Multiscale Adaptive Inference on Conditional Moment Inequalities

Timothy B. Armstrong[1] Hock Peng Chan[2]

[1]Yale University

[2]National University of Singapore

June 2013

Conditional moment inequality models

- Numerous applications in empirical economics
- Relax assumptions at the cost of getting bounds rather than point estimates
- Also useful in mapping revealed preference inequalities to data
- If the bounds are good, we can have more confidence in our analysis and still reach the same conclusions.

Main results

- Inference on conditional moment inequalities using "multiscale" test statistic
- Derive asymptotic distribution - requires new methods for supremum of nonstationary, nongaussian random process without intermediate gaussian approximation
- Critical values based on asymptotic distribution are easy to compute - no simulation. Bootstrap validity also shown.
- Derive power results showing that the test achieves optimal power against local alternatives in set identified case "adaptively"
- Test is less sensitive to moment selection procedures than many available procedures: achieves optimal power even using "least favorable" critical values

Setup

- Inference on points in $\Theta_0 \equiv \{\theta | E(m(W,\theta)|X) \geq 0 \text{ a.s.}\}$.
- For each θ, test null of $\theta \in \Theta_0$
- $m(W,\theta)$ can be multivariate, in which case the inequality is interpreted as holding for each element
- θ can be multivariate.
- Multiscale statistic:

$$T_{n,j} = T_{n,j}(\theta) \equiv \left| \inf_{s,s+t \in \hat{\mathcal{X}}, t \geq t_n} \frac{1}{n} \sum_{i=1}^{n} \frac{m_j(W_i,\theta) I(s < X_i < s+t)}{\hat{\sigma}_{n,j}(s,t,\theta)} \right|_-$$

where t_n is a sequence going to zero, $\hat{\mathcal{X}}$ is the convex hull of $\{X_i\}_{i=1}^n$, $|x|_- \equiv |\min\{x,0\}|$ and

$$\hat{\sigma}^2_{n,j}(s,t,\theta) \equiv \hat{var}(m_j(W_i,\theta) I(s < X_i < s+t))$$

- Reject if $S_n(\theta) \equiv \max_j T_{n,j}(\theta)$ is greater than some critical value.

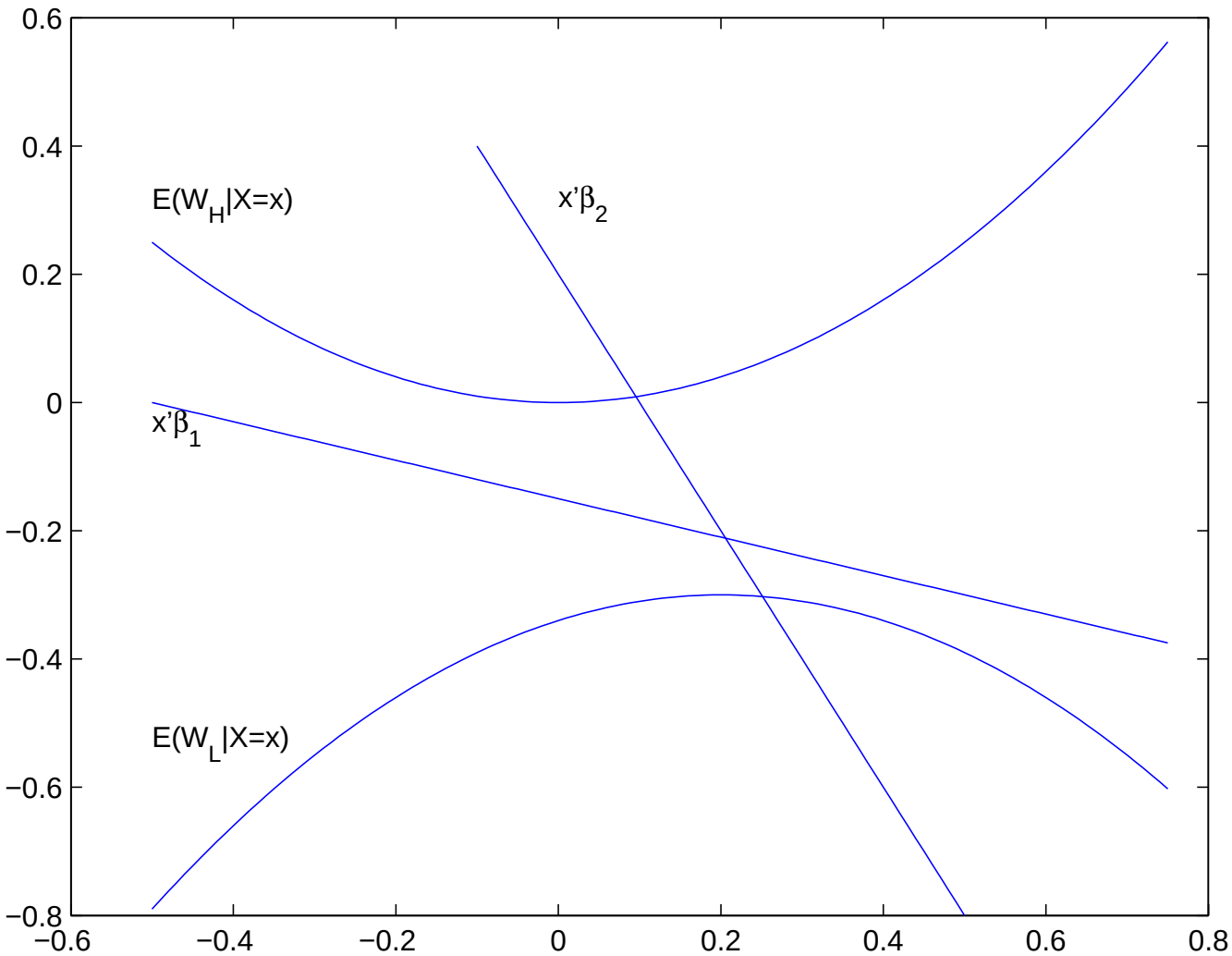
Comparison to nonparametric testing

- To control size, could simply think of this as a nonparametric testing problem. Let $Y = m(W, \theta)$, test $E(Y|X) \geq 0$.
- However, the model tells us how the distribution of $(m(W, \theta), X)$ relates to θ.
- This gives us information about *power* against alternatives *within the model*.
- This research uses additional information from the model to design tests that are more powerful in the generic set identified case.
- Use distance on $E(Y|X)$ that corresponds to distance on θ.
- Other approaches may be better in other situations (e.g. regular point identification, testing stochastic dominance, testing affiliation, etc.).

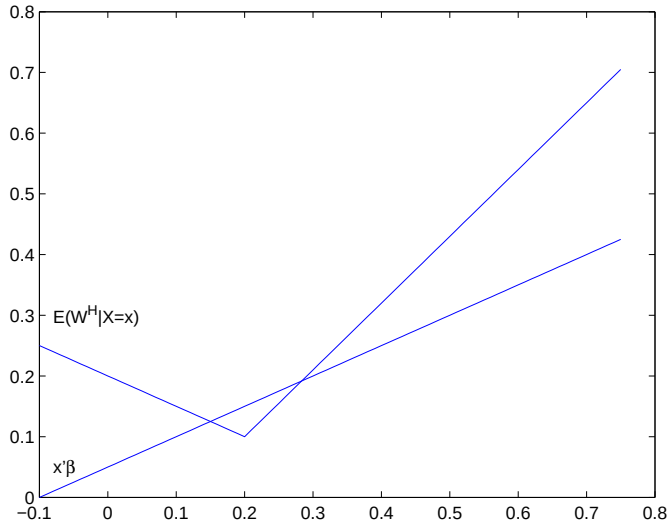
Example: Interval Regression (Manski and Tamer, 2002)

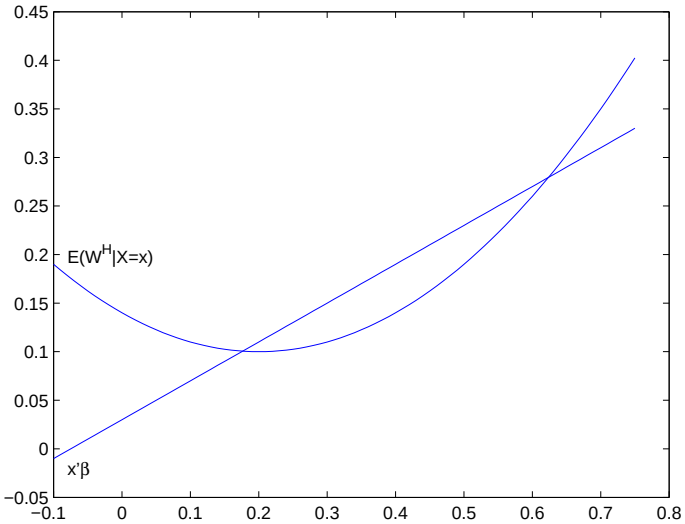
- W^* (health expenditures, say) follows a linear regression on X (income): $E(W^*|X) = X'\theta_0$
- Don't observe W^*, but observe X and interval $[W^L, W^H]$ with $W^* \in [W^L, W^H]$.
- No restrictions on censoring process (can depend on W^* and X)
- Conditional moment inequalities: $E(W^L|X) \leq X'\theta \leq E(W^H|X)$ a.s.

Example: Interval Regression (Manski and Tamer, 2002)



Intuition for optimal power





- Take supremum over all "bandwidths" t to find optimal one without prior knowledge of smoothness of dgp.
- Weight by $\hat{\sigma}(s,t,\theta)$ to optimize tradeoff between variance and drift

Conditions for asymptotic distribution

- Consider *least favorable* dgp in null with $\bar{m}(\theta, x) \equiv E(m(W_i, \theta)|X_i = x) = 0$ all x.
- Test statistic is greatest when all moments bind, so using critical values based on this distribution controls size.
- Can extend to moment selection, but achieve rate optimal power even without moment selection.
- Results hold uniformly over suitable classes of underlying distributions P (shown in paper), but stated for a single P here for simplicity.

Conditions for asymptotic distribution

Assumption 1

a.) The distribution of $m(W_i, \theta)$ conditional on X_i satisfies the following conditions.

i.) There exists a $\Lambda > 0$ such that

$$E(\exp(\lambda |m_j(W_i, \theta)||X_i) < \infty \text{ a.s. all } 1 \leq j \leq d_Y, 0 \leq \lambda \leq \Lambda$$

ii.) $var(m_j(W_i, \theta)|X_i = x)$ is positive and continuous in x for all j.

iii.) $E[(m_j(W_i, \theta) - \bar{m}_j(\theta, x))^3|X_i]$ is finite and bounded for all j.

iv.) $corr(m_j(W_i, \theta), m_k(W_i, \theta)|X_i)$ is bounded away from 1 for all $j \neq k$.

b.) The support $\mathcal{X}$ of X_i is compact and convex and X_i has a density f that is bounded away from zero on $\mathcal{X}$.

c.) $t_n \to 0$ and $nt_n^{d_X} / |\log t_n|^4 \to \infty$.

Notes on regularity conditions

- Exponential moment could potentially be relaxed with self-normalized moderate deviations results
- t_n allowed to go to zero at fastest possible rate (up to a $\log n$ term)
  - Needed for good power in higher dimensions (i.e. $d_X > 1$) or with less "smoothness"
  - Conditions on t_n for intermediate gaussian approximations too stringent to obtain good power in more than one dimension in realistic settings.
- Convexity of $\mathcal{X}$ could be relaxed by truncating $\hat{\sigma}$ rather than t

Theorem 1

Suppose that the null hypothesis and Assumption 1 hold for θ and the data are iid. Let $\hat{c}_n = vol(\hat{\mathcal{X}})/t_n^{d_X}$ and let $a(\hat{c}_n) = (2n \log \hat{c}_n)^{1/2}$ and $b(\hat{c}_n) = 2 \log \hat{c}_n + (2d_X - 1/2) \log \log \hat{c}_n - \log(2\sqrt{\pi})$. Then, for any vector $r \in \mathbb{R}^{d_Y}$,

$$\liminf_n P\left(a(\hat{c}_n) T_n - b(\hat{c}_n) \leq r\right) \geq P(Z \leq r)$$

where Z is a d_Y dimensional vector of independent standard type I extreme value random variables. If, in addition $\bar{m}_j(\theta, x) = 0$ for all x and j, then

$$a(\hat{c}_n) T_n - b(\hat{c}_n) \xrightarrow{d} Z.$$

Comments

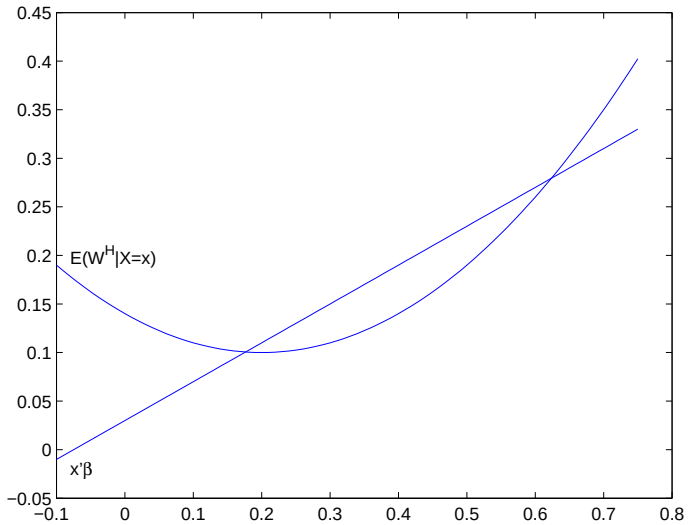
- Distribution is extreme value despite dependence on large bandwidths t.
- Depends only on $vol(\mathcal{X})$ despite complicated covariance function of approximating process - easy to compute feasible critical values.
- Good in finite samples when t_n is large enough for normal approximation to work for fixed (s, t), small enough for tail approximations to work well.
- If approximation is poor for "optimal" t_n, can use smaller t_n without losing much power (still achieve optimal rate).
- Can bootstrap when approximation does not work well.

▸ sketch of proof ▸ moment selection

Power

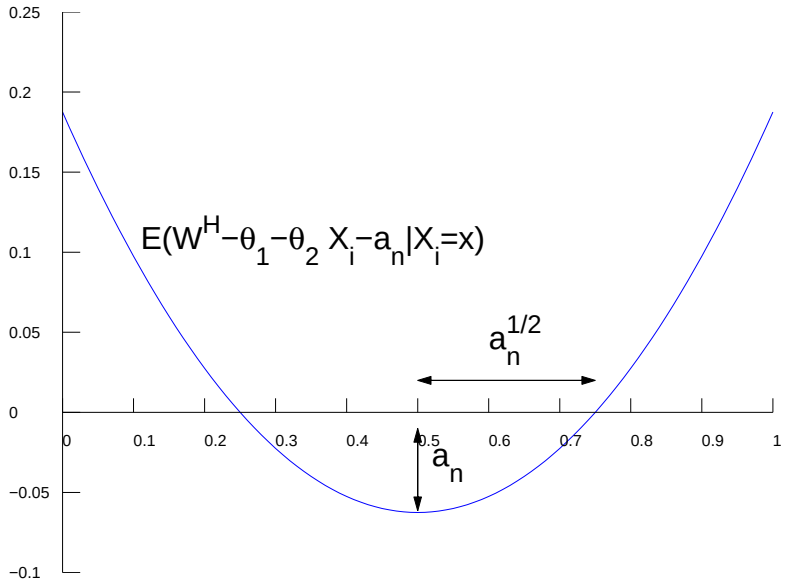
- We consider power against sequences of alternatives θ_n that approach the boundary $\delta\Theta_0$ of the identified set Θ_0.
- Let $\theta_n = \theta_0 + ar_n$ for θ_0 on the boundary of Θ_0 and a vector a and a sequence of scalars $r_n \to 0$ such that $\theta_n \notin \Theta_0$ (at least for large enough n).
- Derive fastest sequence r_n for which the test has nontrivial power.
- Adaptivity: test achieves optimal rate for r_n for a range of smoothness conditions for dgp without prior knowledge of smoothness conditions
- Conditions are "generic" under set identification. Test is close to optimal (up to $\log n$ term) in other settings.

Power



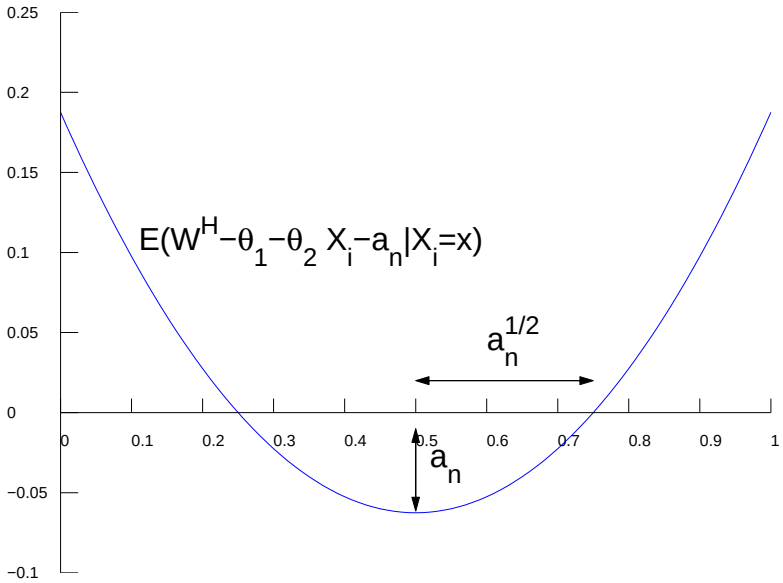
- To understand the results in the general case, consider the interval regression model.
- Consider alternative with a_n added to the intercept term. Need $E[(W_i^H - \theta_1 - \theta_2 X_i - a_n) I(s < X_i < s + t)]/\sigma(s,t)$ more negative than $(n/\log n)^{-1/2}$ critical value to get power.

Power (one dimensional case with 2 deriviatives)



- With 2 derivatives, $E(m(W_i, \theta)|X_i = x)$ is bounded from above by a constant times $(x - x_0)^2 - a_n$ for some x_0
- For $(s, s + t)$ near x_0 with $t \approx a_n^{1/2}$, $E(m(W_i, \theta_n)I(s < X_i < s + t)) \approx -a_n^{3/2}$ and $\sigma(s, t) \approx t^{1/2} = a_n^{1/4}$

Power (one dimensional case with 2 deriviatives)



- So $E(m(W_i, \theta_n) I(s < X_i < s + t))/\sigma(s, t) \approx -a_n^{3/2}/a_n^{1/4} = -a_n^{3/2-1/4} = -a_n^{5/4}$
- Needs to be larger in magnitude than $(\log n)^{1/2}/n^{1/2}$, so need $a_n \approx [(\log n)/n]^{-2/5}$

Conditions for local power

Assumption 2

a.) $\bar{m}(\theta, x) \equiv E(m(W_i, \theta)|X_i = x)$ is differentiable in θ with derivative $\bar{m}_\theta(\theta, x)$ that is continuous as a function of θ uniformly in (θ, x).

b.) For some γ, C, j and $x_0 \in \mathcal{X}$, we have $\bar{m}_j(\theta_0, x_0) = 0$ and, for all x in a neighborhood of x_0,

$$|\bar{m}_j(\theta_0, x) - \bar{m}_j(\theta_0, x_0)| \leq C\|x - x_0\|^\gamma.$$

- γ describes smoothness of the dgp - typically, $\gamma \leq 2$

Theorem 2

Suppose that Assumptions 1 and 2 hold for θ_0. Let $\theta_n = \theta_0 + ar_n$ for some $a \in \mathbb{R}^{d_\theta}$ and a sequence of scalars $r_n \to 0$. Suppose that, for some index j such that part (b) of Assumption 2 holds for j,

$$\liminf r_n \left(\frac{n}{2 \log t_n^{-d_X}} \right)^{\gamma/(d_X+2\gamma)} > c(a)$$

Where $c(a)$ is a constant that depends on a, $\frac{d}{d\theta}\bar{m}(\theta, x_0)\big|_{\theta=\theta_0}$, C, γ and the pdf and conditional variance at x_0. Then, if $t_n < \eta(n/\log n)^{-1/(d_X+2\gamma)}$ for small enough η, we will have

$$P(S_n(\theta_n) > \hat{q}_{1-\alpha}) \to 1.$$

Comments

- $(n/\log n)^{-\gamma/(d_X+2\gamma)}$ is the best possible rate even if γ is known (see Stone, 1982)
- Need $t_n < (n/\log n)^{-1/(d_X+2\gamma)}$ - can obtain this simultaneously for all γ by choosing, say, $t_n = [(\log n)^5/n]^{1/d_X}$.
- In practice, can choose t_n based on smallest (least smooth) γ that seems plausible
- Effect of t_n on critical value: if $t_n = n^{-\delta}$ for some δ, the critical value is proportional to $[2\delta d_X(\log n)/n]^{1/2}$

▸ comparison to other tests

Monte Carlo

- Median regression with potentially endogenously missing data.
- Conditional median given by $q_{1/2}(W_i^*|X_i) = \theta_1 + \theta_2 X_i$, W_i^* not always observed, missingness may be correlated with W_i^* itself.
- Conditional moment inequality: $E(I(\theta_1 + \theta_2 X_i \leq W_i^H) - 1/2|X_i) \geq 0$ where $W_i^H = W_i^*$ when W_i^* is observed, ∞ otherwise (similar lower bound, but not used in monte carlos for simplicity).
- Examine finite sample size under least favorable null, power under alternatives corresponding to different "missingness" processes.

	t_n	$n = 100$	$n = 500$	$n = 1000$
nominal size .1	$n^{-1/5}$	0.2510	0.1840	0.1770
	$n^{-1/3}$	0.1640	0.1160	0.1150
	$n^{-1/2}$	0.0890	0.0770	0.0880
nominal size .05	$n^{-1/5}$	0.1020	0.0650	0.0790
	$n^{-1/3}$	0.0750	0.0410	0.0550
	$n^{-1/2}$	0.0340	0.0220	0.0350

Table : False Rejection Probabilities for Least Favorable Null

Power

- Fix slope parameter θ_2 at 0, let $\overline{\theta}_1$, be largest value of intercept parameter θ_1 in identified set with θ_2 fixed.
- Consider alternatives with $\theta_1 > \overline{\theta}_1$
- Vary "missingness" process to get different shapes of conditional moment inequalities.
  - Design 1: flat conditional mean ($\sqrt{n/\log n}$ convergence)
  - Design 2: continuous, kinked at minimum ($\gamma = 1$)
  - Design 3: two derivatives with interior minimum ($\gamma = 2$)
- Use finite sample critical values from least favorable distribution.

t_n	$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
$n^{-1/5}$	.1	0.2070	0.5030	0.7290
	.2	0.4800	0.9540	1.0000
	.3	0.7590	1.0000	1.0000
	.4	0.9560	1.0000	1.0000
	.5	0.9970	1.0000	1.0000
$n^{-1/3}$	.1	0.1440	0.4530	0.6300
	.2	0.3780	0.9390	0.9980
	.3	0.6910	1.0000	1.0000
	.4	0.8860	1.0000	1.0000
	.5	0.9820	1.0000	1.0000
$n^{-1/2}$	.1	0.1560	0.3580	0.5020
	.2	0.3480	0.8980	0.9910
	.3	0.6490	0.9990	1.0000
	.4	0.8620	1.0000	1.0000
	.5	0.9740	1.0000	1.0000

Table : Power for Level $\alpha = .05$ Test with Critical Values Based on Finite Sample Least Favorable Distribution (Design 1: flat conditional mean)

t_n	$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
$n^{-1/5}$	.1	0	0	0
	.2	0.0060	0.0160	0.0320
	.3	0.0260	0.1380	0.2950
	.4	0.0640	0.4490	0.8310
	.5	0.1750	0.8480	0.9950
$n^{-1/3}$	.1	0.0070	0.0120	0.0050
	.2	0.0160	0.0620	0.1000
	.3	0.0410	0.2150	0.4560
	.4	0.1190	0.6040	0.8760
	.5	0.2100	0.9020	0.9960
$n^{-1/2}$	.1	0.0060	0.0140	0.0100
	.2	0.0230	0.0570	0.0860
	.3	0.0380	0.2290	0.3890
	.4	0.1190	0.5320	0.7910
	.5	0.2030	0.8500	0.9820

Table : Power for Level $\alpha = .05$ Test with Critical Values Based on Finite Sample Least Favorable Distribution (Design 2: $\gamma = 1$)

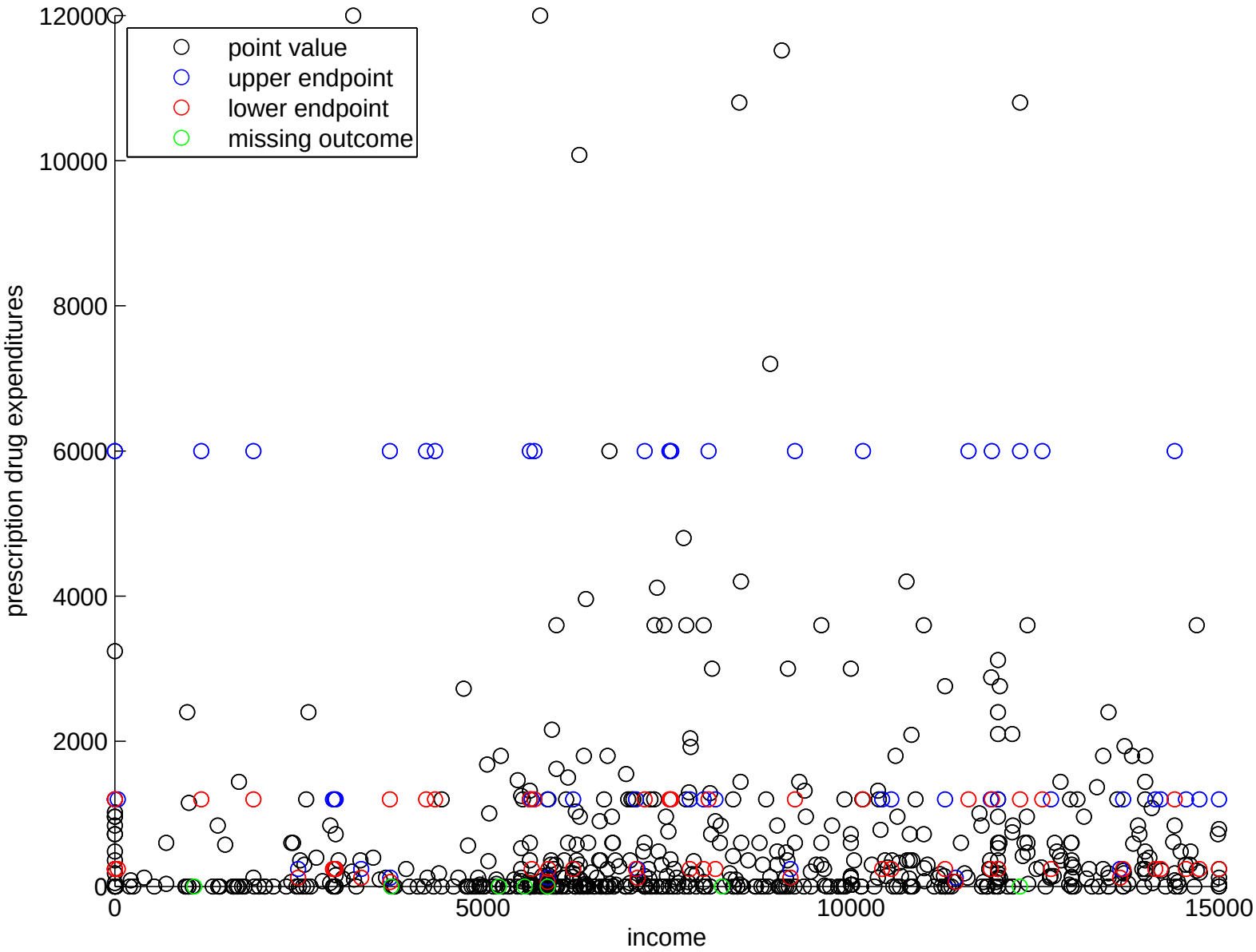
t_n	$\theta_1 - \overline{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
$n^{-1/5}$	.1	0.0340	0.0640	0.1200
	.2	0.0930	0.4660	0.7040
	.3	0.2720	0.8690	0.9900
	.4	0.5010	0.9940	1.0000
	.5	0.7670	1.0000	1.0000
$n^{-1/3}$	.1	0.0390	0.1040	0.1160
	.2	0.1120	0.4290	0.6400
	.3	0.2570	0.8380	0.9790
	.4	0.4630	0.9940	1.0000
	.5	0.7170	1.0000	1.0000
$n^{-1/2}$	.1	0.0300	0.0830	0.0870
	.2	0.1210	0.3250	0.5230
	.3	0.2400	0.7620	0.9670
	.4	0.3970	0.9840	1.0000
	.5	0.6690	1.0000	1.0000

Table : Power for Level $\alpha = .05$ Test with Critical Values Based on Finite Sample Least Favorable Distribution (Design 3: $\gamma = 2$)

Empirical Illustration: Income and Out of Pocket Prescription Drug Spending in the Health and Retirement Study (HRS)

- HRS presents series of brackets to those who do not report a point value.
- Results in interval measurements.
- $n = 636$ observations
- Compute confidence regions for an interval median regression model of prescription drug spending on income.

▸ details


point value
upper endpoint
lower endpoint
missing outcome
prescription drug expenditures
income
0
2000
4000
6000
8000
10000
12000
0
5000
10000
15000

Results

	θ_1	θ_2
multiscale test (this paper)	$[-30, 109]$	$[0.0053, 0.0320]$
bounded weights	$[-60, 138]$	$[0.0030, 0.0372]$
LAD with Points	$[-63, 63]$	$[0.0100, 0.0244]$

Table : 95% Confidence Intervals for Components of θ

Conclusion

- Inference on conditional moment inequalities using multiscale test statistic
- Derived asymptotic distribution - requires different argument than elsewhere in the literature
- Critical values can be computed analytically
- Validity of (modified) bootstrap follows from same result.
- Test achieves optimal power adaptively even with least favorable critical value

Sketch of proof

- Need to overcome two problems in proof: (1) gaussian approximations not good enough for this statistic and (2) covariance function of process nonstationary in ways not previously dealt with
- Solution: use results of Chan and Lai (2006) to obtain tail approximations for nonstationary, nongaussian processes without using strong approximations
- Uses moderate deviations combined with conditions on the tail - equicontinuity, etc.
- Show that (s, t) with t much larger than t_n do not matter, use independence over (s, t) with t small and s along with tail approximations for suprema over small blocks.
- Similar ideas to proof for gaussian process, but using approximations directly on finite sample objective function

back

Moment selection

- Test uses critical values based on "least favorable" dgp with $E(m(W_i, \theta)|X_i = x) = 0$ for all x, but this may be too pessimistic if the inequality binds only on a subset $\tilde{\mathcal{X}}$.
- Can extend to allow for pre-tests that find a set $\tilde{\mathcal{X}}$ such that the inequalities bind only on this set - replace $vol(\hat{\mathcal{X}})$ with $vol(\tilde{\mathcal{X}})$ in the definition of $\hat{c}_n$.
- Perhaps surprisingly, this leads to no first order power improvement unless $vol(\tilde{\mathcal{X}}) \to 0$
- Why? Note that, regardless of $\mathcal{X}$, the critical value $\hat{q}_{1-\alpha}$ satisfies

$$\hat{q}_{1-\alpha} \sim b(\hat{c}_n)/a(\hat{c}_n) \sim (2 \log \hat{c}_n)^{1/2}/n^{1/2}$$
$$\sim [2 \log t_n^{-d_X} + \log vol(\mathcal{X})]^{1/2}/n^{1/2} \sim (2 \log t_n^{-d_X})^{1/2}/n^{1/2}$$

back

Comparison to other approaches

- The "multiscale statistic" can be thought of as using (1) an instrument based approach with (2) a KS (supremum) statistic, (3) a variance weighting (4) a "least favorable" critical value (or moment selection)
- For (1), can use kernels or sieves to estimate $E(m(W_i, \theta)|X)$ instead of transforming to $E(m(W_i, \theta)I(s < X_i < s + t))$.
- For (2), can use a CvM statistic that integrates the negative part of the sample moments.
- For (3), can use a bounded weight
- For (4), can use moment selection or asymptotic distribution on the boundary of Θ_0

Our approach is best in "generic set identified case" with no prior knowledge of smoothness, close to best in other cases (see Armstrong (2011a,b, 2012) for local power of these procedures under set identification).

Comparison to bounded weighting (see Andrews and Shi, 2012, Kim, 2009, Armstrong, 2011)

- Variance weighting improves power and achieves optimal rate
  - Can think of this as analogous to optimal weighting in GMM
- Weighting allows statistic to perform bias-variance tradeoff optimally.
- Asymptotic distribution and derivation are different
  - Quantiles of asymptotic distribution can be computed analytically in our case.
- Critical values are less sensitive to moment selection.
- Power improvement applies to conditional moment inequality models in set identified case. Our statistic is less powerful by a $\log n$ term in certain other cases, but could be modified to do just as well.

Comparison to kernel statistic (see Chernozhukov, Lee and Rosen, 2012, Ponomareva, 2011, Armstrong 2012)

- Supremum is taken over all bandwidths as well as over x
- Allows for adaptivity - don't need to know optimal bandwidth
- No first order power loss relative to using kernels with t_n as bandwidth
- Our approach doesn't do too much worse than Andrews-Shi-Kim in cases where it is better (e.g. regular point identification), while kernels can do much worse.

Comparison to CvM (integration based) statistics (see Lee, Song and Whang, 2012, Andrews and Shi, 2012, Kim, 2009, Armstrong, 2012)

- KS statistics do better in generic set identified case.
- Why? Critical value is less sensitive to which inequalities are close to holding with equality.
- Supremum of k normal variables increases like $(\log k)^{1/2}$, while sum of positive part increases like k. Intuition carries over to a continuum of moments.
- Applied work typically uses a finite number of moments and a statistic that sums the negative parts (CvM). These results suggest that many of these papers should be using a supremum statistic.

Comparison to other approaches (in generic set identified case)

statistic	weighting	critical value	local power	robust
inst-KS	variance	least fav	$\left(\frac{n}{\log n}\right)^{-\frac{\gamma}{d_X+2\gamma}}$	yes
inst-KS	bounded	mom. selec.	$n^{-\frac{\gamma}{2(d_X+\gamma)}}$	yes
inst-KS	bounded	dist. on bndry	$n^{-\frac{\gamma}{d_X+2\gamma}}$	no
kern-KS	-	mom. selec.	$\left(\frac{nh^{d_X}}{\log n}\right)^{-1/2} \vee h^{\gamma}$	yes
inst-CvM	bounded	mom. selec.	$n^{-\frac{\gamma}{2[d_X+\gamma+(d_X+1)/p]}}$	yes
inst-CvM	variance	mom. selec.	$n^{-\frac{\gamma}{2[d_X/2+\gamma+(d_X+1)/p]}}$	yes
kern-CvM	-	mom. selec.	$h^{\gamma} \vee \left(nh^{d_X}\right)^{-\frac{1}{[2(1+d_X/(p\gamma))]}}$	yes

back

Empirical Illustration: Income and Out of Pocket Prescription Drug Spending in the Health and Retirement Study (HRS)

- HRS presents series of brackets to those who do not report a point value.
- Results in interval measurements.
- Both variables are measured as intervals for some observations, but I use those with point values for income and focus on interval reporting for spending (valid under interval-reporting-at-random assumption for income, possibly endogenous interval reporting for spending).
- Restrict sample to women with \$15,000 of income or less who report using prescription medications ($n = 636$).
- Compute confidence regions for an interval median regression model of prescription drug spending on income.

Setup

- Out of pocket prescription spending W_i^* follows linear median regression model given income X_i: $q_{1/2}(W_i^*|X_i) = \theta_1 + \theta_2 X_i$.
- Observe (X_i, W_i^L, W_i^H) where $W_i^* \in [W_i^L, W_i^H]$.
- Leads to conditional moment inequalities
  $E(I(\theta_1 + \theta_2 X_i \leq W_i^H) - 1/2|X_i) \geq 0$ and
  $E(1/2 - I(\theta_1 + \theta_2 X_i \leq W_i^L)|X_i) \geq 0$.
- Similar intuition for faster rates of convergence to interval mean regression.
- Also report results for least absolute deviations regression (LAD) restricted to point values (requires additional interval reporting at random assumption for spending)

back