L2: AEA Continuing Education
Econometric Tools for Analyzing Moment Inequalities. *

Lecturer; Ariel Pakes, Harvard University.

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

Econometrics for Moment Inequalities.

Our model delivers the conditional moment restriction

$$\mathcal{E}\left[\Delta r(d_i, d', \mathbf{d}_{-i}, \mathbf{z}_i^o, \theta_0)|x\right] \leq 0, \text{ for almost every } x.$$

Note that we are starting with conditional moment inequalities, conditional on "x", since this is what theory typically delivers (the literature does not always make this distinction, but it will become important below).

Despite the fact that this typically generates an infinite set of moments, what we will do is derive a finite set of unconditional moments,

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

and use them in estimation. We are not going to ask how to chose these moments in an optimal way (for an attempt at doing so see, for e.g. Andrews and Shi,2013), though in any given application you should think through what you need in the way of moments to estimate different coefficients (to get bounds that are as tight as possible).

Estimator. Form sample analog and looks for values of θ that satisfy these moment inequalities. Note that these are inequalities and we generate set estimators.

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.

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

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

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

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

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

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

Let

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

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 $f(\cdot)_+ \equiv max(0, f(\cdot))$ 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

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

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

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.

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 θ_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 θ is available. Call that estimate $\widehat{D}_J(\theta)$ (a diagonal matrix). Then, estimation proceeds as follows. Set

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

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 negative 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

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

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 θ say $\theta_{0,1} \in [\underline{\theta}_1, \overline{\theta}_1]$.

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

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

Then

$$\mathsf{Pr}\left\{[\underline{\theta}_1, \overline{\theta}_1] \subset [\widehat{a}, \widehat{b}]\right\} \geq$$

$$1 \ - \ \mathsf{Pr}\left\{\widehat{a} > \underline{\theta}_1\right\} \ - \ \mathsf{Pr}\left\{\widehat{b} < \overline{\theta}_1\right\} = 1 - \alpha.$$

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

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

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 θ. 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 at $\underline{\theta}_1$, i.e. $\mathcal{P}m_0(w, \underline{\theta}_1) = 0$, and

- Non-binding at $\underline{\theta}_1$:i.e. $\mathcal{P}m_1(w, \underline{\theta}_1) < 0$.

With probability approaching one $\Theta_J = \{\theta : \mathbf{P}_J m(w,\theta) \leq 0\}$. So stochastic equicontinuity, and $\underline{\widehat{\theta}} - \underline{\theta} = O_p(1/\sqrt{J})$, *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

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

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

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 $\underline{\theta}$.

It suffices to consider only the binding moments (the $m_0(w,\underline{\theta}_1)$). Note that $\mathcal{P}m_0(w,\underline{\theta}) = 0$ (in the second and third terms). So the binding moments can be expressed as

$$\sqrt{J}\mathbf{P}_J m_0(w,\underline{\hat{\theta}}) = \sqrt{J}\mathcal{P}m_0(w,\underline{\hat{\theta}}) + \sqrt{J}\mathbf{P}_J m_0(w,\underline{\theta}) + o_p(1).$$

If we linearize the first term and consider the implications of the theory on the rhs approximation we have

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

where

$$\Gamma_0 \equiv \frac{\partial \mathcal{P}m_0(w,\theta)}{\partial\theta}|_{\theta=\underline{\theta}}$$

which we assume has full column rank. The fact that the estimator is $\sqrt{J}$ consistent insures that the approximation error from the expansion is $o_p(1)$.

This is still an inequality so we cannot solve for the distribution of the estimator directly (as we would do with m.o.m.). We do know that $\sqrt{J}\mathbf{P}_J m_0(w,\underline{\theta}) \sim N(0,\boldsymbol{\Sigma}_0)$. This implies that the distribution of $\sqrt{J}(\underline{\widehat{\theta}}-\underline{\theta})$ is is given by the following theorem.

Theorem.

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

where

$$\widehat{\tau} = arg \min_{\left[0 \leq \Gamma_0 \tau + Z\right]} \tau_1; \quad \text{and} \;\; Z \sim N(0,\boldsymbol{\Sigma}_0) \quad ♠.$$

There is no ”pivotal” distribution for the solution to this problem, but it is easy to simulate its distribution.

Take random draws on a normal with the appropriate covariance matrix, and solve a linear programming problem for each one to determine the set of values that it accepts. We can then form a confidence set for a point (a point is in the CS if it is covered by 95% of the simulated draws.).

Consider two cases.

- $dim(m_0) = dim(\theta)$. One might think this is the leading case (just as many parameters as binding moments) 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). It is the absence of the derivative to $\mathcal{P}m_0(x,\theta)$ that violates our standard regularity conditions for this case.

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 (Γ_0, Σ_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 θ_0 that covers the true parameter with probability at least $1 - \alpha$ (and the CS's we find will tend to be ”conservative”; i.e. they will tend to cover with probability greater than $1 - \alpha$) . Moreover, because the expectation of the objective function is non-differentiable at θ_0, there is no longer a reason to think that any estimate of a function of Θ_0, for example $\underline{\theta}_1$, distributes normally (or for that matter has any ”pivotal” distribution). 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” over $P \in \mathcal{P}$ is over all data generating processes, including ones which generate a Θ_I where many more moments bind than there are parameters being estimated.

Least favorable Confidence Sets for the Point, θ_0.

Intuition. What we do is assume that all the moments of the model are exactly zero at each θ, 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 θ. If the sample moment evaluation is greater than that of the $1-\alpha$ quantile, then the value of θ 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 $m_J(\theta) = m(P_J, \theta)$. Define

$$R(m_J(\theta), \Omega_J(\theta)) = max_k(max[\frac{m_k(\theta)}{\sqrt{\Sigma_{J(k,k)}(\theta)}}, 0])$$

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

For each θ we look for a number, the critical value, $C_\alpha(m(\theta), \Omega_J(\theta))$, where $\Omega_J(\theta)$ is the correlation matrix of the data such that

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

To obtain the least favorable $C_\alpha(m(\theta), \Omega_J(\theta))$ we simulate from a normal with mean zero and a covariance matrix equal to the the correlation matrix variance of the data many times and compute $R(0, \Omega_J(\theta))$ for each simulation run. The α quantile of that statistic over the simulated samples is $C_\alpha(0, \Omega_J(\theta))$. We then go back to the data and find out if

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

The particular θ 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 generates a conservative CS, and we call it 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 Θ. 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 $C_\alpha(0, \Omega_J(\theta))$.

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 θ). I.e. we hold the random draws fixed as we look over alternative θ.

Step 3. Go back to the data. Compute the value of the objective function at θ_l. Accept all θ_l for which $R(m_J(\theta_l), \Omega_J(\theta_l)) \leq C_\alpha(0, \Omega_J(\theta_l))$

Note that no matter what value the true θ_0 is, it will be in this set with probability $1 - \alpha$. Hence, it is a confidence set with significance level at most α.

Computational burden. The simulation is easy enough for a fixed θ. 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 Θ_L. Most would estimate $\hat{\Theta}_I$ (the estimate of the identified set) first, and then use that as a basis for defining Θ_L. You need a Θ_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) ϵ (and at least in non-linear models this may be hard to determine).

For a large dimensional θ 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 θ requires a fixed point calculation. We come back to ways of alleviating the computational burden below as their are cases where this limits the use of moment inequality estimators.

The number of moments and the 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 θ will (weakly) increase the estimate of $C_\alpha(0, \Omega_J(\theta))$. This is a bit counterintuitive; adding moments, which should be adding 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 early versions of these processes required a ”tuning” parameter much like the bandwidth used in nonparametric estimation. The paper by Romano, Shaikh, and Wolf (2014) does not require a “tuning” parameter, and so, at least initially, 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.

[*] 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) for a discussion of these alternatives.

The number of moments and the precision of the estimated variance-covariance matrix. In what follows, and in most of econometrics, we are going to ignore issues that might arise as a result of the imprecision of the estimate of Σ (or Ω). However as we increase the number of moments, we increase the number of components of Ω we are estimating at the rate of the square of the number of moments. At some point those estimates are going to become imprecise. You should keep this in mind when choosing the number of moments.

Romano, Shaikh, and Wolf and Shifted Moments.

They do an initial step which finds the least favorable critical value for size β. That step uses the max norm for obtaining the critical value. They then form the following "shifted" mean

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

$$\min\{m_k(\theta) + \Sigma_{J(k,k)}^{1/2} C_\beta(0, \Omega_J(\theta)), 0\}.$$

Note if the original moment was negative, this moment can be negative, and it will be more negative the more negative the initial mean.

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

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

A θ is put in the CS if and only if

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

Since we now have shifted means negatively, less random draws will be positive, so we expect the critical value to fall.

They prove that if the CS is formed in this way

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

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

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

A Note on Testing.

Testing is likely to be quite important in the moment inequality context when there are many inequalities and the sample underlying the calculation of each is "small". If the model is correct and we had unlimited data all of the inequalities would converge to their limit values (uniformly in θ), and we would find values of θ 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 misspecified. 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 θ for which the sample moments are all positive) then you will never reject any test.

- If one has estimated a confidence set by the "pointwise" methods discussed above, then you have already computed a test statistic. I.e. if there is no θ 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 θ_0 such that $Pm(z, \theta_0) \geq 0$. For this θ_0, our critical values are constructed such that θ_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 $\Pr(Q_n(\theta_0) \geq c(\alpha, \theta_0) \mid H_0) \leq \alpha$.

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

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 in formating the critical value we can simulate from the the average of the conditional variances. From standard probability theory

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

and we have a similar decomposition for the sample covariance

$$\Sigma_J(\theta) = Var(\frac{1}{\sqrt{J}}\sum_j m_{k,j}(w_j,\theta)) =$$

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

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

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

So if we let

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

and use it instead of $\Sigma_J(\theta)$ in the formula above 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 Abadie et al (2014)[*]. Let

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

where

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

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

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

Here the "2" takes account of the variance in both observations. Note that we do not have to do this for

[*] This requires compact support and conditional expectations, i.e. $E[m(w,\theta)|X]$, that are sufficiently smooth in X (Lipshitz in X).

each θ but rather just once. So though there is an added computational step, it is not too onerous.

• 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 pushes us to think we will do better using the conditional variance than the total variance (since the conditional variance is smaller we should get 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 as then we might not be able to tell which of the parameters we need to change to satisfy the objective function.).

Inference on Functions of Parameters.

We typically want to find a CS for $\beta = f(\theta)$. The most frequent applied case is β is a component of θ for then we would be obtaining a CI for this parameter. This may be because we are focused on a particular parameter (or a linear combination thereof). However regardless of what we are interested in, if we are writing for a journal we are going to have to report summary statistics for measures of precision for each parameter and the most familiar of these would be CI's for the parameters.

Projection Method. If we already have a confidence set for θ, then to find out if a particular β is acceptable all we need do is find out is 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, i.e. we search directly for

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

When $f(\theta) = \theta_1$, then the minimum in the above expression is over values of $(\theta_2, \ldots \theta_K)$ s.t. $\theta \in \Theta$.

As noted earlier the computational burden for computing the whole CS can be immense for large dimensional parameter vectors, so if we can decrease the computational burden by going after particular functions of θ it can be a very good thing.

The next section shows that In the linear case we can essentially make the computational problem disappear. Moreover it also shows that in the linear case we also have ways of obtaining sharper confidence intervals then we would obtain using the techniques described above.

The Linear Case; Andrews and Pakes, 2016.

We can offer substantial gains in computational and statistical efficiency in the linear case. These gains are only available when we use the conditional variance formula above (so our variance is V and correlation is Ω_V).

If we let K be the cardinality of the partition of any given axis, the linear program will drops the computational burden of finding CI's from being geometric in the number of parameters (i.e. $K^{\#\theta}$) to being a low order polynomial in the number of parameters $(\#\theta)^3$? times a constant which depends on the number of inequalities.

Notation. Say $\theta = (\beta, \delta)$, and we are interested in testing whether a given value of β, say β_0 is in the identified set (so δ is being treated as a "nuisance" parameter). If we need CI's for all parameters we are going to have to do the test below separately for each component of θ.

We are trying to find a confidence interval for the interval

$$I_\beta(P) = \{\beta : \exists \delta \; s.t. \; E_P[m(d, \beta, \delta)|z] \leq 0, \text{ a. e. } z\}.$$

If our moment is "linear" i.e. we can write it as is

$$m(z, d, \beta, \delta) = m(d, \beta, 0) - x(z, \beta)\delta,$$

so we are generalizing the conventional linear regression model (set $m(d, \beta, 0) = y$) to allow for models that generate inequalities. Note that our y and x variables can depend on parameters, and in that sense we can have constructed regressors and dependent variables. Below we omit the dependence of the $m(\cdot)$ on β for notational convenience and just say $m(d, \beta, 0) = y$.

We assume throughout the asymptotic approximation

$$Y_J|_{\mu_J, X_J} \sim \mathcal{N}(\mu_J - X_J\delta, V_J),$$

and we want to test the null

$$H_0 : \exists\delta \text{ such that } \mu_J - X_J\delta \leq 0,$$

The set of μ that are accepted by the test is our CI.

We are going to base our test on the maximum moment.

$$R(Y_J - X_J\delta, \Omega_V) = max_k\{\frac{max[Y_{J,k} - X_{J,k}\delta, 0]}{\sqrt{V_{J,k,k}}}\}.$$

Where I have used Ω_V to indicate that we are using normalized moments. In what follows I am going to omit the index J for notational convenience.

Projection Method in the Linear Case. To test if we can accept a particular $\mu \in H_0$ we look for a $C_\alpha(\mu - X\delta(\mu), \Omega_V)$ such that

$$Pr\{\min_\delta \max_k[\frac{Y_k - X_k\delta}{\sqrt{V_{k,k}}}] \geq C_\alpha(\mu - X\delta(\mu), \Omega_V)] = \alpha.$$

Both sides of this inequality look difficult to calculate. Starting with the left hand side, we note that it can be turned into a linear programming problem by noting that if we solve

$$\widehat{\eta}^* = \min_{\delta,k} \eta(\delta, k),$$

such that

$$\frac{Y_k - X_k\delta}{\sqrt{V_{k,k,}}} \leq \eta(\delta, k),$$

then $\widehat{\eta}^*$ is the solution to the problem on the lhs of our inequality and we accept H_0 if and only if

$$\widehat{\eta}^* \leq C_\alpha(\mu - X\delta(\mu), \Omega_V).$$

This problem is a linear programming problem can be solved quickly even for high dimensional problems (typically problems thousands of variables and thousands of moments can be solved very quickly).

We still need to find $C_\alpha(\mu - X\delta(u), \Omega_V)$, the α level critical value for a normal with the appropriate mean and variance.

Observation 1. $\forall\tilde{\delta}, \quad C_\alpha(\mu - X\delta(\mu), \Omega_V) = C_\alpha(\mu - X\delta(\mu) + X\tilde{\delta}, \Omega_V)$. That is the critical value of the test statistic does not change if we add $X\tilde{\delta}$ to the mean. To see this let

$$\widehat{\eta}(\mu,\epsilon) = \min_\delta \max_k [\frac{\mu_k - X\delta + \epsilon}{\sqrt{V_{k,k}}}] \equiv \max_k [\frac{\mu_k - X\widehat{\delta}(\mu,\epsilon) + \epsilon}{\sqrt{V_{k,k}}}].$$

But for any $\tilde{\delta}$

$$\hat{\eta}(\mu + X\tilde{\delta}, \epsilon) = \min_{\delta} \max_{k} [\frac{\mu_k + X_k\tilde{\delta} - X_k\delta + \epsilon}{\sqrt{V_{k,k}}}]$$

and replacing δ with $\hat{\delta}(\mu, \epsilon) - \tilde{\delta}$, which is one candidate value of δ, this has to be

$$\leq max_k [\frac{\mu_k + X_k\tilde{\delta} - X_k(\hat{\delta}(\mu, \epsilon) - \tilde{\delta}) + \epsilon}{\sqrt{V_{k,k}}}]$$

$$= \max_{k} [\frac{\mu_k - X_k\hat{\delta}(\mu, \epsilon) + \epsilon}{\sqrt{V_{k,k}}}] = \hat{\eta}(\mu, \epsilon)$$

and we could make the analogous argument for adding $-\tilde{\delta}$ to μ. So for any $\tilde{\delta}$ the critical value is the same as for δ.

Intuition. The minimization gives us freedom to chose any linear combinations of X. So you can add or subtract a linear combination to start, and then undo it in the minimization process.

Observation 2. $C_\alpha(\mu - X\delta(\mu), \Omega_V)$ is increasing in μ (element by element). This follows from the fact that $\eta(\mu, \Omega_V) = \max_k \min_\delta [\mu_k - X_k\delta + \epsilon]/\sqrt{V_{k,k}}]$ is stochastically increasing in μ in the first order dominance sense. I.e.

$$\mu_1 \geq \mu_2, \Rightarrow \ \eta(\mu_1, \Omega_V) = max_k[\mu_k - X_k\delta(\mu_1) + \epsilon]/\sqrt{V_{k,k}}] \geq$$

$$max_k[\mu_k - X_k\delta(\mu_1) + \epsilon]/\sqrt{V_{k,k}}] \geq \eta(\mu_2, \Omega_V)$$

Proposition. Here "LF" refers to the least favorable critical value. We have

$$C_\alpha(LF, \Omega_V) = sup_{\mu \in H_0} C_\alpha(\Omega_V)$$

$$= sup_{\mu - X\delta(\mu) \leq 0} C(\Omega_V) = \sup_{\mu \leq 0} C_\alpha(\Omega_V) = C_{\mu=0}(\Omega_V).$$

Proof. The second equality follows from the fact that

$$\forall \mu \in H_0, \quad \exists \delta(\mu) \; s.t. \; \mu - X\delta(\mu) \leq 0.$$

The third equality follows from observation 1, and the fourth from observation 2. ♠.

So we can find $C_\alpha(LF, \Omega_V)$, by taking random draws from $\mathcal{N}(0, \Omega)$, solving the linear program

$$\widehat{\eta} = \min_{\delta} \eta(\delta)$$

subject to

$$max_k[\epsilon_k - \frac{X_k}{V_{k,k,}}\delta] \leq \eta(\delta)$$

many times and setting $C_\alpha(LF, \Omega_V)$ equal to the 95th of the resulting distribution.

Note: This will be a smaller critical value than we would have gotten from the earlier methods.

To see this consider what would happen if we had calculated a critical value for $\mu = 0$ by simulating the objective function for different values of δ. Say we were using the critical value from NS simulation draws. We would have drawn $\{\epsilon_{ns}\}_{ns=1}^{NS}$, calculated $\{\eta(\epsilon_{ns}(\delta))\}$ for each one of them for different values of δ and if $Q^{.95}(z)$

is notation for the 95th quantile of the vector z, we would set for each δ

$$\eta^{.95}(\delta) = Q^{.95}\Big\{\eta(\epsilon_1(\delta)), \ldots, \eta(\epsilon_{NS}(\delta)\Big\},$$

and then solve

$$\eta^* = \min_{\delta} \eta^{.95}(\delta).$$

The current procedure sets our

$$\eta = Q^{.95}\left\{\eta(\min_{\delta}(\epsilon_1(\delta)), \ldots, \eta(\min_{\delta}(\epsilon_{NS}(\delta))\right\}. \ \spadesuit$$

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.

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

- 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 Confidence 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, “Confidence Intervals for Partially Identified 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.