,j}(\theta^*(r),x)ra]k((x-\tilde{x})/h)f_X(x)\,dx\right|_-^p f_\mu(\tilde{x},h)\,d\tilde{x}\,dh$$

where the integrals are taken over $\|\tilde{x}-x_k\|<Cr^{1/\gamma}, h<Cr^{1/\gamma}$ and $\theta^*(r)$ is between θ_0 and θ_0+ra (we suppress the dependence of $\theta^*(r)$ on x in the notation). Using the change of variables $u=(x-x_k)/r^{1/\gamma}$, $v=(x-x_k)/r^{1/\gamma}$, $\tilde{h}=h/r^{1/\gamma}$, this is equal to

$$\int\int \left|\int [\|r^{1/\gamma}u\|^\gamma\tilde{\psi}_{j,k}(r^{1/\gamma}u)+\bar{m}_{\theta,j}(\theta^*(r),x_k+r^{1/\gamma}u)ra]k((u-v)/\tilde{h})f_X(x_k+r^{1/\gamma}u)r^{d_X/\gamma}\,du\right|_-^p$$
$$f_\mu(x_k+r^{1/\gamma}v,r^{1/\gamma}\tilde{h})r^{d_X/\gamma}\,dv r^{1/\gamma}\,d\tilde{h}$$
$$=r^{[d_X+1+p(\gamma+d_X)]/\gamma}\int\int \left|\int [\|u\|^\gamma\tilde{\psi}_{j,k}(r^{1/\gamma}u)+\bar{m}_{\theta,j}(\theta^*(r),x_k+r^{1/\gamma}u)a]k((u-v)/\tilde{h})f_X(x_k+r^{1/\gamma}u)\,du\right|_-^p$$
$$f_\mu(x_k+r^{1/\gamma}v,r^{1/\gamma}\tilde{h})\,dv\,d\tilde{h}$$

where the integrals are taken over $\|v\|<C,\tilde{h}<C$. The result now follows from the dominated convergence theorem (here, and in subsequent results involving sequences of the form $\int|\int g_n(z,w)\,d\mu(z)|_-^p\,d\nu(w)$, the dominated convergence theorem is applied to the inner integral for each w, and again to the outer integral).

□

Lemma B.12. *Under the conditions of Theorem 4.3, for any* $a \in \mathbb{R}^{d_\theta}$*,*

$$r^{-[d_X+p(d_X/2+\gamma)+1]/\gamma} \int\int \sum_{j=1}^{d_Y} |Em_j(W_i,\theta_0+ra)k((X_i-\tilde{x})/h)/(\sigma_j(\theta_0+ra,\tilde{x},h)\vee\sigma_n)|_-^p f_\mu(\tilde{x},h)\,d\tilde{x}\,dh$$

$$\leq \sum_{k=1}^{\mathcal{X}_0} \sum_{j\in\tilde{J}(k)} \lambda_{var}(a,j,k,p) + o(1)$$

for any $r = r_n \to 0$*. If, in addition,* $\sigma_n r_n^{-d_X/(2\gamma)} \to 0$*, the above display will hold with the inequality replaced by equality.*

Proof. As in the previous lemma, the following argument assumes, for simplicity, that $\gamma(j,k) = \gamma$ for all (j,k) with $j \in J(k)$. Let $\tilde{s}_j(r,\tilde{x},h) = \sigma_j(\theta_0+ra,\tilde{x},h)/h^{d_X/2}$. As before, for large enough C, the integrand will be zero unless $\max\{\|x-x_k\|, h\} < Cr^{1/\gamma}$ for some k with $j \in J(k)$. Thus, it suffices to prove the result for, fixing (j,k) with $j \in J(k)$,

$$\int\int |Em_j(W_i,\theta_0+ra)k((X_i-\tilde{x})/h)(h^{-d_X/2}\tilde{s}_j^{-1}(r,\tilde{x},h)\wedge\sigma_n^{-1})|_-^p f_\mu(\tilde{x},h)\,d\tilde{x}\,dh$$

$$= \int\int \sum_{j=1}^{d_Y} \left| \int [\|x-x_k\|^\gamma \tilde{\psi}_{j,k}(x-x_k) + \bar{m}_{\theta,j}(\theta^*(r),x)ra] \right.$$

$$\left. k((x-\tilde{x})/h)(h^{-d_X/2}\tilde{s}_j^{-1}(r,\tilde{x},h)\wedge\sigma_n^{-1})f_X(x)\,dx \right|_-^p f_\mu(\tilde{x},h)\,d\tilde{x}\,dh$$

where the integral is taken over $\|\tilde{x}-x_k\| < Cr^{1/\gamma}$, $h < Cr^{1/\gamma}$ and $\theta^*(r)$ is between θ_0 and θ_0+ra. Using the change of variables $u = (x-x_k)/r^{1/\gamma}$, $v = (\tilde{x}-x_k)/r^{1/\gamma}$,$\tilde{h} = h/r^{1/\gamma}$, this is equal to

$$\int\int \left| \int r[\|u\|^\gamma \tilde{\psi}_{j,k}(r^{1/\gamma}u) + \bar{m}_{\theta,j}(\theta^*(r),x_k+ur^{1/\gamma})a]k((u-v)/\tilde{h}) \right.$$

$$\left. (((r^{1/\gamma}\tilde{h})^{-d_X/2}\tilde{s}_j^{-1}(r,x_k+vr^{1/\gamma},r^{1/\gamma}\tilde{h}))\wedge\sigma_n^{-1})f_X(x_k+ur^{1/\gamma})r^{d_X/\gamma}\,du \right|_-^p$$

$$f_\mu(x_k+vr^{1/\gamma},r^{1/\gamma}\tilde{h})r^{d_X/\gamma}\,dv r^{1/\gamma}\,d\tilde{h}$$

$$= r^{[p(\gamma+d_X/2)+d_X+1]/\gamma} \int\int \left| \int [\|u\|^\gamma \tilde{\psi}_{j,k}(r^{1/\gamma}u) + \bar{m}_{\theta,j}(\theta^*(r),x_k+ur^{1/\gamma})a]k((u-v)/\tilde{h}) \right.$$

$$\left. ((\tilde{h}^{-d_X/2}\tilde{s}_j^{-1}(r,x_k+vr^{1/\gamma},r^{1/\gamma}\tilde{h}))\wedge(r^{d_X/(2\gamma)}\sigma_n^{-1}))f_X(x_k+ur^{1/\gamma})\,du \right|_-^p f_\mu(x_k+vr^{1/\gamma},r^{1/\gamma}\tilde{h})\,dv\,d\tilde{h}.$$

where the integral is taken over $\|v\| < C$, $h < C$. By Lemma B.9 and the dominated convergence theorem, this converges to $\lambda_{var}(a,j,k,p)$ if $\sigma_n r_n^{-d_X/(2\gamma)} \to 0$. If $\sigma_n r_n^{-d_X/(2\gamma)}$ does not converge to zero, the above display is bounded from above by the same expression with σ_n^{-1} replaced by ∞.

□

Lemma B.13. *Under the conditions of Theorem 4.5, for any $a \in \mathbb{R}^{d_\theta}$,*

$$r^{-(\gamma p+d_X)/\gamma}\int \sum_{j=1}^{d_Y} |[Em_j(W_i,\theta_0+ra)k((X_i-x)/h)/Ek((X_i-x)/h)]\omega_j(\theta_0+ra,x)|_-^p \, dx$$
$$\to \sum_{k=1}^{|\mathcal{X}_0|}\sum_{j\in J(k)} \lambda_{kern}(a,c_{h,r},j,k,p)$$

as $r \to 0$ with $h/r^{1/\gamma} \to c_{h,r}$ for $c_{h,r} > 0$. If the limit is zero for $(a, c_{h,r})$ in a neighborhood of the given values, the sequence will be exactly equal to zero for large enough r.

If $h/r^{1/\gamma} \to 0$, then, as $r \to 0$,

$$r^{-(\gamma p+d_X)/\gamma}\int \sum_{j=1}^{d_Y} |[Em_j(W_i,\theta_0+ra)k((X_i-x)/h)/Ek((X_i-x)/h)]\omega_j(\theta_0+ra,x)|_-^p \, dx$$
$$\to \sum_{k=1}^{|\mathcal{X}_0|}\sum_{j\in J(k)} \tilde{\lambda}_{kern}(a,j,k,p).$$

Proof. As before, this proof treats the case where $J(k) = \tilde{J}(k)$ for ease of exposition. As with the proofs of Lemmas B.11 and B.12, it suffices to prove the result for, fixing (j,k) with $j \in J(k)$,

$$\int |[Em_j(W_i,\theta_0+ra)k((X_i-\tilde{x})/h)/Ek((X_i-\tilde{x})/h)]\omega_j(\theta_0+ra,\tilde{x})|_-^p \, d\tilde{x}$$
$$= \int \left| \int [\|x-x_k\|^\gamma \tilde{\psi}_{j,k}(x-x_k) + \bar{m}_{\theta,j}(\theta^*(r),x)ra]k((x-\tilde{x})/h)f_X(x)\,dx h^{-d_X} b(\tilde{x})\omega_j(\theta_0+ra,\tilde{x}) \right|_-^p \, d\tilde{x}$$

where the integral is over $\|\tilde{x}-x_k\| < Cr^{1/\gamma}$ and $b(\tilde{x}) \equiv h^{d_X}/Ek((X_i-\tilde{x})/h)$ converges to $(f_X(x_k))^{-1}$ uniformly over $\tilde{x}$ in any shrinking neighborhood of x_k by Lemma B.10. Let $\tilde{h} = h/r^{1/\gamma}$. By the change of variables $u = (x-x_k)/r^{1/\gamma}$, $v = (\tilde{x}-x_k)/r^{1/\gamma}$, the above display is equal to

$$\int \left| \int [\|ur^{1/\gamma}\|^\gamma \tilde{\psi}_{j,k}(ur^{1/\gamma}) + \bar{m}_{\theta,j}(\theta^*(r), x_k + ur^{1/\gamma})ra]k((u-v)/\tilde{h}) f_X(x_k + ur^{1/\gamma}) r^{d_X/\gamma}\, du \right.$$
$$\left. (r^{1/\gamma}\tilde{h})^{-d_X} b(x_k + vr^{1/\gamma})\omega_j(\theta_0 + ra, x_k + r^{1/\gamma}v) \right|_-^p r^{d_X/\gamma}\, dv$$
$$= r^{p+d_X/\gamma} \int \left| \int [\|u\|^\gamma \tilde{\psi}_{j,k}(ur^{1/\gamma}) + \bar{m}_{\theta,j}(\theta^*(r), x_k + ur^{1/\gamma})a]k((u-v)/\tilde{h}) f_X(x_k + ur^{1/\gamma})\, du \right.$$
$$\left. \tilde{h}^{-d_X} b(x_k + vr^{1/\gamma})\omega_j(\theta_0 + ra, x_k + r^{1/\gamma}v) \right|_-^p dv \tag{15}$$

where the integral is over $v < C$. The first display of the lemma (the case where $h/r^{1/\gamma} \to c_{h,r}$ for $c_{h,r} > 0$) follows from this and the dominated convergence theorem.

To show that the sequence is exactly zero for small enough r when the limit is zero in a neighborhood of $(a, c_{h,r})$, note, that, if the limit is zero in a neighborhood of $(a, c_{h,r})$, we will have, for all $(\tilde{a}, \tilde{c}_{h,r})$ in this neighborhood and any v,

$$\int \left[\|u\|^\gamma \psi_{j,k}\left(\frac{u}{\|u\|}\right) + \bar{m}_{\theta,j}(\theta_0, x_k)\tilde{a} \right] k((u-v)/\tilde{c}_{h,r})\, du$$
$$= \int \left[\tilde{c}_{h,r}^\gamma \|\tilde{u}\|^\gamma \psi_{j,k}\left(\frac{u}{\|u\|}\right) + \bar{m}_{\theta,j}(\theta_0, x_k)\tilde{a} \right] k(\tilde{u} - \tilde{v})\, \tilde{c}_{h,r}^{d_X} d\tilde{u} \geq 0.$$

Evaluating this at $(\tilde{c}_{r,h}, \tilde{a})$ such that $\tilde{c}_{h,r}^\gamma \leq c_{h,r}^\gamma(1-\varepsilon)$ and (for the case where $\bar{m}_{\theta,j}(\theta_0, x_k)a$ is negative) $\bar{m}_{\theta,j}(\theta_0, x_k)\tilde{a} \leq (\bar{m}_{\theta,j}(\theta_0, x_k)a)(1+\varepsilon)$ shows that

$$\int \left[c_{h,r}^\gamma \|\tilde{u}\|^\gamma \psi_{j,k}\left(\frac{u}{\|u\|}\right) \cdot (1-\varepsilon) + (\bar{m}_{\theta,j}(\theta_0, x_k)a)(1+\varepsilon) \right] k(\tilde{u} - \tilde{v})\, d\tilde{u} \geq 0$$

for all v for some $\varepsilon > 0$. The above display is, for small enough r, a lower bound for the inner integral in (15) times a constant that does not depend on r, so that, for small enough r, the inner integral in (15) will be nonnegative for all v and (15) will eventually be equal to zero.

For the case where $\tilde{h} = h/r^{1/\gamma} \to 0$, multiplying (15) by $r^{-(p+d_X/\gamma)}$ gives, after the change of variables $\tilde{u} = (u-v)/\tilde{h}$,

$$\int \left| \int [\|\tilde{h}\tilde{u} + v\|^\gamma \tilde{\psi}_{j,k}((\tilde{h}\tilde{u} + v)r^{1/\gamma}) + \bar{m}_{\theta,j}(\theta^*(r), x_k + (\tilde{h}\tilde{u} + v)r^{1/\gamma})a]k(\tilde{u}) f_X(x_k + (\tilde{u}\tilde{h} + v)r^{1/\gamma})\, d\tilde{u} \right.$$
$$\left. b(x_k + vr^{1/\gamma})\omega_j(\theta_0 + ra, x_k + r^{1/\gamma}v) \right|_-^p dv$$

which converges to

$$\int |[\|v\|^\gamma \psi_{j,k}(v/\|v\|) + \bar{m}_{\theta,j}(\theta_0, x_k)a]\omega_j(\theta_0, x_k)|_-^p \, dv$$

as required by the dominated convergence theorem.

□

We are now ready for the proofs of the main results.

proof of Theorem 4.1. The result follows immediately from Lemmas B.7 and B.11 since $(n^{-\gamma/\{2[d_X+\gamma+(d_X+1)/p]\}})^{-[d_X+p(d_X+\gamma)+1]/(\gamma p)} = n^{1/2}$. □

proof of Theorem 4.3. The result follows immediately from Lemmas B.6, B.7 and B.12 since $(n^{-\gamma/\{2[d_X/2+\gamma+(d_X+1)/p]\}})^{-[d_X+p(d_X/2+\gamma)+1]/(\gamma p)} = n^{1/2}$. □

proof of Theorem 4.5. The result follows from Lemmas B.6, B.8 and B.13. Note that $(nh^{d_X})^{p/2}/(n^{1-d_X s})^{p/2}$ $c_h^{d_X p/2}$, and that, for the case where $s \geq 1/[2(\gamma + d_X/p + d_X/2)$,

$$(n^{-q})^{-(\gamma p+d_X)/(\gamma p)} = (n^{-(1-sd_X)/[2(1+d_X/(p\gamma))]})^{-(\gamma p+d_X)/(\gamma p)} = n^{(1-sd_X)/2}.$$

For the case where $s < 1/[2(\gamma + d_X/p + d_X/2)]$, it follows from Lemmas B.6, B.8 and B.13 that

$$n^{q(\gamma p+d_X)/(\gamma p)} T_n(\theta_0 + a_n) \xrightarrow{p} \left(\sum_{k=1}^{|\mathcal{X}_0|} \sum_{j \in J(k)} \lambda_{\text{kern}}(a, c_h, j, k, p) \right)^{1/p}$$

so that $(nh^{d_X})^{1/2} T_n(\theta_0 + a_n)$ will converge to ∞ in this case if the limit in the above display is strictly positive. If the limit in the above display is zero in a neighborhood of (a, c_h), it follows from Lemmas B.6 and B.8 that $(nh^{d_X})^{1/2} T_n(\theta_0 + a_n)$ is, up to $o_p(1)$, equal to a term that is zero for large enough n by Lemma B.13.

□

References

Andrews, D. W. K., and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81(2), 609–666.

Aradillas-Lopez, A., A. Gandhi, and D. Quint (2013): "Testing Inequalities of Conditional Moments, with an Application to Ascending Auction Models," .

Armstrong, T. (2011a): "Asymptotically Exact Inference in Conditional Moment Inequality Models," *Unpublished Manuscript.*

——— (2011b): "Weighted KS Statistics for Inference on Conditional Moment Inequalities," *Unpublished Manuscript.*

Armstrong, T., and H. P. Chan (2012): "Multiscale Adaptive Inference on Conditional Moment Inequalities," *Unpublished Manuscript.*

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

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

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

Chetverikov, D. (2012): "Adaptive Test of Conditional Moment Inequalities," *Unpublished Manuscript.*

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

Khan, S., and E. Tamer (2009): "Inference on endogenously censored regression models using conditional moment inequalities," *Journal of Econometrics*, 152(2), 104–119.

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

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

Lepski, O., and A. Tsybakov (2000): "Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point," *Probability Theory and Related Fields*, 117(1), 17–48.

Manski, C. F. (1990): "Nonparametric Bounds on Treatment Effects," *The American Economic Review*, 80(2), 319–323.

Menzel, K. (2010): "Consistent Estimation with Many Moment Inequalities," *Unpublished Manuscript*.

Pollard, D. (1984): *Convergence of stochastic processes*. Springer, New York, NY.

Stone, C. J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10(4), 1040–1053.

$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
0.1	0.196	0.593	0.818
0.2	0.458	0.973	1
0.3	0.775	1	1
0.4	0.952	1	1
0.5	0.995	1	1

Table 3: Power for Unweighted Instrument CvM Test under Design 1

$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
0.1	0.166	0.644	0.835
0.2	0.442	0.989	1
0.3	0.781	1	1
0.4	0.957	1	1
0.5	0.994	1	1

Table 4: Power for Unweighted Instrument KS Test under Design 1

σ_n^2	$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
	0.1	0.198	0.567	0.859
	0.2	0.49	0.977	1
$\frac{1}{4}n^{-1/5}$	0.3	0.77	1	1
	0.4	0.955	1	1
	0.5	0.997	1	1
	0.1	0.208	0.62	0.851
	0.2	0.475	0.983	1
$\frac{1}{4}n^{-1/3}$	0.3	0.808	1	1
	0.4	0.958	1	1
	0.5	0.994	1	1
	0.1	0.203	0.591	0.822
	0.2	0.474	0.981	1
$\frac{1}{4}n^{-1/2}$	0.3	0.804	1	1
	0.4	0.946	1	1
	0.5	0.996	1	1

Table 5: Power for Weighted Instrument CvM Test under Design 1

t_n	$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
	0.1	0.207	0.503	0.729
	0.2	0.48	0.954	1
$n^{-1/5}$	0.3	0.759	1	1
	0.4	0.956	1	1
	0.5	0.997	1	1
	0.1	0.144	0.453	0.63
	0.2	0.378	0.939	0.998
$n^{-1/3}$	0.3	0.691	1	1
	0.4	0.886	1	1
	0.5	0.982	1	1
	0.1	0.156	0.358	0.502
	0.2	0.348	0.898	0.991
$n^{-1/2}$	0.3	0.649	0.999	1
	0.4	0.862	1	1
	0.5	0.974	1	1

Table 6: Power for Weighted Instrument KS Test under Design 1 (from Armstrong and Chan (2012))

h_n	$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
	0.1	0.186	0.547	0.858
	0.2	0.453	0.97	1
$n^{-1/5}$	0.3	0.729	1	1
	0.4	0.934	1	1
	0.5	0.994	1	1
	0.1	0.188	0.663	0.843
	0.2	0.452	0.987	1
$n^{-1/3}$	0.3	0.794	1	1
	0.4	0.947	1	1
	0.5	0.997	1	1
	0.1	0.185	0.582	0.848
	0.2	0.443	0.977	1
$n^{-1/2}$	0.3	0.78	1	1
	0.4	0.942	1	1
	0.5	0.997	1	1

Table 7: Power for Kernel CvM Test under Design 1

h_n	$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
	0.1	0.16	0.439	0.625
	0.2	0.343	0.92	0.997
$n^{-1/5}$	0.3	0.62	0.999	1
	0.4	0.883	1	1
	0.5	0.975	1	1
	0.1	0.095	0.266	0.481
	0.2	0.201	0.715	0.929
$n^{-1/3}$	0.3	0.382	0.976	1
	0.4	0.606	0.999	1
	0.5	0.809	1	1
	0.1	0	0.094	0.138
	0.2	0	0.255	0.404
$n^{-1/2}$	0.3	0	0.508	0.773
	0.4	0	0.812	0.982
	0.5	0	0.976	1

Table 8: Power for Kernel KS Test under Design 1

$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
0.1	0	0	0
0.2	0.001	0	0
0.3	0.005	0	0
0.4	0.008	0.001	0.004
0.5	0.023	0.054	0.119

Table 9: Power for Unweighted Instrument CvM Test under Design 2

$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
0.1	0	0	0
0.2	0.003	0.002	0.001
0.3	0.007	0.022	0.037
0.4	0.01	0.145	0.412
0.5	0.039	0.596	0.884

Table 10: Power for Unweighted Instrument KS Test under Design 2

σ_n^2	$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
	0.1	0	0	0
	0.2	0	0	0
$\frac{1}{4}n^{-1/5}$	0.3	0.003	0	0
	0.4	0.007	0.006	0.013
	0.5	0.04	0.118	0.294
	0.1	0	0	0
	0.2	0	0	0
$\frac{1}{4}n^{-1/3}$	0.3	0.001	0.001	0
	0.4	0.011	0.009	0.016
	0.5	0.032	0.139	0.371
	0.1	0	0	0
	0.2	0.001	0	0
$\frac{1}{4}n^{-1/2}$	0.3	0.003	0	0
	0.4	0.009	0.003	0.014
	0.5	0.034	0.114	0.288

Table 11: Power for Weighted Instrument CvM Test under Design 2

t_n	$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
	0.1	0	0	0
	0.2	0.006	0.016	0.032
$n^{-1/5}$	0.3	0.026	0.138	0.295
	0.4	0.064	0.449	0.831
	0.5	0.175	0.848	0.995
	0.1	0.007	0.012	0.005
	0.2	0.016	0.062	0.1
$n^{-1/3}$	0.3	0.041	0.215	0.456
	0.4	0.119	0.604	0.876
	0.5	0.21	0.902	0.996
	0.1	0.006	0.014	0.01
	0.2	0.023	0.057	0.086
$n^{-1/2}$	0.3	0.038	0.229	0.389
	0.4	0.119	0.532	0.791
	0.5	0.203	0.85	0.982

Table 12: Power for Weighted Instrument KS Test under Design 2 (from Armstrong and Chan (2012))

h_n	$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
	0.1	0	0	0
	0.2	0.001	0.002	0
$n^{-1/5}$	0.3	0.008	0.007	0.024
	0.4	0.012	0.108	0.369
	0.5	0.074	0.484	0.923
	0.1	0	0.001	0
	0.2	0.001	0	0
$n^{-1/3}$	0.3	0.003	0.009	0.011
	0.4	0.023	0.126	0.273
	0.5	0.062	0.519	0.848
	0.1	0	0	0
	0.2	0.001	0	0
$n^{-1/2}$	0.3	0.001	0	0
	0.4	0.005	0.007	0.023
	0.5	0.023	0.089	0.308

Table 13: Power for Kernel CvM Test under Design 2

h_n	$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
	0.1	0.001	0.001	0.001
	0.2	0.009	0.029	0.049
$n^{-1/5}$	0.3	0.044	0.185	0.386
	0.4	0.082	0.524	0.867
	0.5	0.18	0.879	0.997
	0.1	0.007	0.015	0.014
	0.2	0.015	0.067	0.129
$n^{-1/3}$	0.3	0.029	0.18	0.454
	0.4	0.087	0.525	0.856
	0.5	0.167	0.825	0.98
	0.1	0	0.014	0.006
	0.2	0	0.025	0.032
$n^{-1/2}$	0.3	0	0.057	0.123
	0.4	0	0.163	0.286
	0.5	0	0.321	0.604

Table 14: Power for Kernel KS Test under Design 2

$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
0.1	0.005	0	0.001
0.2	0.031	0.046	0.058
0.3	0.131	0.454	0.743
0.4	0.359	0.914	0.997
0.5	0.619	0.999	1

Table 15: Power for Unweighted Instrument CvM Test under Design 3

$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
0.1	0.006	0.015	0.013
0.2	0.027	0.231	0.402
0.3	0.117	0.737	0.959
0.4	0.34	0.982	1
0.5	0.568	1	1

Table 16: Power for Unweighted Instrument KS Test under Design 3

σ_n^2	$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
	0.1	0.006	0	0.001
	0.2	0.037	0.079	0.136
$\frac{1}{4}n^{-1/5}$	0.3	0.133	0.515	0.837
	0.4	0.341	0.941	1
	0.5	0.636	1	1
	0.1	0.006	0.003	0.001
	0.2	0.029	0.065	0.173
$\frac{1}{4}n^{-1/3}$	0.3	0.143	0.514	0.872
	0.4	0.375	0.961	1
	0.5	0.642	1	1
	0.1	0.006	0.003	0
	0.2	0.043	0.059	0.101
$\frac{1}{4}n^{-1/2}$	0.3	0.161	0.52	0.845
	0.4	0.335	0.935	0.999
	0.5	0.63	0.999	1

Table 17: Power for Weighted Instrument CvM Test under Design 3

t_n	$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
	0.1	0.034	0.064	0.12
	0.2	0.093	0.466	0.704
$n^{-1/5}$	0.3	0.272	0.869	0.99
	0.4	0.501	0.994	1
	0.5	0.767	1	1
	0.1	0.039	0.104	0.116
	0.2	0.112	0.429	0.64
$n^{-1/3}$	0.3	0.257	0.838	0.979
	0.4	0.463	0.994	1
	0.5	0.717	1	1
	0.1	0.03	0.083	0.087
	0.2	0.121	0.325	0.523
$n^{-1/2}$	0.3	0.24	0.762	0.967
	0.4	0.397	0.984	1
	0.5	0.669	1	1

Table 18: Power for Weighted Instrument KS Test under Design 3 (from Armstrong and Chan (2012))

h_n	$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
	0.1	0.013	0.017	0.018
	0.2	0.05	0.229	0.446
$n^{-1/5}$	0.3	0.187	0.757	0.965
	0.4	0.411	0.98	1
	0.5	0.698	1	1
	0.1	0.007	0.012	0.01
	0.2	0.044	0.167	0.323
$n^{-1/3}$	0.3	0.173	0.676	0.932
	0.4	0.377	0.986	1
	0.5	0.657	1	1
	0.1	0.002	0.001	0
	0.2	0.029	0.03	0.049
$n^{-1/2}$	0.3	0.082	0.326	0.654
	0.4	0.21	0.866	0.991
	0.5	0.47	0.996	1

Table 19: Power for Kernel CvM Test under Design 3

h_n	$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
	0.1	0.043	0.087	0.161
	0.2	0.099	0.487	0.722
$n^{-1/5}$	0.3	0.261	0.876	0.99
	0.4	0.48	0.995	1
	0.5	0.746	1	1
	0.1	0.037	0.086	0.122
	0.2	0.079	0.297	0.528
$n^{-1/3}$	0.3	0.164	0.646	0.912
	0.4	0.296	0.937	0.999
	0.5	0.507	0.996	1
	0.1	0	0.035	0.026
	0.2	0	0.087	0.118
$n^{-1/2}$	0.3	0	0.195	0.385
	0.4	0	0.427	0.703
	0.5	0	0.716	0.952

Table 20: Power for Kernel KS Test under Design 3