# Griliches Lectures, Kyoto, August 2016.

# Lecture 2: Econometric Tools for Analyzing Moment Inequalities. *

by Ariel Pakes

*This is a preliminary version of these notes, and no doubt contains many errors and omissions.

## Econometrics for Moment Inequalities.

Our model delivers the condition that

$$\mathcal{E}\left[\Delta r(d_i, d', \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0) \otimes h(x_i)\right] \geq 0.$$

**Estimator.** Form sample analog and looks for values of $\theta$ that satisfy these moment inequalities (can be a set).

**Formalities.** $j = 1, \ldots, J$ markets with observations on $(z, x, d)$ for individual agents. Markets' observations are independent draws from a population with a distribution, say $\mathcal{P}$, that respects our assumptions.

**Sample Moments.**

$$m(z^j, d^j, x^j, \theta) =$$

$$\frac{1}{n_j} \sum_i \Delta r^j(d_i^j, d', d_{-i}^j, z_i^{o,j}, \theta) \otimes h(x_i^j),$$

$$m(\mathbf{P}_J, \theta) = \frac{1}{J} \sum_{j=1}^{J} m(z^j, d^j, x^j, \theta),$$

$$\Sigma(\mathbf{P}_J, \theta) = Var(m(z^j, d^j, x^j, \theta)).$$

**Population Moments.** $(m(\mathcal{P}, \theta), \Sigma(\mathcal{P}, \theta))$ with

$$m(\mathcal{P}, \theta_0) \geq 0.$$

Let

$$\Theta_0 = \{\theta \ : \ m(\mathcal{P}, \theta) \ \geq 0 \ \},$$

which is called the identified set.

**Estimator.** For now I am going to discuss estimation where we do not adjust for the differential variances of the moments. I will come back to an adjustment for differential variance below.

Two different metrics on the negative part of the distance between

$$m(\mathbf{P}_J, \theta) \equiv [m_1(\mathbf{P}_J, \theta), \ldots, m_K(\mathbf{P}_J, \theta)]'$$

and zero are commonly used in the literature. If $f(\cdot)_- \equiv min(0, f(\cdot))$ then one is

$$\Theta_J = \arg\min_{\theta} \|m(\mathbf{P}_J, \theta)_-\|,$$

and at least initially I will focus on it, though analogous reasoning applies when we use

$$\Theta_J = \arg\min_{\theta}[\min_{k} m_k(\mathbf{P}_J, \theta), 0].$$

If all the moments are positive this metric is zero, and if one or more is negative we take the most negative.

## Inference.

**Consistency of Set Estimator.** Several papers provide conditions for the consistency of the estimator, usually in Hausdorff metric

$$d_H(sup_{\theta_j \in \Theta_j} inf_{\theta_0 \in \Theta_0} d(\theta_j, \theta_0) + sup_{\theta_0 \in \Theta_0} inf_{\theta_j \in \Theta_j} d(\theta_j, \theta_0))$$

where $d(\cdot, \cdot)$ is taken to be a norm (usually the sup norm) on points in Euclidean space.

**Large Measures of Precision.** There are several different ways of conceptualizing measures of the precisions of your (set) estimator. We could attempt to:

- Get a confidence set for the set; i.e. a set which would cover the identified set 95% of the time (starts with Chernozhukov, Hong, and Tamer, *Econometrica* 2007). I will not go over this, as it has not been used intensively.

- Get a confidence set for the point $\theta_0$ (starts with Imbens and Manski, *Econometrica*, 2004). This is what you see most often, and I will focus on it.

- Get confidence interval for intervals defined for a particular direction in the parameter space; simplest case is directions defined by each component of $\theta = [\theta_1, \ldots, \theta_K]$ as this gives us the analogue of standard

confidence intervals produced by moment equality estimators. I will consider this, as this is what is often needed for applied articles.

There are a number of ways of providing estimates of appropriate size for each concept. I will briefly discuss some of the alternatives.

**Adjust for Different Variance of Different Moments.** Assume that a consistent estimator of the diagonal matrix consisting of the square root of the moments evaluated at each $\theta$ is available. Call that estimate $\hat{D}_J(\theta)$ (a diagonal matrix). Then, estimation proceeds as follows. Set

$$\hat{\Theta}_J = \arg\min_{\theta \in \Theta} \|\hat{D}_J(\theta)^{-1/2} \mathbf{P}_J m(w, \theta)_-\| \qquad (1)$$

*Note* the difference between the weighting being done here and the weighting that is done for m.o.m. with equality constraints. In the equality case we weight with the full covariance matrix. Here we do not do that because the weighting by the Cholesky transform of the covariance matrix might imply multiplying a moment by a negative number, and then the weighted moment inequalities at $\theta = \theta_0$ need not have positive expectation.

## Intuition for why standard limiting arguments do not work.

Look to one parameter that we are particularly interested in. Define

$$\underline{\theta} = argmin_{\theta \in \Theta_0} \theta_1,$$

Note that $\underline{\theta} \in \mathcal{R}^k$. Analogously define

$$\widehat{\underline{\theta}} = argmin_{\theta \in \Theta_J} \theta_1.$$

This, and the analogous procedure for the upper bound, will give me my estimates for the upper and lower bound of the first component of the vector $\theta$ say $\theta_{0,1} \in [\underline{\theta}_1, \overline{\theta}_1]$.

If we could obtain "good" estimates of the limiting distributions of $(\underline{\hat{\theta}}, \overline{\hat{\theta}})$, we could use them to build *conservative* confidence intervals as follows. Use the limiting distributions of the boundary estimators to obtain $\hat{a}$ and $\hat{b}$ such that

$$\Pr(\hat{a} > \underline{\theta}_1) = \alpha/2 \quad and \quad \Pr(\hat{b} < \overline{\theta}_1) = \alpha/2.$$

Then

$$\Pr\{[\underline{\theta}_1, \overline{\theta}_1] \subset [\hat{a}, \hat{b}]\} \geq$$

$$1 - \Pr\{\hat{a} > \underline{\theta}_1\} - \Pr\{\hat{b} < \overline{\theta}_1\} = 1 - \alpha.$$

Two points to come back to

• First, the interval CI is conservative for the point,$\theta_{0,1}$. I.e.since

$$Pr\{\theta_{0,1} \in [\hat{a}, \hat{b}]\} \geq \Pr\{[\underline{\theta}_1, \overline{\theta}_1] \subset [\hat{a}, \hat{b}]\},$$

If the $[\widehat{a}, \widehat{b}]$ satisfy the inequality above $Pr\{\theta_{0,1} \in [\widehat{a}, \widehat{b}]\} \geq 1 - \alpha$.

• Second, we could improve on the interval slightly by finding the joint distribution of the upper and lower bound and then account for the covariance between them.

**Note.** This assumes we know the true limiting distributions for $(\underline{\widehat{\theta}}, \overline{\widehat{\theta}})$. We now consider the problem of determining these distributions.

**Note.** I will need to construct an approximation to the distribution of the objective function at different values of $\theta$. I will use simulation to do this. An alternative

would be to use subsampling, but I will not pursue that further here.

**Limit Distribution.** Intuition: split moments up into those that are

- Binding: $\mathcal{P}m_0(w, \underline{\theta}) = 0$, and
- Non-binding: $\mathcal{P}m_1(w, \underline{\theta}) > 0$.

With probability approaching one $\Theta_J = \{\theta : \mathbf{P}_J m(w, \theta) \geq 0\}$. So stochastic equicontinuity, and $\widehat{\underline{\theta}} - \underline{\theta} = O_p(1/\sqrt{J})$, neither of which require differentiability of the objective function at $\theta = \underline{\theta}$ (for these arguments see, for e.g. Pakes and Pollard, 1989), imply

$$\sqrt{J}\mathbf{P}_J m(w,\underline{\widehat{\theta}}) = \sqrt{J}\Big(\mathcal{P}m(w,\underline{\widehat{\theta}}) - \mathcal{P}m(w,\underline{\theta})\Big)$$

$$+\sqrt{J}\Big(\mathbf{P}_J m(w,\underline{\theta}) - \mathcal{P}m(w,\underline{\theta})\Big) \; + \sqrt{J}\mathcal{P}m(w,\underline{\theta}) + o_p(1) \geq 0.$$

where

$$o_p(1) \equiv \sqrt{J}\Big(\mathbf{P}_J m(w,\underline{\widehat{\theta}}) - \mathcal{P}m(w,\underline{\widehat{\theta}})\Big) - \sqrt{J}\Big(\mathbf{P}_J m(w,\underline{\theta}) - \mathcal{P}m(w,\underline{\theta})\Big).$$

Now $\sqrt{J}\mathcal{P}m_1(w,\underline{\theta}) \to \infty$ and hence, when $J$ is large enough, will never bind and can be ignored when solving for $\underline{\theta}$. If we linearize the binding moments (the first term) and note that $\mathcal{P}m_0(w,\underline{\theta}) = 0$ (in the second and third terms), we get

$$\Gamma_0\sqrt{J}(\underline{\hat{\theta}} - \underline{\theta}) + \sqrt{J}\mathbf{P}_J m_0(w, \underline{\theta}) + o_p(1) \geq 0.$$

where $\Gamma_0 \equiv \frac{\partial \mathcal{P} m_0(w,\theta)}{\partial \theta}|_{\theta=\underline{\theta}}$, which we assume has full column rank.

**Theorem**.

$$\sqrt{J}(\underline{\hat{\theta}} - \underline{\theta}) \rightarrow_d \hat{\tau}$$

where

$$\hat{\tau} = arg \min_{\left[0 \leq \Gamma_0 \tau + Z\right]} \tau_1$$

and

$$Z \sim N(0, \Sigma_0) \quad \spadesuit.$$

**Consider two cases.**

- $dim(m_0) = dim(\theta)$. One might think this is the leading case. It produces a normal limit distribution.

- $dim(m_0) > dim(\theta)$. This case leads to a non-normal distribution as there is no derivative of the limit function (in any given direction we will have a limit normal, but depending on the realization of the sampling error we will move away from $\underline{\theta}$ in different directions with different derivatives).

*Though the first assumption seems to be generically the "right" assumption for models, the second most often produces a more accurate picture of the true small*

*sample distribution for the size of samples we use.* This is because our samples typically have enough variance so that different realizations of the sample moments will generate different binding moments, so we need an "asymptotic" approximation that mimics that behavior. More formally the first case may be the limit case, but the asymptotic distribution has a "limiting discontinuity" .

**Estimate Limit Distribution.** When this literature discusses building CI's which are uniform over possible DGP's it means that it can cover the case where the parameters are such that the second case is relevant. When the second case is relevant the limit function (i.e. the population moments) are not differentiable at

$\theta = \underline{\theta}$. The estimator will still be $\sqrt{N}$ consistent, but the form of the limit distribution is not normal. However, note that if we *knew* which moments were binding we could obtain a parametric bootstrap by substituting consistent estimates of $(\Gamma_0, \Sigma_0)$ into the formula in the theorem and solving the linear program for different draws of the $Z$. The problem occurs because we do not know which moments to focus on.

So we need a "new" way of finding a confidence set for a multidimensional $\theta_0$ that covers the true parameter with probability $1-\alpha$. Moreover, because the expectation of the objective function is non-differentiable at $\theta_0$, there is no longer a reason to think that any estimate of a function of $\Theta_0$, such as $\underline{\theta}_1$, distributes normally. So we are going to have to simulate test statistics.

Formally we want to test $H_0 : \theta \in \Theta_I(P)$ where $\Theta_I(P)$ is the identified set. We look for a confidence set with the property that

$$\lim_{J\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta\in\Theta_I(P)} Pr\{\theta \in CS\} = 1-\alpha.$$

where the "inf" is over all data generating processes, including ones which generate a $\Theta_I$ where many more moments bind than there are parameters being estimated.

## Least favorable Confidence Sets for the Point, $\theta_0$.

**Intuition.** What we do is assume that all the moments of the model are exactly zero at each $\theta$, and then simulate a distribution for the objective function many times given that fact. We then find the $1-\alpha$ quantile of the simulated distribution of the objective function. Then go back to the data and evaluate the sample moments at that $\theta$. If the sample moment evaluation is greater than that of the $1-\alpha$ quantile, then the value of $\theta$ would be rejected even if all the moments were exactly zero. They must therefore be rejected when the true moments are less than zero.

Let $m_J(\theta) = m(P_J, \theta)$ and to conform to the literature assume that the null is $H_0 : m(\theta_0) \leq 0$. Define

$$R(m_J(\theta), \Sigma_J(\theta)) = max_k(\frac{m_k(\theta)}{\sqrt{\Sigma_{J(k,k)}(\theta)}})$$

We could also do the analogous procedure using, as the objective function, $\|m(P_J, \theta)_+\|$.

For each $\theta$ we look for a $C_\alpha(m(\theta), \Sigma_J(\theta))$ such that

$$Pr\{R(m_J(\theta), \Sigma_J(\theta)) \geq C_\alpha(m(\theta), \Sigma_J(\theta))\} = \alpha.$$

To obtain the least favorable $C_\alpha(m(\theta), \Sigma_J(\theta))$ we simulate from a mean zero variance $\Sigma_J(\theta)$ normal many times and compute $R(0, \Sigma_J(\theta))$ for each simulation run.

The $\alpha$ quantile of that statistic over the simulated samples is $C_\alpha(0, \Sigma_J(\theta))$. We then go back to the data and find out if

$$R(m_J(\theta), \Sigma_J(\theta)) \leq C_\alpha(0, \Sigma_J(\theta)).$$

The particular $\theta$ is in the confidence set if and only if this condition is satisfied. Clearly if we were to simulate from a normal with any acceptable mean (acceptable meaning all the moments are less than zero), the critical value would be less than this, so this is the least favorable critical value.

The steps for obtaining a CS in this way are as follows.

**Step 1.** In principal we are now searching over every point in $\Theta$. In fact, we are going to have to start with some grid, call it $\Theta_L = \{\theta_l, l = 1, \ldots L\}$.

**Step 2.** For each $\theta_l \in \Theta_L$ construct a normal with mean zero and the correlation matrix of $m(\theta_l)$. Simulate many times and calculate the $(1-\alpha)$ quantile of the distribution of $\{z(\theta_l)_{ns}\}_{ns=1}^{NSIM}$, where $z(\theta_l)_{ns}$ is a simulation draw from the normal. This becomes $C_\alpha(0, \Sigma_J(\theta))$.

You should do this from a single set of i.i.d. independent vectors of normal draws and apply that to the Cholesky factorization (which differs by $\theta$). I.e. we hold the random draws fixed as we look over alternative $\theta$.

**Step 3.** Go back to the data. Compute the value of the objective function at $\theta_l$. Accept all $\theta_l$ for which $R(m_J(\theta_l), \Sigma_J(\theta_l)) \leq C_\alpha(0, \Sigma_J(\theta_l))$ The true $\theta_0$ will be in this set with probability $1-\alpha$. Hence, it is a confidence set with significance level $\alpha$.

**Computational burden.** The simulation is easy enough for a fixed $\theta$. However, we should be doing the test at each point in the entire parameter space. Typically what is done is we divide the parameter space into cells and do the test for each cell. There is a question of how you determine $\Theta_L$. Most would estimate $\hat{\Theta}_I$ (the estimate of the identified set) first, and then use that as a basis for defining $\Theta_L$. You need a $\Theta_L$ that is larger than $\hat{\Theta}_I$; perhaps a set where the points yield values

of the objective function less than some (fairly large) $\epsilon$ (and at least in non-linear models this may be hard to determine).

For a large dimensional $\theta$ this can generate a computational burden which is large enough to limit the applicability of the estimator. This will be particularly computationally difficult if the calculation of the moments for each $\theta$ requires a fixed point calculation. We come back to ways of alleviating the computational burden below.

**Size of the confidence set.** As we add moments here two things happen. If the new moments bind (in some direction) it will help us make the confidence set

smaller. However, if they do not bind they will just increase the CS. I.e. adding a moment that does not bind at a particular value of $\theta$ will (weakly) increase the estimate of $C_\alpha(0, \Sigma_J(\theta))$. This is a bit counterintuitive; adding moments, which should be added information, is likely to give you less precise estimates, even if the moment is well specified.

More generally there is a source of conservativeness in the approximation we are using. Some moments will be well below zero, and hardly likely to bind. Still in the simulation we center them to zero, which will imply that they are as likely to bind as the moments that are near zero. A number of modifications designed for utilizing the information in the sample means to make

the procedure less conservative have been suggested. Examples;

- Use a pre-test which throws out the moments which are far away from binding and then adjust significance levels accordingly (moment selection techniques).

- Center the simulated means at a point which reflects the information in the sample mean and adjusts significance levels (the shifted means techniques)*

*The shifted means technique starts with the "long-version" of Pakes, Porter, Ho and Ishii (2015). See Andrews and Guggenburger (2009) and Andrews and Soares (2010) and to Romano, Shaikh, and Wolf (2014) , for discussions of alternatives.

The early versions of these processes required a "tuning" parameter much like the bandwidth used in non-parameteric estimation. The paper by Romano, Shaikh, and Wolf (2014) does not require a "tuning" parameter, and so I am going to focus on it. This despite the fact that it is among the more computationally intensive techniques. Romano, Shaikh, and Wolf starts with a pre-test and then moves to a "moment shifting" technique.

**Romano, Shaikh, and Wolf and Shifted Moments.**

They do an initial step which finds the least favorable critical value for size $\beta$. They then form the following "shifted" mean

$$\tilde{m}_k(m_k(\theta), \Sigma_J(\theta)) =$$

$$\min\{m_k(\theta) + \Sigma^{-1/2}_{J(k,k)} C_\beta(0, \Sigma_J(\theta)), 0\}.$$

They then simulate from a normal with mean $\tilde{m}(m_J(\theta), \Sigma_J(\theta))$ and variance $\Sigma_J(\theta)$, and use that simulation to form the critical value

$$C_{\alpha-\beta}\Big(\tilde{m}(m_J(\theta), \Sigma_J(\theta)), \Sigma_J(\theta)\Big).$$

A $\theta$ is put in the CS if and only if

$$R(m_J(\theta_l), \Sigma_J(\theta_l)) \leq C_{\alpha-\beta}\Big(\tilde{m}(m(\theta), \Sigma_J(\theta)), \Sigma_J(\theta)\Big).$$

They prove that if the CS is formed in this way

$$\lim_{J\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta\in\Theta_I(P)} Pr\Big\{R(m_J(\theta),\Sigma_J(\theta))$$

$$\leq C_{\alpha-\beta}\Big(\tilde{m}(m(\theta),\Sigma_J(\theta)),\Sigma_J(\theta)\Big)\Big\} \geq 1-\alpha+\beta.$$

They;
(i) restrict their test to not reject if $\min_k\{m_k(\theta) + \Sigma^{-1/2}_{J(k,k)}C_\beta(0,\Sigma_J(\theta))\} < 0$ and
(ii) suggest using $\beta$ a small fraction of $\alpha$, say $1/10 \times \alpha$ and adjusting $\alpha$ to insure the desired size.

Note this procedure is less sensitive to the inclusion of irrelevant moments, and hence is an improvement in that sense. However, adding non-binding moments still (weakly) increase the CS, and the computational demands are worse than the least favorable case.

## Conditional and Unconditional Variance-Covariance Matrices, $\Sigma_J(\theta)$.

The following is from Andrews and Pakes (2016), see also Chetverikov (2013). These papers note that if the moment inequalities generated from the model are conditional moment inequalities, conditional say on $X$, then the variance covariance matrix to be used for $\Sigma(\theta)$ can also be the average of the conditional variances. Moreover since

$$Var(m(\cdot,\theta)) = E[Var(m(\cdot,\theta)|x)] + Var(E[m(\cdot,\theta)|x]).$$

we have for the sample

$$Var(\frac{1}{\sqrt{J}}\sum_j m_{k,j}(w_j,\theta) \mid x_1,\ldots,x_J)$$

$$\frac{1}{J}\sum_{j=1}^{J}\left\{m_{k,j}(w_j,\theta) - E[m_{k,j}(w_j,\theta)|x_j]\right\}^2 +$$

$$\frac{1}{J}\sum_{j=1}^{J}\left\{E[m_{k,j}(w_j,\theta)|x_j] - Em_{k,j}(w_j,\theta)\right\}^2 \equiv$$

$$\frac{1}{J}\sum_{j=1}^{J} Var(m_{k,j}(w_j,\theta)|x_j) + Var(E[m_{k,j}(w_j,\theta)|x_j]).$$

So if we set

$$\Sigma_J(\theta) = \frac{1}{J}\sum_{j=1}^{J} Var(m_{k,j}(w_j, \theta)|x_j)$$

we use a smaller (in the matrix sense) variance covariance matrix.

Three points on this are worth noting.

- First, to do this we have to obtain estimates of

$$Var(m_{k,j}(w_j, \theta)|x_j).$$

The suggestion here is to use an estimator suggested in Chetverikov (2013) and Abadie et al (2014). This requires compact support and conditional expectations, i.e. $E[m(w, \theta)|X]$, that are sufficiently smooth in $X$ (Lipshitz in $X$). Let

$$l(X_j) = arg \min_{s \neq j} [(X_s - X_j)' \widehat{Var}(X)^{-1} (X_s - X_j)$$

where

$$\widehat{Var}(X) \equiv J^{-1} \sum_j [X_j - \overline{X}_J][X_j - \overline{X}_J]'$$

and $\overline{X}_J$ is the sample mean vector, and then set

$$\widehat{\Sigma}_J \equiv \frac{1}{2J} \sum_{j=1}^{J} (Y_j - Y_{l(X_j)})(Y_j - Y_{l(X_j)})'.$$

Here the "2" takes account of the variance in both observations. Note that we do not have to do this for each $\theta$ but rather just once. So though there is an added computational step, it is not too onerous. On the other hand, there is a question of just how accurate the resulting estimate of $\Sigma_J(\theta)$ will be in small samples; though we do know that $\widehat{\Sigma}_J$ converges to the average of the conditional covariance matrices.

• Second, it is interesting to compare this to the moment equality case. The variances used for the equality

case do not depend on whether or not you condition on $X$. The reason it does here, is because in the moment equality case all the conditional moments are (or at least are supposed to be, and are treated as) mean zero. So the variance in the conditional moments is zero. Here the conditional moments are not mean zero, and taking out there mean reduces variance.

• Third our intuition that says we will do better using the conditional variance than the total variance (better in the sense of a smaller CS). However this is not necessarily true. The reason is that the conditional covariance (or rather correlation) matrix has different off-diagonals then the unconditional covariance, and those off diagonals could make things worse (especially if they are much more severely positively correlated).

## A Note on Testing.

Testing is likely to be quite important in the moment inequality context when there are many inequalities and the sample is "small". If the model is correct and we had unlimited data all of the inequalities would converge to their limit values (uniformly in $\theta$), and we would find values of $\theta$ which makes all the sample moments non-negative. However, in finite samples the distribution of each moment will be approximately normal. If there are enough moments then, even if in the limit they would all be positive, in finite samples we are likely find one which violates an inequality (actually as we increase the number of moments this will happen with arbitrarily large probability).

When this occurs we will want to find out whether the violation can be attributed to sampling error; if not the model is mis-specified. The issue is easiest to see when estimating an interval.

- If there are many moments which estimate the lower bound, the estimation algorithm will pick out the greatest lower bound.

- Since the expectation of a max is greater than the max of an expectation, use of the glb will generate a positively biased estimate of the upper bound.

- Analogously when we take the least upper bound for the estimator of the upper bound for the interval

we will be obtaining a negatively biased estimate of that bound.

- If these two biases cause the estimated bounds to cross each other, there will not be a value of the parameter which satisfies all the constraints.

One can derive tests in a number of ways, and there is an active literature about this. A few comments are in order.

- If the identified set is non-empty (there is some value of $\theta$ for which the sample moments are all positive) then you will never reject any test.

- If one has estimated a confidence set for a point, then you have already computed a test statistic. I.e. if there is no $\theta$ which is less than the simulated $C_\alpha(\theta)$ level.

Ask more formally if this test has the right *size*?
I.e. what is the probability of rejecting under the null?

Under $H_0$, there is some $\theta_0$ such that $Pm(z, \theta_0) \geq 0$. For this $\theta_0$, our critical values are constructed such that $\theta_0$ is "covered" by the confidence set with probability at least $1-\alpha$. In other words, $\Pr(Q_n(\theta_0) \leq c(\alpha, \theta_0) \mid H_0) \geq 1-\alpha$, or $\Pr(Q_n(\theta_0) \geq c(\alpha, \theta_0) \mid H_0) \leq \alpha$.

$$
\begin{aligned}
\Pr(\mathit{Reject} \mid H_0) &= \Pr(Q_n(\theta) \geq c(\alpha, \theta)\ \forall\theta \mid H_0) \\
&\leq \Pr(Q_n(\theta_0) \geq c(\alpha, \theta_0) \mid H_0) \\
&\leq \alpha. \quad \spadesuit
\end{aligned}
$$

## Inference on Functions of Parameters.

We typically want to find a CS for $\beta = f(\theta)$. The most frequent case is $\beta$ is a component of $\theta$ for then we can univariate CI's for all parameters.

**Projection Method.** If we already have a confidence set for $\theta$, then to find out if a particular $\beta$ is acceptable all we need do is find out if $\exists \theta \in CS(\theta),\ s.t.\ \beta = f(\theta)$. If all we need is the $CS(\beta)$, then sometimes it will be computationally easier to look for this directly (see below) i.e.

$$CS(\beta) = \Big\{ \min_{\theta \in \Theta : f(\theta) = \beta} R(\cdot) - C_{\alpha,\cdot}(\cdot, \theta) \leq 0 \Big\}.$$

## Critical Values That Condition on Binding Moments; Andrews and Pakes (2016)

This is also a tuning parameter free method. It is motivated by two facts

- When there are many moments we are likely to get very large CS's. As in RSW we mitigate this problem but we don't have a comparison of when one procedure works better than the other.

- There are simplifying computational procedures when the model is linear in "nuisance" parameters when

we use the conditional moments that are not currently available for other procedures.

For a fixed $\theta$ let

$$\frac{m_k(w,\theta)}{\Sigma_{k,k}(\theta)} \equiv h_k(\theta), \;\; \Omega_{k,i}(\theta) \equiv \frac{\Sigma_{i,k}(\theta)}{\sqrt{\Sigma_{i,i}(\theta), \Sigma_{k,k}(\theta)}},$$

that is we use the means relative to their standard error, and their correlation (here it is understood that all these terms are implicitly indexed by $J$ the sample size).

Let $k(max) \equiv arg\max_k h_k(w,\theta)$. So $h_{k(max)}(w,\theta)$ is the binding constraint and if it is negative we accept the null. So we are only concerned with cases where it is positive. We now derive a distribution for the binding

constraint *conditional on it being binding*. It will depend on a statistic, $V^{lo}_{k(max)}(\theta)$, which we now define.

Assume $\Sigma(\theta)$ is of full rank and define

$$V^{lo}_{k(max)}(\theta) \equiv max_{i \neq k(max)} \frac{h_i(\theta) - \Omega_{i,k}(\theta) h_{k(max)}(\theta)}{1 - \Omega_{i,k}(\theta)}$$

which, if we let

$$\Delta_{k(max)}(\theta) \equiv max_{i \neq k(max)} \frac{h_i(\theta) - h_{k(max)}(\theta)}{1 - \Omega_{i,k}(\theta)}$$

can be written as

$$V^{lo}_{k(max)}(\theta) = \Delta_{k(max)}(\theta) + h_{k(max)}(\theta),$$

which, since we are only concerned with the case where $h_{k(max)} \geq 0, \quad \Rightarrow \quad V^{lo}_{k(max)} \leq h_{k(max)}$, with probability one.

Also $V^{lo}_{k(max)}$is increasing (in absolute value) in the: (i) correlation and (ii) distance between, the binding constraint and the (covariance adjusted) next to binding, constraint.

Note that $\forall i \neq k$

$$\xi_{i,k}(\theta) = \frac{h_i(\theta) - \Omega_{i,k} h_k(\theta)}{1 - \Omega_{i,k}(\theta)} \Rightarrow E[h_k(\theta)\xi_{i,k}(\theta)] = 0.$$

Moreover, since both the $\xi_{i,k}$ and $h_k(\theta)$ are normal, this means that $h_k(\theta)$ is independent of the $K-1$ dimensional vector $[\xi_{1,k}, \ldots \xi_{K,k}]$. But $V^{lo}_{k,max}$ is just a linear function of $[\xi_{1,k(max)}, \ldots \xi_{K,k(max)}]$, so it is independent of $h_{k(max)}$ also. Moreover since $\Delta_{k(max)} < 0$, we have constructed $k(max)$ so that

$$h_{k(max)}(\theta) > V^{lo}_{k(max)}(\theta).$$

So the distribution of $h_{k(max)}(\theta)$ is the distribution of a truncated normal random variable*, truncated at $V^{lo}_{k(max)}$.

This implies that our critical value we look for a number, say $C_{\alpha,C}(\theta)$ such that the probability that the truncated normal distribution is greater than $C_{\alpha,C}(\theta)$ is less than $\alpha$ or

$$Pr(h_{k(max)}(\theta)) > C_{\alpha,C}(\theta)) = \frac{1 - \Phi(C_{\alpha,C}(\theta))}{1 - \Phi(V^{lo}_{k}(\theta))} = \alpha,$$

or

$$C_{\alpha,C}(\theta) = \Phi^{-1}\Big(1 - \alpha + \alpha\Phi(V^{lo}_{k}(\theta))\Big).$$

*This result follows from the work of R. Tibshirani, J. Taylor, R. Lockhart, and R.J. Tibsirani (2014). Exact Post-Selection Inference for Sequential Regression Procedures, Unpublished Manuscript.

**Notes.**

• Once we have $V_k^{lo}(\theta)$ the truncation point $C_{\alpha,C}(\theta)$ can be gotten from a standard computer program.

• The test will be powerful if $V_k^{lo}(\theta)$ is very small which will happen if the second highest (covariance adjusted) constraint is far from the first, and the correlation is very positive.

• The extreme case is when $\Omega_{i,k}(\theta) \to 1$, as then $V_k^{lo}(\theta) \to -\infty$ and we are back to a standard normal test. In the extreme case we need not worry about the possibility of other binding constraints.

• On the other hand if $V_k^{lo}(\theta)$ is very large (when $\Omega_{i,k,}(\theta)$ is very negative or the difference between the two moments is small) we will not reject much.

• $V^{lo}(\theta)$ is increasing in $h_{k(max)}(\theta)$, and so is less powerful the larger the binding moment (it has no power

when $h_{k(max)}(\theta) \to \infty$).

• This test is not sensitive to the addition of extraneous moments; as long as the extra moments do not bind at $\theta$ and are not the second highest (covariance adjusted) moment, the extra moments have no effect at all on the test statistic.

• Similar to RSW we can form hybrid critical values $C_{\alpha-\beta,C} = min(C_{\alpha-\beta,C}(\theta), C_{\beta,LF}(\Sigma(\theta)))$ which sacrifices a bit of power when there is a lot of power in return for insurance against cases where $V^{lo}_{k(max)}(\theta)$ is large.

**Simplifications in the linear case with conditional moments; Andrews and Pakes (2016).**

Let $\theta = (\beta, \delta)$ and our moment be

$$m(\cdot, \beta, \delta) = m(D, \beta, 0) + X\delta,$$

with

$$E[m(\cdot, \beta^0, \delta^0)|X_t] \leq 0.$$

The linearity here will help us with computation while the fact that we have conditional expectations will, as before, reduce the variance used in calculating critical values, and is likely to reduce the size of the CS.

Define

$$m(X) \equiv E_P[m(\cdot, \beta, \delta)|X],\ \mu_J = \sqrt{J}^{-1} \sum_j m(X_j)$$

$$\Sigma(\beta) = E_P(Var_P[m(\cdot, \beta, \delta)|X])$$

$$X_J = \sqrt{J}^{-1} \sum_j X_j,\ Y_J(\beta) = \sqrt{J}^{-1} \sum_j m(D_j, \beta, 0).$$

Then

$$Y_J(\beta^0) + X_J \delta^0 \to \mathcal{N}(\mu_J, \Sigma(\beta^0)) \text{ with } \mu_J \leq 0.$$

This formulation lets us use linear programming to find a CS for $\beta^0$.

First, find a critical value, say $C_\alpha(\Sigma(\beta))$. Note that since we are using conditional variances the critical value does not depend on $\delta$.

Then solve the LP

$$\Delta(\beta) \equiv min_{\delta}\Delta(\delta, \beta),$$

subject to

$$Y_T(\beta) + X_T\delta + \Delta \leq C_\alpha(\Sigma(\beta)).$$

Then

$$\beta \in CS(\theta) \;\; iff \;\; \Delta(\beta) \leq 0.$$

There are fast linear programming solutions that do not require you to cycle through the $\delta$'s – indeed the number of evaluations is tied to the number of moments, so this will be quick. Of course if everything was linear we could do this for every component of $\theta$ separately.

## References


- Abadie, A., G. Imbens, and F. Zheng, "Inference for Misspecified Models With Fixed Regressors," *Journal of the American Statistical Association*, 2014, 109(508): 1601-1614.

- Andrews, D. and P. Guggenberger, "Validity of Subsampling and 'Plug-in Asymptotic' Inference for Parameters Defined by Moment Inequalities," *Econometric Theory*, 2009, 25(3): 669-709.

- Andrews, D. and G. Soares, "Inference for Models Defined by Moment Inequalities Using Generalized Moment Selection Procedures," *Econometrica*, 2010, 78(1): 119-157.

- Andrews, I. and A. Pakes, "Linear Moment Inequalities" *in process*, Harvard, 2016.

- Armstrong, T., "A Note on Minimax Testing and Confidence Intervals in Moment Inequality Models, Yale working paper, 2014.

- Chernozhukov, V., H. Hong, and E. Tamer, "Estimation and Condence Regions for Parameter Sets in Econometric Models, *Econometrica*, 2007, 75(5): 1243-1284.

- Chetverikov, D., "Adaptive Tests of Conditional Moment Inequalities," UCLA working paper, 2013.

- Imbens, G. and C. Manski, "Condence Intervals for Partially Identied Parameters, *Econometrica*, 2004, 72(6): 1845-1857.

- Pakes A., and D. Pollard, "Simulation and the Asymptotics of Optimization Estimators", *Econometrica* 1989.

- Pakes, A., J. Porter, K. Ho, and J. Ishi, "Moment Inequalities and Their Application," *Econometrica*, 2015, 83(1): 315-334.

- Romano, J., A. Shaikh, and M. Wolf, "A Practical Two-Step Method for Testing Moment Inequalities," *Econometrica*, 2014, 82 (5): 1979-2002.

- R. Tibshirani, J. Taylor, R. Lockhart, and R.J. Tibsirani (2014): "Exact Post-Selection Inference for Sequential Regression Procedures" Unpublished Manuscript.