ECON 556 Lecture Notes

Tim Armstrong

last updated: September 26, 2017

1 Introduction

- This section of the course will cover two recent literatures in econometrics: *weak instruments/weak identification* and *moment inequalities*.
- We'll give some intuition for theoretical motivation, but will focus on practical aspects, particularly computation.
- In "standard" setting (GMM under the usual regularity conditions), estimation and inference are handled using the same machinery:
  - Form an estimate asymptotically normal $\hat{\theta}$ of a parameter θ.
  - Report $\hat{\theta}$ as a point estimate and form confidence regions for θ based on this asymptotic distribution (e.g. $\hat{\theta}_j \pm z_{1-\alpha/2} \cdot \text{se}_j$ as a confidence interval for θ_j).
  - Easy to define relative efficiency using the same criterion for estimation, CI construction and testing: smaller asymptotic variance is always better.
- This breaks down in the weak IV and moment inequality settings.
  - Can't form CI as estimate plus-or-minus a constant times standard error (for one thing, estimates aren't asymptotically normal).
  - Rather, these literatures have focused on inverting tests for

    $$H_\theta : \text{ true parameter value is } \theta$$

    leading to confidence regions of the form $\mathcal{C} = \{\theta : T(\theta) \leq c(\theta)\}$ where $T(\theta)$ is a test statistic and $c(\theta)$ is a critical value.

    - The confidence region $\mathcal{C}$ typically does not simplify further. This contrasts with "standard setting" in which the CI $\hat{\theta}_j \pm z_{1-\alpha/2}\text{se}_j$ inverts the two-sided z-test (based on $T(\theta) = |\hat{\theta}_j - \theta_j|/\text{se}_j$ and $c(\theta) = z_{1-\alpha/2}$).
    - We will focus (more than in a typical econometrics course) on issues that arise in computing and reporting these confidence sets.
  - The weak IV and moment inequality literatures have focused less on point estimation than on inference (although there are some exceptions). Point estimation in these settings is an important topic as well (people usually report point estimates!), but we won't spend much time on it in this course.

2 Weak Identification

- Consider the GMM model based on $g(\theta) = Eg(w_i, \theta)$ where θ is an unknown parameter in $\mathbb{R}^{d_\theta}$, $g(w_i, \theta)$ is a known function taking values in $\mathbb{R}^{d_g}$ and we observe $\{w_i\}_{i=1}^n$.
- The true parameter value θ_0 satisfies $g(\theta_0) = 0$.
- Let

$$\hat{g}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(w_i, \theta)$$

  denote the sample mean. In the "standard" setting, we would use the GMM estimator $\hat{\theta} = \arg\min_\theta \hat{g}(\theta)' W_n \hat{g}(\theta)$ where W_n is a sequence of weighting matrices.

  - Under "standard" regularity conditions,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, (G(\theta_0)'WG(\theta_0))^{-1} G(\theta_0)'W\Sigma(\theta_0)WG(\theta_0)(G(\theta_0)'WG(\theta_0))^{-1})$$

    where

$$G(\theta) = \frac{d}{d\theta'} g(\theta), \quad \Sigma(\theta) = \text{var}(g(w_i, \theta))$$

  - To form a CI, we then estimate $G(\theta)$ and $\Sigma(\theta)$ using

$$\hat{G}(\theta) = \frac{d}{d\theta'} \hat{g}(\theta), \quad \hat{\Sigma}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(w_i, \theta) g(w_i, \theta)' - \hat{g}(\theta)\hat{g}(\theta)'$$

(this form of $\hat{\Sigma}(\theta)$ is for the iid case; if there is dependence, then an estimate taking this into account will be needed) and form the CI for θ_j as $\hat{\theta}_j \pm z_{1-\alpha/2}\text{se}_j$ where se_j is $1/\sqrt{n}$ times the square root of the j, jth element of

$$(\hat{G}(\hat{\theta})'W_n\hat{G}(\hat{\theta}))^{-1}\hat{G}(\hat{\theta})'W_n\hat{\Sigma}(\hat{\theta})W_n\hat{G}(\hat{\theta})(\hat{G}(\hat{\theta})'W_n\hat{G}(\hat{\theta}))^{-1}.$$

- A key regularity condition needed for this is "strong identification:" we need the equations $g(\theta) = 0$ to be solved uniquely at θ_0, and we need $G(\theta_0)$ to be full rank.
- If $G(\theta_0)$ is not full rank, then the CI may not have correct coverage.
  - Note that it is not clear a priori whether there will be overcoverage or undercoverage: the estimate will be inaccuate, but the standard errors will also be large with high probability, since $\hat{G}(\hat{\theta})$ will be close to rank deficient with high probability.
  - However, it turns out that severe undercoverage is possible. One way of seeing this is to note that, in the case where $g(\theta) = 0$ for arbitrarily large values of θ_j (i.e. the data cannot rule out arbitrarily large values of θ_j), the CI should be infinite with high probability. However, typically $\hat{G}(\hat{\theta})$ will be full rank with probability one (even when $G(\theta)$ is not full rank) due to sampling error, so se_j will be finite with probability one.
- The literature on *weak identification* or *weak instruments* focuses on cases where $G(\theta_0)$ is not full rank (including lack of identification), or is "close to" not being full rank. The latter case is modeled using sequences of data generating processes (dgps) that change with the sample size, and we will not pursue it formally here. Note, however, that a test is called level α when the rejection probability is less than α for *all* null distributions. Thus, for the level of a test to be less than $\alpha+\varepsilon$ for large n, we need it to be asymptotically level $\alpha + \varepsilon$ for any sequence of null dgps. For this reason, sequences of dgps that change with the sample size are relevant for controlling the level of a test according to the "usual" definition of significance level.
- We will discuss some tests that from this literature, which are designed to be robust to weak identification. First, we will give some examples of models where weak identification or lack of identification can be an issue.

2.1 Examples

- Linear IV The linear IV model is GMM with $g(x_i, y_i, z_i, \theta) = (y_i - x_i'\theta)z_i$ where y_i is scalar valued and z_i takes values in $\mathbb{R}^{d_g}$. In this case, $G(\theta) = -Ez_i x_i'$, so the identification assumption required for "standard" asymptotics is that $Ez_i x_i'$ must be full rank. This amounts to an assumption that the instruments z_i have to be correlated with the endogenous variables x_i (in the case where there are multiple variables in x_i that are not in z_i, this also requires that the correlation with each instrument be different in the right way to make this matrix nonsingular).

- Nonlinear IV Nonlinear IV is GMM with $g(x_i, y_i, z_i, \theta) = \rho(x_i, y_i, \theta)z_i$ for some known function ρ. In this case, $G(\theta) = Ez_i \frac{d}{d\theta'}\rho(x_i, y_i, \theta)$, so the identification assumption is similar to the linear case, but with $\frac{d}{d\theta'}\rho(x_i, y_i, \theta)$ in place of x_i, so that z_i needs to be correlated with this nonlinear function of x_i, y_i rather than with x_i.

2.2 Anderson-Rubin Test

- The issues with the "standard" approach arise because the normal approximation for $\hat{\theta}$ is bad when $G(\theta_0)$ is close to singular.

- To get around this, we can form a test based on $\hat{g}(\theta)$, which is asymptotically normal even when $G(\theta) = 0$. In particular, we have

$$\sqrt{n}(\hat{g}(\theta) - g(\theta)) \xrightarrow{d} N(0, \Sigma(\theta)), \quad \hat{\Sigma}(\theta) \xrightarrow{p} \Sigma(\theta)$$

  so long as a central limit theorem (CLT) and law of large numbers (LLN) hold. In the iid case, it is sufficient that $g(w_i, \theta)$ have a finite second moment. No conditions on $G(\theta)$ are needed.

- Note that, if the true parameter value is equal to θ, then $g(\theta) = 0$. Thus, we can base tests about θ on estimates of $g(\theta)$.

- The Anderson-Rubin (AR) test of the null that the true parameter value is equal to θ is based on the statistic

$$S(\theta) = n\hat{g}(\theta)'\hat{\Sigma}(\theta)^{-1}\hat{g}(\theta).$$

- Under the null, $g(\theta) = 0$, so

$$S(\theta) = [\sqrt{n}(\hat{g}(\theta) - g(\theta))]'\hat{\Sigma}(\theta)^{-1}[\sqrt{n}(\hat{g}(\theta) - g(\theta))] \xrightarrow{p} \chi^2_{d_g},$$

  where $\chi^2_{d_g}$ denotes the chi-square distribution with d_g degrees of freedom (recall that a quadratic form in a normal vector with the matrix given by the inverse of its variance is distributed χ^2 with the dimension of the vector as the degrees of freedom).

- Let $q_{\chi^2,d_g,1-\alpha}$ denote the $1-\alpha$ quantile of the $\chi^2_{d_g}$ distribution. The AR test rejects when $S(\theta) > q_{\chi^2,d_g,1-\alpha}$. The resulting confidence region is given by

$$\mathcal{C} = \{\theta : S(\theta) \leq q_{\chi^2,d_g,1-\alpha}\}.$$

- The Anderson-Rubin test for GMM was proposed by Stock and Wright (2000), and is an extension of the test proposed by Anderson and Rubin (1949) for the IV model with normal and homoskedastic errors.

- The confidence set $\mathcal{C}$ is a d_θ dimensional confidence set for the parameter θ. If $d_\theta > 3$, then it is not clear how to report it. In practice, one can report *projections* of the set $\mathcal{C}$ onto each component. The projection onto the jth component is

$$\{\theta_j : \theta \in \mathcal{C}\}.$$

  We will discuss computing this later on. Note that $\mathcal{C}$ is not necessarily connected, and the projection is not necessarily an interval (although we can report the smallest interval containing it if we want to report an interval).

2.3 Kleibergen's K Statistic

- An issue with the AR statistic is that it is inefficient under the "usual" ("strong identification") asymptotics when $d_g > d_\theta$ (i.e. when the model is potentially overidentified).
  - Intuition: under strong identification, GMM with optimal weighting uses an estimate of $G(\theta_0)$ to choose the optimal combination of instruments. In contrast, AR statistic does not use this information.

- To fix this, one can base inference on the joint asymptotic distribution of $\hat{g}(\theta)$ and

$\hat{G}(\theta)$:

$$\sqrt{n}\begin{pmatrix} \hat{g}(\theta) - g(\theta) \\ \text{vec}(\hat{G}(\theta)) - \text{vec}(G(\theta)) \end{pmatrix} = \sqrt{n}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}[g(w_i,\theta) - Eg(w_i,\theta)] \\ \frac{1}{n}\sum_{i=1}^{n}\left[\text{vec}\left(\frac{d}{d\theta'}g(w_i,\theta)\right) - E\text{vec}\left(\frac{d}{d\theta'}g(w_i,\theta)\right)\right] \end{pmatrix}$$

$$\xrightarrow{d} N\left(0, \begin{pmatrix} \underbrace{\Sigma(\theta)}_{d_g \times d_g} & \underbrace{V_{gG}(\theta)}_{d_g \times (d_g \cdot d_\theta)} \\ \underbrace{V_{Gg}(\theta)}_{(d_g \cdot d_\theta) \times d_g} & \underbrace{V_{GG}(\theta)}_{(d_g \cdot d_\theta) \times (d_g \cdot d_\theta)} \end{pmatrix}\right)$$

where $\text{vec}(A)$ stacks the columns of a matrix A into a vector. In the iid case, this holds by the CLT so long as $g(w_i, \theta)$ and its derivatives have bounded second moments, and the asymptotic variance is

$$\begin{pmatrix} \Sigma(\theta) & V_{gG}(\theta) \\ V_{Gg}(\theta) & V_{GG}(\theta) \end{pmatrix} = \text{var}\begin{pmatrix} g(w_i,\theta) \\ \text{vec}\left(\frac{d}{d\theta'}g(w_i,\theta)\right) \end{pmatrix}$$

which can be estimated using

$$\begin{pmatrix} \hat{\Sigma}(\theta) & \hat{V}_{gG}(\theta) \\ \hat{V}_{Gg}(\theta) & \hat{V}_{GG}(\theta) \end{pmatrix}$$
$$= \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} g(w_i,\theta) \\ \text{vec}\left(\frac{d}{d\theta'}g(w_i,\theta)\right) \end{pmatrix}\begin{pmatrix} g(w_i,\theta) \\ \text{vec}\left(\frac{d}{d\theta'}g(w_i,\theta)\right) \end{pmatrix}' - \begin{pmatrix} \hat{g}(\theta) \\ \text{vec}(\hat{G}(\theta)) \end{pmatrix}\begin{pmatrix} \hat{g}(\theta) \\ \text{vec}(\hat{G}(\theta)) \end{pmatrix}'$$

- Recall the score/Lagrange multiplier statistic of the null that the true value is equal to θ. It is based on

$$\frac{d}{d\theta}\frac{1}{2}\hat{g}(\theta)'W\hat{g}(\theta) = \hat{G}(\theta)'W\hat{g}(\theta)$$

and is given by a quadratic form of this quantity, using an estimate of its variance:

$$LM(\theta) = n\left[\hat{G}(\theta)'W\hat{g}(\theta)\right]'\left[\hat{G}(\theta)'W\hat{\Sigma}(\theta)W\hat{G}(\theta)\right]^{-1}\left[\hat{G}(\theta)'W\hat{g}(\theta)\right].$$

- Note that

$$\sqrt{n}\hat{G}(\theta)W\hat{g}(\theta) \xrightarrow{d} G(\theta)WZ_g \quad \text{and} \quad \hat{G}(\theta)'W\hat{\Sigma}(\theta)W\hat{G}(\theta) \xrightarrow{p} G(\theta)'W\Sigma(\theta)WG(\theta)$$

where $Z_g \sim N(0, \Sigma(\theta))$ follows the asymptotic distribution of $\hat{g}(\theta)$. If $G(\theta)$ is full rank, then the latter matrix will be invertible, so that the LM statistic will have an asymptotic $\chi^2_{d_\theta}$ distribution.

- If $G(\theta)$ is not full rank, then $\hat{G}(\theta)'W\hat{\Sigma}(\theta)W\hat{G}(\theta)$ will not converge to an invertible matrix. This will lead to sampling error in $\hat{G}(\theta)$ playing a role in the asymptotic distribution.

- The K statistic, proposed by Kleibergen (2005), gets around this by using an estimate of $G(\theta)$ that is asymptotically independent of $\hat{g}(\theta)$.

- Recall that, if x and y are joint normal, then $\tilde{y} = y - cov(y,x)var(x)^{-1}x$ and x are normal and independent (since $cov(\tilde{y},x) = cov(y,x) - cov(cov(y,x)var(x)^{-1}x, x) = cov(y,x) - cov(y,x)var(x)^{-1}cov(x,x) = 0$). We can apply this (asymptotically) to $\hat{G}(\theta)$ and $\hat{g}(\theta)$: let $\hat{D}(\theta)$ be the $d_g \times d_\theta$ matrix such that

$$\text{vec}(\hat{D}(\theta)) = \text{vec}(\hat{G}(\theta)) - \hat{V}_{Gg}(\theta)\hat{\Sigma}^{-1}(\theta)\hat{g}(\theta).$$

- The K statistic is formed by replacing $\hat{G}(\theta)$ with $\hat{D}(\theta)$ in the LM statistic:

$$\begin{aligned} K(\theta) &= n\left[\hat{D}(\theta)'W\hat{g}(\theta)\right]'\left[\hat{D}(\theta)'W\hat{\Sigma}(\theta)W\hat{D}(\theta)\right]^{-1}\left[\hat{D}(\theta)'W\hat{g}(\theta)\right] \\ &= n\left[\hat{D}(\theta)'\hat{\Sigma}^{-1}(\theta)\hat{g}(\theta)\right]'\left[\hat{D}(\theta)'\hat{\Sigma}^{-1}(\theta)\hat{D}(\theta)\right]^{-1}\left[\hat{D}(\theta)'\hat{\Sigma}^{-1}(\theta)\hat{g}(\theta)\right] \end{aligned}$$

  where we use $W = \hat{\Sigma}^{-1}(\theta)$ as the weighting matrix.

- It can be shown that, even when $G(\theta)$ is not full rank,

$$K(\theta) \xrightarrow{d} \chi^2_{d_\theta}$$

  when the null hypothesis $g(\theta) = 0$ holds. This leads to the confidence set

$$\mathcal{C} = \{\theta : K(\theta) \leq q_{\chi^2, d_\theta, 1-\alpha}\}.$$

- To understand this asymptotic distribution result, consider the case where $G(\theta)$ is a matrix of zeros. Then, under the null hypothesis, $\sqrt{n}\hat{g}(\theta) \xrightarrow{d} Z_g$ and $\sqrt{n}\hat{D}(\theta) \xrightarrow{d} Z_D$

jointly where Z_g and Z_D are independent with $Z_g \sim N(0, \Sigma(\theta))$. Thus,

$$\begin{aligned} K(\theta) &= \left[\sqrt{n}\hat{D}(\theta)'\hat{\Sigma}^{-1}(\theta)\sqrt{n}\hat{g}(\theta)\right]' \left[\sqrt{n}\hat{D}(\theta)'\hat{\Sigma}^{-1}(\theta)\sqrt{n}\hat{D}(\theta)\right]^{-1} \left[\sqrt{n}\hat{D}(\theta)'\hat{\Sigma}^{-1}(\theta)\sqrt{n}\hat{g}(\theta)\right] \\ &\xrightarrow{d} \left[Z_D'\Sigma^{-1}(\theta)Z_g\right]' \left[Z_D'\Sigma^{-1}(\theta)Z_D\right]^{-1} \left[Z_D'\Sigma^{-1}(\theta)Z_g\right]. \end{aligned}$$

By independence of Z_D and Z_g, the conditional distribution of $Z_D'\Sigma^{-1}(\theta)Z_g$ given Z_D is $N(0, Z_D'\Sigma^{-1}(\theta)\Sigma(\theta)\Sigma^{-1}(\theta)Z_D) = N(0, Z_D'\Sigma^{-1}(\theta)Z_D)$. Thus, the asymptotic distribution in the above display is $\chi^2_{d_\theta}$ conditional on Z_D, which means that it is $\chi^2_{d_\theta}$ unconditionally.

2.4 Tests Based on Both $K(\theta)$ and $S(\theta)$

- The K statistic can be shown to be equal to a quadratic form in the derivative of the continuous updating estimator (CUE) objective function $\hat{g}(\theta)\hat{\Sigma}^{-1}(\theta)\hat{g}(\theta)$. This reveals an issue with the power of the K statistic: we will fail to reject at any local minimum of the CUE objective.

- Another way of seeing this is to note that the K statistic is a quadratic form in $\hat{D}(\theta)'\hat{\Sigma}^{-1}\hat{g}(\theta)$, so it will fail to reject when $\hat{D}(\theta)$ is close to not being full rank even when $\hat{g}(\theta)$ is far from zero. So we want to have some way of rejecting when $\hat{D}(\theta)$ is close to not being full rank.

- To ameliorate this, one can form a test statistic based on both $K(\theta)$ (the K statistic) and $S(\theta)$ (the AR statistic). The idea is to put more weight on $S(\theta)$ when $\hat{D}(\theta)$ is close to reduced rank.

- Note that

$$\begin{aligned} K(\theta) &= n\hat{g}(\theta)'\hat{\Sigma}^{-1}(\theta)\hat{D}(\theta)\left[\hat{D}(\theta)'\hat{\Sigma}^{-1}(\theta)\hat{D}(\theta)\right]^{-1}\hat{D}(\theta)'\hat{\Sigma}^{-1}(\theta)\hat{g}(\theta) \\ &= n\hat{g}(\theta)'\hat{\Sigma}^{-1/2}(\theta)P_{\hat{\Sigma}^{-1/2}\hat{D}(\theta)}\hat{\Sigma}^{-1/2}(\theta)'\hat{g}(\theta) \end{aligned}$$

where $P_X = X(X'X)^{-1}X$ denotes the projection matrix for X. Thus,

$$J(\theta) \equiv S(\theta) - K(\theta) = n\hat{g}(\theta)'\hat{\Sigma}^{-1/2}(\theta)[I - P_{\hat{\Sigma}^{-1/2}\hat{D}(\theta)}]\hat{\Sigma}^{-1/2}(\theta)'\hat{g}(\theta)$$

(we follow the literature in using the notation $J(\theta)$ for this statistic; note, however, that this "J statistic" is different from the test of overidentifying restrictions given

by the minimum of the GMM objective function, which is also referred to as a "J statistic").

- Thus, $K(\theta)$ and $J(\theta)$ are quadratic forms of the asymptotically $N(0, I_{d_g})$ vector $\hat{\Sigma}^{-1/2}(\theta)\hat{g}(\theta)$ onto orthogonal subspaces with dimension d_θ and $d_g - d_\theta$. Using this and the fact that the matrix $\hat{D}(\theta)$ defining the subspaces is asymptotically independent of $\hat{g}(\theta)$, it can be shown that (even when $G(\theta)$ is singular)

$$\begin{aligned}&(K(\theta), J(\theta), \text{vec}(\sqrt{n}(\hat{D}(\theta) - D(\theta))') \\ &\xrightarrow{d} (\chi^2_{d_\theta}, \chi^2_{d_g - d_\theta}, \text{vec}(Z_D)) \quad \text{where } \chi^2_{d_\theta}, \chi^2_{d_g-d_\theta} \text{ and } Z_D \text{ are independent.}\end{aligned}$$

- Using this asymptotic distribution, one can form tests that are a function of both $K(\theta)$, $J(\theta)$ and $\hat{D}(\theta)$ (or, equivalently, of $K(\theta)$, $S(\theta)$ and $\hat{D}(\theta)$). For some function $b(J, K, D)$, we reject when $b(K(\theta), J(\theta), \hat{D}(\theta))$ is greater than $c_\alpha(\hat{D}(\theta))$, where $c_\alpha(D)$ is the $1-\alpha$ quantile of $b(\chi^2_{d_\theta}, \chi^2_{d_g-d_\theta}, D)$ (where $\chi^2_{d_\theta}$ and $\chi^2_{d_g-d_\theta}$ are the independent chi-square variables from the asymptotic distribution result above).

- Recall that we want to put more weight on $J(\theta)$ when $\hat{D}(\theta)$ is close to not being full rank. Let $r(D, \Sigma_D)$ be a scalar valued statistic that is small when D is close to reduced rank. The GMM-M (also called a quasi-conditional likelihood ratio or quasi-CLR) test statistic is given by

$$\frac{1}{2}\left[K(\theta) + J(\theta) - r(\hat{D}(\theta), \hat{\Sigma}_D(\theta)) \right. \\ \left. + \sqrt{\left[K(\theta) + J(\theta) + r(\hat{D}(\theta), \hat{\Sigma}_D(\theta))\right]^2 - 4J(\theta)r(\hat{D}(\theta), \hat{\Sigma}_D(\theta))}\right]$$

  where $\hat{\Sigma}_D(\theta)$ is an estimate of the variance of $\hat{D}(\theta)$.

- The GMM-M/quasi-CLR statistic was proposed by Kleibergen (2005) for nonlinear GMM. It is an extension of the conditional likelihood ratio (CLR) test of Moreira (2003). Other tests of this form have been proposed by Andrews (2016).

2.5 Inference when Some Parameters are Strongly Identified

- Suppose that $g(w_i, \theta)$ takes the form

$$g(w_i, z_i, \theta) = (\rho(w_i, \beta) - z_{1i}'\gamma)z_i \quad \text{where} \quad z_i = (z_{1i}', z_{2i}'),$$

so that z_{1i} are included instruments (that enter linearly) and z_{2i} are excluded instruments. This holds in the linear IV model

$$y_i = z_{1i}'\gamma + x_{2i}'\beta + \varepsilon_i, \quad E(\varepsilon_i|z_i) = 0$$

with $\rho(x_{2i}, y_i, \beta) = y_i - x_{2i}'\beta$.

- We can obtain a moment condition that does not depend on γ by projecting out z_{1i}. Let

$$\gamma(\beta) = (Ez_{1i}z_{1i}')^{-1}Ez_{1i}\rho(w_i, \beta) \text{ and } B = (Ez_{1i}z_{1i}')^{-1}Ez_{1i}z_{2i}'$$

denote the population projections of $\rho(w_i, \beta)$ and z_{2i} on z_{1i}. Then β satisfies the moment conditions

$$E(\rho(w_i, \beta) - z_{1i}'\gamma(\beta))(z_{2i} - B'z_{1i}) = 0$$

(note that the true value of γ is given by $\gamma_0 = \gamma(\beta_0)$ where β_0 is the true value of β).

- To get a feasible version of this moment condition, we simply replace $\gamma(\beta)$ with its sample analogue $\hat{\gamma}(\beta) = (\sum_{i=1}^n z_{1i}z_{1i}')^{-1}\sum_{i=1}^n z_{1i}\rho(w_i, \beta)$ and we replace B with its sample analogue $\hat{B} = (\sum_{i=1}^n z_{1i}z_{1i}')^{-1}\sum_{i=1}^n z_{1i}z_{2i}'$:

$$\tilde{g}(\beta) = \frac{1}{n}\sum_{i=1}^n(\rho(w_i, \beta) - z_{1i}'\hat{\gamma}(\beta))(z_{2i} - \hat{B}'z_{1i}) = \frac{1}{n}\sum_{i=1}^n \tilde{\rho}(w_i, \beta)\tilde{z}_i$$

where $\tilde{\rho}(w_i, \beta) = \rho(w_i, \beta) - z_{1i}'\hat{\gamma}(\beta)$ and $\tilde{z}_i = z_{2i} - \hat{B}'z_{1i}$ are residuals from the projections of $\rho(w_i, \beta)$ and z_{2i} on z_{1i}.

- With this form of the moment conditions, one can check that, under the true β, estimation error in $\hat{\gamma}(\beta)$ and $\hat{B}$ does not affect the asymptotic variance, and the same holds for the derivative estimate $\frac{d}{d\theta'}\tilde{g}(\beta)$. So, we can just take $\tilde{\rho}(w_i, \beta)\tilde{z}_i$ to be our

moment function and proceed as usual.

  - In the linear regression model where $\rho(x_{2i}, y_i, \beta) = y_i - x_{2i}'\beta$, this just corresponds to projecting out z_{1i} from y_i, x_{2i} and z_{2i}.

- This works in the linear regression model with a single endogenous variable, if we are interested in the coefficient of the endogenous variable. However, in most other cases, we do not get such a nice dichotomy between the parameter of interest and nuisance parameters that can be "projected out." In such cases, we face a problem of *subvector inference* (i.e. inference on the subvector of θ that we are interested in).

2.6 Subvector Inference

- Suppose that we are interested in some function $h(\theta)$, where $h : \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_h}$ is a function specified by the researcher. For example, if we are interested in the jth component of θ, we can specify $h(\theta) = \theta_j$.

- We can construct a CI for $h(\theta)$ from a confidence region $\mathcal{C}$ for θ: $\{h(\theta) : \theta \in \mathcal{C}\}$. However, this will be conservative if we base $\mathcal{C}$ on the statistics above, since they are designed for inference on the whole vector, and therefore "waste power" by rejecting values of θ that are not relevant for inference on $h(\theta)$. (Draw figure.)

- Chaudhuri and Zivot (2011) proposed a test that does not suffer from such inefficiencies. Here, we will present a version of this test that incorporates some extensions from Andrews (2017).

- To get some intuition for this test, note that, under strong identification, the efficient GMM estimator satisfies

$$\sqrt{n}(h(\hat{\theta}) - h(\theta)) = \sqrt{n}H(\theta)(G(\theta)'\Sigma(\theta)^{-1}G(\theta))^{-1}G(\theta)'\Sigma(\theta)^{-1}\hat{g}(\theta) + o_P(1)$$

  where $H(\theta) = \frac{d}{d\theta'}h(\theta)$ is the $d_h \times d_\theta$ derivative matrix of $h(\theta)$.

- In other words, the efficient linear combination of moments for inference on $h(\theta)$ is given by

$$H(\theta)(G(\theta)'\Sigma(\theta)^{-1}G(\theta))^{-1}G(\theta)'\Sigma(\theta)^{-1}\hat{g}(\theta).$$

Following the ideas behind the construction of the K statistic, we can use a sample analogue of these moments that uses $\hat{D}(\theta)$ to estimate $G(\theta)$. Let

$$M_h(\theta)' = H(\theta)(\hat{D}(\theta)'\hat{\Sigma}(\theta)^{-1}\hat{D}(\theta))^{-1}\hat{D}(\theta)'\Sigma(\theta)^{-1}.$$

A version of the K statistic that is efficient for projection inference for $h(\theta)$ is then given by a quadratic form in $M_h(\theta)'\hat{g}(\theta)$:

$$K_h(\theta) = n\hat{g}(\theta)M_h(\theta)(M_h(\theta)'\hat{\Sigma}(\theta)M_h(\theta))^{-1}M_h(\theta)'\hat{g}(\theta).$$

- Similar to the issues with the K statistic in Section 2.4, we will want to combine this test with a test based on $S(\theta)$ to ameliorate issues with power at certain types of alternatives. Chaudhuri and Zivot (2011) proposed the test

$$\text{reject } \theta \text{ if } K_h(\theta) > c_K \text{ or } S(\theta) > c_S$$

for some critical values c_S.

- It can be shown that

$$K_h(\theta) \xrightarrow{d} \chi^2_{d_h}, \quad S(\theta) \xrightarrow{d} \chi^2_{d_g}$$

where the asymptotic χ^2 random variables are independent. Thus, one can take

$$c_K = q_{\chi^2,d_h,1-\alpha+\gamma} \quad c_S = q_{\chi^2,d_g,1-\gamma}$$

where γ is some small number less than α - say, $\gamma = .01$ when $\alpha = .05$.

  - This is actually conservative since it doesn't take into account the dependence structure of $K_h(\theta)$ and $S(\theta)$. Indeed, it can be shown (Andrews, 2017) that

$$(K_h(\theta), S(\theta) - K_h(\theta)) \xrightarrow{d} (\chi^2_{d_h}, \chi^2_{d_g-d_h}).$$

    One can use this to decrease c_K and/or c_S.

- Under the usual (strong identification) asymptotics, the Chaudhuri-Zivot CI is asymptotically equivalent to the efficient CI $\{h(\hat{\theta}) \pm z_{1-\alpha/2}\text{se}_h\}$ if γ is taken to zero with the sample size.

2.7 Testing Lack of Identification

- A common approach in applied work is to first test the null of lack of identification, and then proceed using the usual asymptotic approximations so long as the test rejects.

- In the linear IV model with a single endogenous variable, the model is identified iff. at least one excluded instrument has a nonzero coefficient in the OLS regression of the endogenous variables on all exogenous variables (including the ones that are included in the structural equation). See, e.g., Sections 10.4 and 10.12 in Hansen (2017). Thus, a valid test of the null of lack of identification in this case can be formed by testing the null that all of these coefficients are zero. This can be done using an F-test, which, in this case, is called the *first stage F test*.

- In the linear IV model with multiple endogenous variables, identification holds iff. the matrix of reduced form coefficients is full rank (see Section 10.4 in Hansen, 2017). Tests for lack of identification in this setting were proposed by Cragg and Donald (1993).

- In the general GMM setup, Wright (2003) proposed a test for local identification.

- Staiger and Stock (1997) and Stock and Yogo (2002) consider tests for "weak identification" in the homoskedastic linear model. They model "weak identification" using sequences of dgps that change with the sample size, and they provide two possible definitions of "weak identification:"

  - Estimation: the model is "weakly identified" if the (asymptotic) bias of 2SLS is greater than 10% of the bias of OLS.
  - Inference: the model is "weakly identified" if the coverage of a nominal 95% CI is less than 90%.

- A popular rule of thumb is to find "weak identification" if the first stage F statistic is below 10. This corresponds to the estimation interpretation of weak IV: Stock and Yogo (2002) show that it corresponds to a test that the asymptotic bias of 2SLS is greater than 10% of the bias of OLS (at level .05, in the homoskedastic case; the critical value actually depends on the number of instruments and is slightly above 10 in most cases, but it is below 11.6 in all cases that they report).

  - If one is worried about bias in estimation with a single endogenous regressor and the sign of the first stage coefficients are known, one can use the estimator of

Andrews and Armstrong (2017), which is asymptotically unbiased under both weak and strong IV (using knowledge of the sign of the coefficients of the first stage).

  - For the inference interpretation, the critical value for the first stage F test depends to a greater extent on the number of instruments, and is typically larger (e.g. with four excluded instruments, one rejects when $F > 24.6$).
  - With multiple endogenous variables, Stock and Yogo (2002) report critical values for the Cragg and Donald (1993) test.

- In contrast to these tests, the confidence sets we have considered in these notes do not attempt to distinguish between cases where the model is identified or not: if the model is not identified, they will be large with high probability, thereby reflecting the inherent uncertainty in such models.

- If one uses that Stock and Yogo (2002) critical values based on the inference interpretation, one can interpret the procedure as forming a *two stage CI*:

  - Step 1: test the null of weak identification at level 5%.
  - Step 2: if the test rejects, report the usual CI. If not, report the entire real line as the CI.

  This CI has at least 90% coverage.

- This interpretation relies on the researcher reporting the entire real line in step 2 if the test in step 1 fails to reject. Another possibility is that, if the test in step 1 fails to reject, the paper will not be published. In this case, the coverage probability conditional on the paper being published can be arbitrarily small. Another possibility is that the researcher reports a weak instrument robust CI when the test in step 1 fails to reject, which leads to coverage at least 85%. See Chioda and Jansson (2005) and Andrews (2017) for more on these issues.

- Another concern is that the Stock and Yogo (2002) critical values only work for the homoskedastic linear model. For the estimation interpretation of weak IV, Montiel Olea and Pflueger (2013) consider the heteroskedastic linear model. However, a version of the Stock and Yogo (2002) approach for the inference interpretation of weak IV that works under heteroskedasticity does not appear to be available in the literature.

Extensions of Stock and Yogo (2002) to nonlinear models appear to be an open question as well.

3 Computing Projection CIs

- We have constructed confidence regions for the parameter θ. These take the form of a set $\mathcal{C} \subseteq \mathbb{R}^{d_\theta}$.

- In practice, we often want to construct a confidence region $\mathcal{C}_h$ for a one dimensional parameter $h(\theta)$ where $h : \mathbb{R}^{d_\theta} \to \mathbb{R}$ is a known function.

- As we discussed above, the projection confidence region

$$\mathcal{C}_h = \{h(\theta) : \theta \in \mathcal{C}\}$$

  is a valid confidence region for $h(\theta)$ so long as $\mathcal{C}$ is a valid confidence region for θ since, letting θ_0 be the true parameter value, the event $\theta_0 \in \mathcal{C}$ implies $h(\theta_0) \in \mathcal{C}_h$, so $P(h(\theta_0) \in \mathcal{C}_h) \geq P(\theta_0 \in \mathcal{C})$.

- To characerize this set, first note that the smallest interval containing this set is given by

$$[\underline{h}, \overline{h}] \quad \text{where} \quad \overline{h} = \sup\{h(\theta) : \theta \in \mathcal{C}\}, \quad \underline{h} = \inf\{h(\theta) : \theta \in \mathcal{C}\}. \tag{1}$$

  Computing $\overline{h}$ and $\underline{h}$ amounts to solving constrained optimization problems. We will discuss this further below.

- Once $\overline{h}$ and $\underline{h}$ are computed, one can take a grid of values $[h_1^*, \ldots, h_s^*]$ where $h_1^* = \underline{h}$ and $h_s^* = \overline{h}$ and check whether $h_j^* \in \mathcal{C}_h$ for each j. This can also be done by solving the feasibility problem

$$\inf_\theta 0 \quad \text{s.t.} \quad h(\theta) = h_j^*, \quad \theta \in \mathcal{C} \tag{2}$$

  where we take the solution to be ∞ if no value of θ satisfies the constraints. We then include h_j^* if the constraints are feasible (i.e. the value of the above problem is 0) and we do not include h_j^* if the constraints are not feasible (i.e. the value of the above problem is ∞).

- This gives a grid approximation to the confidence set $\mathcal{C}_h$. Alternatively, one can simply report $[\underline{h}, \overline{h}]$ as a confidence set for $\mathcal{C}_h$ (this will be simpler to report if $\mathcal{C}_h$ is disconnected).

- In the examples we have seen, the confidence set $\mathcal{C}$ takes the form

$$\mathcal{C} = \{\theta : T(\theta) \leq c(\theta)\}$$

for some test statistic $T(\theta)$ and critical value $c(\theta)$. Computing $\overline{h}$ in (1) then amounts to solving

$$\overline{h} = \sup h(\theta) \quad \text{s.t.} \quad T(\theta) \leq c(\theta)$$

and similarly for $\underline{h}$. For the feasibility problem (2), we can solve this, for example, by solving

$$\inf T(\theta) - c(\theta) \quad \text{s.t.} \quad h(\theta) = h_j^*$$

and checking whether the value of this problem is less than or equal to zero.

- If $T(\theta)$ and $c(\theta)$ are smooth and can be computed quickly (or, more generally, if the constraint $T(\theta) \leq c(\theta)$ can be expressed using a moderate number of smooth constraints), then we can solve these constrained optimization problems using a solver such as Knitro. If analytic formulas can be provided for the first and second derivatives of $T(\theta)$ and $c(\theta)$ (or, at least, a function to compute these derivatives that can be computed quickly), then this will further speed up optimization.

- We now discuss this optimization problem in the context of the confidence regions introduced so far. Note that, since we are constructing projection CIs for $h(\theta)$, the Chaudhuri-Zivot statistic discussed in Section 2.6 would be best suited to this purpose. For the Chaudhuri-Zivot test, $\overline{h}$ can be computed as

$$\sup_\theta h(\theta) \quad \text{s.t.} \quad K_h(\theta) \leq c_K, \quad S(\theta) \leq c_S$$

where $c_K = q_{\chi^2, d_h, 1-\alpha+\gamma}$ and $c_S = q_{\chi^2, d_g, 1-\gamma}$. However, for the purpose of completeness, we also discuss the other weak IV robust confidence regions we have considered so far.

- The Anderson-Rubin, Kleibergen, and Chaudhuri-Zivot test statistics are all smooth

functions of $\hat{g}(\theta)$, $\hat{G}(\theta)$, $\hat{\Sigma}(\theta)$ and $\hat{V}_{Gg}(\theta)$. The GMM-M statistic is also a smooth function of these quantities, so long as $r(\cdot)$ is smooth. Note that $\hat{g}(\theta)$ and $\hat{\Sigma}(\theta)$ are smooth functions of $g(w_i, \theta)$, and $\hat{G}(\theta)$ and $\hat{\Sigma}(\theta)$ are smooth functions of $g(w_i, \theta)$ and its first derivative. Thus, all of these statistics will be smooth as a function of θ so long as $g(w_i, \theta)$ is smooth in θ.

  - Ideally, one would provide the optimization routine with a function to compute the first and second derivatives of each of these test statistics. Analytic formulas for these derivatives can be obtained as functions of the first four derivatives of $g(w_i, \theta)$ by routine (but perhaps tedious) applications of matrix calculus.

- The critical values for all of these test statistics except for the GMM-M statistic are constant, so they can be computed once and then given to the optimization routine. However, the critical value for the GMM-M statistic depends on θ through $r(\hat{D}(\theta), \hat{\Sigma}_D(\theta))$: it is given by $\tilde{c}(r(\hat{D}(\theta), \hat{\Sigma}_D(\theta)))$ where $\tilde{c}(r)$ is the $1-\alpha$ quantile of

$$\frac{1}{2}\left[\chi^2_{d_\theta} + \sqrt{\left[\chi^2_{d_\theta} + \chi^2_{d_g - d_\theta} + r\right]^2 - 4\chi^2_{d_g - d_\theta} r}\right]$$

  where $\chi^2_{d_\theta}$ and $\chi^2_{d_g - d_\theta}$ are the independent chi-square variables.

  - One can compute approximations to $\tilde{c}(r)$ and its derivatives through simulation. This can be done at low computational cost, since it only needs to be done once, and since r is one-dimensional.

  - In Section 2.4, we also discussed more general statistics of the form $b(K, J, D)$, with critical values given by the $1-\alpha$ quantile of $b(\chi^2_{d_\theta}, \chi^2_{d_g - d_\theta}, \hat{D}(\theta))$. The GMM-M statistic takes this form, but depends on $\hat{D}(\theta)$ only through the scalar valued function $r(\cdot)$. This allows for a routine where critical values are computed beforehand for each value of r, and $\hat{D}(\theta)$ is simply plugged in to obtain the critical value for a given θ. However, if $b(K, J, D)$ depends on D through all $d_\theta \cdot d_g$ elements without any such simplification, this will not be feasible (since computing something on a grid over $d_\theta \cdot d_g$ dimensional space is not feasible for d_θ and d_g above, say, 3). For example, Andrews (2016) proposes tests of this form where the function $b(K, J, D)$ itself needs to be simulated for each value of D. In such cases, $c(\theta)$ will have to be computed by simulation for each value of θ, which will be computationally costly, and will lead to the critical value being nonsmooth as

a function of θ; this will also lead to difficulties in incorporating the test statistic and critical value into a constrained optimization routine.

  - In general, critical values that require simulation for each value of θ lead to difficulties in the constrained optimization problems (1) and (2). We will return to these issues after introducing confidence regions for moment inequalities.

- What do the sets $\mathcal{C}$ and $\mathcal{C}_h$ look like? Under what conditions are they connected, etc.? In the general nonlinear case, it is difficult to provide a general characterization. In the case where $g(w_i, \theta)$ is linear, Dufour and Taamouti (2005) and Mikusheva (2010) provide results for some of the tests we have considered here.

3.1 MPEC Approach

- In some cases the functions $\hat{g}(\theta)$ and $\hat{\Sigma}(\theta)$ take the form $\tilde{g}(\xi(\theta))$ and $\tilde{\Sigma}(\xi(\theta))$ where $\xi(\theta)$ is the solution to $s(\xi, \theta) = S$ for some known function s and some known S.

- Then we can state the optimization problem in the form of a mathematical program with equilibrium constraints (MPEC):

$$\overline{h} = \sup_{\theta} h(\theta) \quad \text{s.t.} \quad \tilde{T}(\xi) \leq \tilde{c}(\xi), \quad s(\xi, \theta) = S$$

  where $\tilde{T}$ and $\tilde{c}$ are the statistic and critical value written as a function of ξ.

- For each of the optimization problems discussed above (and below for moment inequalities), this simply amounts to replacing $\hat{g}(\theta)$ and $\hat{\Sigma}(\theta)$ with $\tilde{g}(\xi)$ and $\tilde{\Sigma}(\xi)$ and adding the constraint $s(\xi, \theta) = S$.

- This approach was applied to GMM estimation by Dubé et al. (2012) and Su and Judd (2012). I am not aware of any papers applying this approach to the settings considered in these notes (weak IV robust confidence sets and moment inequalities).

4 Moment Inequalities

- In *moment inequality* models, the true parameter value θ_0 satisfies

$$g(\theta_0) \geq 0 \tag{3}$$

where $g(\theta) = Eg(w_i, \theta)$ is a known function of data w_i and the parameter θ. As with the GMM setting, g is a function from $\mathbb{R}^{d_\theta}$ to $\mathbb{R}^{d_g}$. Here, we interpret inequality elementwise: $s \leq t$ for vectors $s, t \in \mathbb{R}^\ell$ iff. $s_i \leq t_i$ all i.

- Let Θ_0 denote the values of θ_0 such that (3) holds:

$$\Theta_0 = \{\theta_0 : g(\theta_0) \geq 0\}.$$

  If Θ_0 contains more than one value, then the model is said to be *set identified*.

  - For the GMM (moment equality) setting, we can also define the identified set Θ_0 as the set of values of theta such that $g(\theta_0) = 0$. In the GMM setting, we were worried about forming tests that are robust to lack of point identification, but we also considered how the tests behaved under "standard conditions" where θ was point identified.

- Often, moment inequality models are formed from *conditional moment inequality* models, in which the true parameter value θ_0 satisfies

$$E(m(w_i, \theta_0)|z_i = z) \geq 0 \quad \text{for all } z \tag{4}$$

  where z_i is a $\mathbb{R}^{d_z}$ valued random variable and $m(w_i, \theta)$ is a known function that takes values in $\mathbb{R}^m$. Let $f_1(z), \ldots f_p(z)$ be a set of *instrument functions* with $f_k(z) \geq 0$ for all z for each k. Then (4) implies

$$E(m_\ell(w_i, \theta_0) f_k(z_i)) \geq 0 \quad \text{for } k = 1, \ldots p,\ \ell = 1, \ldots, d_m \tag{5}$$

  which takes the form (3) with $g(w_i, z_i, \theta)$ given by the $(p \cdot d_m) \times 1$ vector with elements given by $m_\ell(w_i, \theta_0) f_k(z_i)$ for $k = 1, \ldots p$, $\ell = 1, \ldots, d_m$.

  - Note that, when z_i is continuously distributed, the identified set for (5) will, in general, be larger than the identified set for (3). That is, using instrument functions can lead to a loss of information. To ameliorate this, one typically takes the number of instrument functions p to increase as $n \to \infty$ (or one can take an infinite set of instrument functions).

– For example, one can use the instrument functions

$$f_k(z) = K((z - z_k)/h_k)$$

for some set of locations $z_1, \ldots, z_p \in \mathbb{R}^{d_z}$ and bandwidths $h_1, \ldots, h_p \in \mathbb{R}_+$ where $K(\cdot)$ is a nonnegative kernel. One can view this as an approximation to an infinite set of functions where $(z_k', h_k)'$ varies over the entire set $\mathbb{R}^{d_z} \times \mathbb{R}_+$, in which case the identified set for (5) will not be larger than the identified set for (3). Alternatively one can take these points to "fill in" the support of $\mathbb{R}^{d_z} \times \mathbb{R}_+$ as $n \to \infty$.

– On the other hand, when z_i takes on finitely many values $\{\tilde{z}_1, \ldots, \tilde{z}_p\}$, one can simply take $f_k(z) = I(z = \tilde{z}_k)$, and the identified set for (5) will be the same as the identified set for (3).

4.1 Example

- Interval regression The variables z_i and y_i^* follow the linear regression model

$$E(y_i^*|z_i) = z_i'\theta$$

We observe $\{(z_i', y_i^L, y_i^H)\}_{i=1}^n$ where $y_i^* \in [y_i^L, y_i^H]$, but we do not observe y_i^*. This leads to the inequalities

$$E(y_i^L|z_i = z) \leq z'\theta \leq E(y_i^H|z_i = z) \quad \text{all } z$$

which takes the form (5) with

$$m(z_i, y_i^L, y_i^H, \theta) = \begin{pmatrix} y_i^H - z_i'\theta \\ z_i'\theta - y_i^L \end{pmatrix}.$$

4.2 Confidence Regions for Moment Inequality Models

- As with the weak instrument setting, the moment inequalities literature has, for the most part, focused on confidence regions of the form $\mathcal{C} = \{\theta : T(\theta) \leq c(\theta)\}$ for some test statistic $T(\theta)$ and critical value $c(\theta)$.

- In practice, this means that one will need to compute the projection confidence region $\mathcal{C}_h = \{h(\theta) : T(\theta) \leq c(\theta)\}$ as discussed in Section 3.

- We will introduce some tests from the literature, with a focus on those tests that lead to computationally tractable optimization problems (1) and (2) for computing $\mathcal{C}_h$. We use the same notation for sample moments and the sample variance matrix as for GMM:

$$\hat{g}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(w_i, \theta)$$

and

$$\hat{\Sigma}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(w_i, \theta)g(w_i, \theta)' - \hat{g}(\theta)\hat{g}(\theta)'.$$

The idea behind all of these tests is to use a statistic $T(\theta) = S(\sqrt{n}\hat{g}(\theta), \hat{\Sigma}(\theta))$ that is large when one or more of the elements of $\hat{g}(\theta)$ is negative.

4.3 Max Statistic

- Consider the statistic

$$T_{\max}(\theta) = \max_{j:1\leq j\leq d_g} \frac{-\sqrt{n}\hat{g}_j(\theta)}{\sqrt{\hat{\Sigma}_{jj}(\theta)}}.$$

where $\hat{\Sigma}_{jj}(\theta)$ denotes the jjth element of $\hat{\Sigma}(\theta)$.

- The max statistic is approximately distributed as $\max_{j:1\leq j\leq d_g} Z_j - g_j(\theta)/\sqrt{\Sigma_{jj}(\theta)}$ where $Z_1, \ldots, Z_{d_g}$ are mean zero normal variables with variance one and covariance determined by $\Sigma(\theta)$. Since $g_j(\theta) \geq 0$ under the null, a conservative approximation to this distribution is to assume that $g_j(\theta) = 0$ all j, thereby using the distribution of $\max_{j:1\leq j\leq d_g} Z_j$ to compute the critical value. This is sometimes called a *least favorable null distribution* for this test statistic.

- Unfortunately, the distribution of $\max_{j:1\leq j\leq d_g} Z_j$ depends on θ through $\Sigma(\theta)$, which determines the covariance of the Z_js. A conservative approximation can be obtained using *Bonferroni's inequality*:

$$P\left(\max_{j:1\leq j\leq d_g} Z_j > t\right) = P(\cup_{j=1}^{d_g}\{Z_j > t\}) \leq \sum_{j=1}^{d_g} P(Z_j > t) = d_g[1 - \Phi(t)]$$

where $\Phi(t)$ is the standard normal cdf. Setting this equal to α and solving for t gives the Bonferroni critical value

$$c_{\text{Bonf},\alpha} = z_{1-\alpha/d_g}$$

where $z_t = \Phi^{-1}(t)$ is the tth quantile of the $N(0,1)$ distribution.

- The Bonferroni critical value is conservative, but it turns out that it is not that conservative when the correlation between most of the Z_j's is small.

- The Bonferroni critical value has been considered by Chernozhukov et al. (2014) in the context of moment inequalities (along with other critical values) and Fan et al. (2007) in the case where d_g increases with the sample size. This approach has a long history in the multiple testing literature; see Lehmann and Romano (2005), Chapter 9.

4.3.1 Multiscale Statistic for Conditional Moment Inequalities

- If the moment inequalities have more structure, one can derive an asymptotically non-conservative critical value in certain cases. We now discuss an approach for the conditional moment inequality model (5) based on Armstrong and Chan (2016). The following discussion incorporates some suggestions for tuning parameters and implementation details.

- First, to make sure that the procedure is not affected by the scale of the elements of z_i, we can transform each element of z_i by its empirical cdf. That is, for $j = 1, \dots d_z$, let $\hat{F}_j(t) = \frac{1}{n}\sum_{i=1}^n I(z_{i,j} \leq t)$, and redefine $z_{i,j}$ to be $\hat{F}_j(z_{i,j})$. This gives n observations (z_i', w_i') where z_i is supported on $[0,1]^{d_z}$.

- Let $\{(s_k', t_k')\}_{k=1}^p$ be elements in the set

$$\{(s', t')' : s, s+t \in [0,1]^{d_z}, t \in [\underline{t}_n, 1]^{d_X}\}. \tag{6}$$

  where $\underline{t}_n$ is a sequence of scalars, chosen by the researcher, which goes to zero slightly more slowly than n^{-1/d_z}.

- The *multiscale statistic* is the max statistic for the moment inequalities using the

functions $f_k(z) = I(s_k \leq z \leq s_k + t_k)$ as instruments

$$T_{\text{multiscale}}(\theta) = \max_{k,\ell:1\leq k\leq p, 1\leq \ell\leq d_m} \frac{-\sqrt{n}\tilde{g}_{k\ell}(\theta)}{\tilde{\sigma}_{k\ell}(\theta)}$$

where

$$\tilde{g}_{k\ell}(\theta) = \frac{1}{n}\sum_{i=1}^{n} m_\ell(w_i, \theta) I(s_k \leq z_i \leq s_k + t_k)$$
$$\tilde{\sigma}_{k\ell}(\theta)^2 = \frac{1}{n}\sum_{i=1}^{n} [m_\ell(w_i, \theta) I(s_k \leq z_i \leq s_k + t_k) - \tilde{g}_{k\ell}(\theta)]^2.$$

- Based on the asymptotic distribution in the least favorable case (where $Em_\ell(w_i, \theta)I(s_k \leq z_i \leq s_k + t_k) = 0$ all ℓ, k), Armstrong and Chan (2016) propose the critical value

$$c_{\text{multiscale},\alpha} = \frac{\log d_m - \log(-\log(1-\alpha)) + 2\log \underline{t}_n^{-d_z} + (2d_z - 1/2)\log\log \underline{t}_n^{-d_z} - \log(2\sqrt{\pi})}{\sqrt{2\log \underline{t}_n^{-d_z}}}$$

  for a $1-\alpha$ confidence region.

- Additional implementation details:

  - After transforming by the empirical cdf, the z_is are distributed on $[0,1]^{d_z}$ with approximate uniform marginal distributions. If the joint pdf were uniform, then we would have approximately $n \cdot \underline{t}^{d_z}$ elements in the smallest "box" with $t = \underline{t}^{d_z}$. A reasonable choice that satisfies the requirements of Armstrong and Chan (2016) is

$$\underline{t}_n = \left[\frac{(\log n)^5}{n} \cdot \frac{100}{(\log 1000)^5}\right]^{1/d_z}.$$

    This sets the smallest box to contain approximately 100 observations when $n = 1000$.

  - The results in Armstrong and Chan (2016) actually allow the maximum over (s'_k, t'_k) to be replaced with a supremum over the entire set given in (6), so here we are limited only by computational resources. One possibility is to choose $\{(s'_k, t'_k)\}_{k=1}^p$ to be the discrete approximation to this set in which each element

of s and t is an integer multiple of $(\log n)^{-1}\underline{t}_n$:

$$\{(s', t')' : s, s + t \in [0, 1]^{d_z}, t \in [\underline{t}_n, 1]^{d_X}\} \cap [(\log n)^{-1}\underline{t}_n\mathbb{Z}]^{2d_z}.$$

4.3.2 Computing $\mathcal{C}_h$

- The upper endpoint of $\mathcal{C}_h$ can be found by solving the optimization problem

$$\overline{h} = \sup_\theta h(\theta) \quad \text{s.t.} \quad \hat{g}_j(\theta) \geq -c_{\text{Bonf},\alpha} \cdot \sqrt{\hat{\Sigma}_{jj}(\theta)}/\sqrt{n} \quad j = 1, \ldots, d_g$$

  for the max statistic with Bonferroni critical values or

$$\overline{h} = \sup_\theta h(\theta) \quad \text{s.t.} \quad \tilde{g}_{k\ell}(\theta) \geq -c_{\text{multiscale},\alpha} \cdot \hat{\sigma}_{k\ell}(\theta)/\sqrt{n} \quad k = 1, \ldots, p,\ \ell = 1, \ldots, d_m$$

  for the multiscale statistic (this is the same optimization problem with $m_\ell(w_i, \theta)I(s_k \leq z_k \leq s_k + t_k)$ subtituted for $g_j(w_i, \theta)$ and the indices k, ℓ used instead of j).

  - If $g(w_i, \theta)$ is smooth in θ (which, in the case of the multiscale statistic, amounts to $m(w_i, \theta)$ being smooth in θ), then the constraints will be smooth in θ, so this amounts to optimizing a smooth function subject to d_g (which is equal to $d_m \cdot p$ in the case of the multiscale statistic) smooth constraints.

4.3.3 Other Max Statistics

- Test statistics of a similar form to the multiscale statistic treated here are considered by Chetverikov (2017) and Dumbgen and Spokoiny (2001). The multiscale statistic uses kernel functions with multiple bandwidths as instruments (hence the name "multiscale"). A max statistic based on a single bandwidth was considered by Chernozhukov et al. (2013); in one version of their test, they use an analytic critical value, which leads to $\overline{h}$ being computable as a solution to an optimization problem that takes a similar form to the one described above for the multiscale statistic.

4.4 Rosen's Approach to Moment Inequalities

- Rosen (2008) considered confidence regions for the moment inequality problem (3)

based on the *quasi-likelihood ratio* statistic

$$T_{QLR}(\theta) = n \min_{t \geq 0} [\hat{g}(\theta) - t]' \hat{\Sigma}(\theta)^{-1} [\hat{g}(\theta) - t].$$

- The conservative version of this confidence region uses the critical value $c^*_{d_g}$, which solves

$$P(\chi^2_{d_g} > c^*_{d_g})/2 + P(\chi^2_{d_g - 1} > c^*_{d_g})/2 = \alpha.$$

- The upper endpoint of the confidence region $\mathcal{C}_h = \{h(\theta) : T_{QLR}(\theta) \leq c^*_{d_g}\}$ can then be computed as

$$\overline{h} = \sup_{\theta} h(\theta) \quad \text{s.t.} \quad n \min_{t \geq 0} [\hat{g}(\theta) - t]' \hat{\Sigma}(\theta)^{-1} [\hat{g}(\theta) - t] \leq c^*_{d_g}.$$

- Note that θ satisfies these constraints iff. there exists a $t \geq 0$ such that $n[\hat{g}(\theta) - t]' \hat{\Sigma}(\theta)^{-1} [\hat{g}(\theta) - t] \leq c^*_{d_g}$. Thus, we can phrase this as a single optimization problem in θ, t:

$$\overline{h} = \sup_{\theta, t} h(\theta) \quad \text{s.t.} \quad n[\hat{g}(\theta) - t]' \hat{\Sigma}(\theta)^{-1} [\hat{g}(\theta) - t] \leq c^*_{d_g}, \quad t \geq 0.$$

- Similar comments apply to the computation of $\underline{h}$, and to checking whether $h^*_j \in \mathcal{C}_h$.

- The quasi-likelihood ratio test is based on the likelihood ratio test for moment inequalities in the finite sample normal model with known covariance. See Rosen (2008) for references.

4.5 Relative Efficiency in Moment Inequality Models

- The theory of relative efficiency is different for moment inequalities than for settings such as GMM. For conditional moment inequalities, Armstrong (2014c) derives the rate at which the confidence region shrinks toward Θ_0 for tests proposed in the literature. It follows from these results and results in Armstrong (2014a) that the multiscale test described in Section 4.3.1 leads to a confidence set with a strictly better rate of convergence than other tests proposed in the literature.

  - Unlike GMM and other "regular" settings, the rate of convergence is slower than $n^{1/2}$, and depends on the "smoothness" of certain conditional means. This leads to relative efficiency results that are more closely related to those in the literature on nonparametric estimation and inference, such as results on optimal kernels and bandwidths (see, e.g., Ichimura and Todd, 2007; Tsybakov, 2009). Indeed, the theory of relative efficiency here is closely related to the literature on nonparametric testing (see Ingster and Suslina, 2003).

- See Armstrong (2014b), Chernozhukov et al. (2014), Chetverikov (2017) and references therein for more on optimality properties of the max statistic (including in the general unconditional case where the moments do not have the multiscale structure).

- Based on these results, and on computational feasibility, I would recommend, as a default approach ...

  - ... using the max statistic with Bonferroni critical value for (unconditional) moment inequality models of the form (3)
  - ... using the multiscale statistic and critical value in Section 4.3.1 for conditional moment inequality models of the form (4)

4.6 Moment Selection and Simulated Critical Values

- The confidence regions $\mathcal{C}$ given above are conservative due to (1) the critical values being based on a "least favorable" distribution, where $g_j(\theta) = 0$ all j and (2) not taking into account the dependence structure in the covariance matrix $\Sigma(\theta)$.

- A solution to (1) is to use *moment selection*. The idea here is to use pre-tests to find indices j for which $g_j(\theta)$ is "too far inside of the null set" to affect the sampling distribution. Such procedures have been considered by Hansen (2005) and others.

  - As an example of moment selection applied to the quasi-likelihood ratio test, Rosen (2008) proposes a version of his test that uses $\hat{b}(\theta)$ in place of d_g to compute the degrees of freedom, where $\hat{b}(\theta) = \sum_{j=1}^{d_g} I(\hat{g}_j(\theta) \leq c\sqrt{(\log n)/n})$ for some constant c.
  - For the max test with Bonferroni critical values, a moment selection procedure of this form would amount to replacing the critical value $z_{1-\alpha/d_g}$ with $z_{1-\alpha/b(\theta)}$.

- Since $\hat{b}(\theta)$ is discontinuous as a function of θ, this will lead to a discontinuous critical value as a function of θ in both cases. However, it may be possible to obtain a smoothed critical value using a smoothed version of this procedure (for example, by replacing the indicator function in the definition of $\hat{b}(\theta)$ with a smooth function).
- For the max statistic, the critical value $z_{1-\alpha/d_g}$ behaves like $\sqrt{2\log d_g}$ when d_g is large. Since this increases slowly with d_g, using moment selection to replace d_g with some $b(\theta) < d_g$ will not to large improvements in the critical value unless $b(\theta)$ is much smaller than d_g.

- Regarding (2) (being conservative about the dependence structure of $\Sigma(\theta)$), consider the max statistic for concreteness. A nonconservative critical value would be the $1-\alpha$ quantile of $\max_{j:1\leq j\leq d_g} Z_j$ where $Z = (Z_1, \ldots, Z_{d_g})'$ follows a normal distribution with variance $[\text{diag}(\Sigma(\theta))]^{-1/2}\Sigma(\theta)[\text{diag}(\Sigma(\theta))]^{-1/2}$. In particular, let

$$J(t,\Sigma) = \int_{z_1=-\infty}^{t} \cdots \int_{z_{d_g}=-\infty}^{t} \phi(z; [\text{diag}(\Sigma)]^{-1/2}\Sigma[\text{diag}(\Sigma)]^{-1/2})\, dz_1 \cdots dz_{d_g}$$

  where $\phi(z;\Omega)$ denotes the pdf of a $N(0,\Omega)$ random variable and $\text{diag}(\Sigma)$ is the matrix with the same diagonal elements as Σ and zeros for all nondiagonal elements. A nonconservative critical value is given by $J^{-1}(1-\alpha, \hat{\Sigma}(\theta))$ where J^{-1} denotes the inverse in the first argument.

  - The high dimensional integral involved in computing $J(t,\Sigma)$ makes the critical value difficult so compute. One can, however, approximate it using simulation or resampling: let $T_1^*(\theta), \ldots, T_B^*(\theta)$ denote draws of the max statistic based on simulated or resampled data, and let $c(\theta)$ denote the $1-\alpha$ empirical quantile of these simulated test statistics. The confidence region is then given by $\mathcal{C} = \{\theta : T(\theta) \leq c(\theta)\}$. Unfortunately, $c(\theta)$ will not be smooth even if $T_1^*(\theta), \ldots, T_B^*(\theta)$ are smooth, so one does not immediately obtain a smooth optimzation problem for computing $\overline{h}$.
  - In general, simulated critical values can be nonsmooth in θ, which can cause problems for computing confidence sets. We saw this issue in the context of weak IV when we discussed computation of confidence sets based on statistics of the form $b(K,J,D)$ that generalize the GMM-M statistic.

– While the simulated critical value described above is not smooth as a function of θ, the critical value that it tries to approximate, $J^{-1}(1-\alpha, \hat{\Sigma}(\theta))$, is a smooth function of θ. If one were to use an approximation to $J^{-1}(1-\alpha, \Sigma)$ that is smooth in Σ, then this could be used to obtain a smooth optimization problem.

- As discussed above, the Bonferroni critical values for the max statistic are actually not that conservative unless $\Sigma(\theta)$ is highly dependent, and moment selection will typically not lead to large changes in the critical value for this statistic. Because of this, and since moment selection and simulated critical values can lead to computational difficulties, I would recommend simply using the Bonferroni critical value (or, for the multiscale statistic, the critical value $c_{\text{multiscale},\alpha}$ described in Section 4.3.1) as described in Section 4.3, without moment selection.

4.7 Confidence Regions for the Identified Set

- The confidence regions we have considered so far have the property

$$P(\theta_0 \in \mathcal{C}) \geq 1-\alpha \quad \text{all } \theta_0 \in \Theta_0 \tag{7}$$

(or, at least, they satisfy this property asymptotically, in the sense that it holds for some sequence $\alpha_n \to \alpha$). This property is sometimes called *coverage of points in the identified set*.

- One may instead ask for the stronger property

$$P(\Theta_0 \subseteq \mathcal{C}) \geq 1-\alpha \tag{8}$$

(or an asymptotic version of this property, in which the above display holds for some sequence $\alpha_n \to \alpha$). This property was considered by Chernozhukov et al. (2007) and Romano and Shaikh (2010), and is sometimes called *coverage of the identified set*.

- The literature seems to have converged on the criterion (7). A rationale for this is that it is the same requirement used in other settings: if there is some "true" θ_0 that generated the data, we want a CI that contains it with probability at least $1-\alpha$, whether or not the confidence region contains other points in the identified set. See Imbens and Manski (2004) for further discussion.

- Nevertheless, let us briefly discuss some methods for attaining the criterion (8) for covering the identified set. As before, we will focus on computing the confidence set $\mathcal{C}_h = \{h(\theta) : \theta \in \mathcal{C}\}$ for $h(\theta)$ based on a confidence set $\mathcal{C}$ for θ.

- Romano and Shaikh (2010) noted that, whereas inverting a family of level α tests of $H_{0,\theta_0} : \theta_0 \in \Theta_0$ gives a CI that satisfies (7), one can obtain a CI that satisfies (8) by inverting a family of tests that controls the *familywise error rate (FWER)*. If the test rejects when $T(\theta) > c$, this means that

$$P\left(\text{there exists } \theta_0 \in \Theta_0 \text{ with } T(\theta) > c\right) \leq \alpha. \tag{9}$$

  The confidence region $\mathcal{C} = \{\theta : T(\theta) \leq c\}$ then satisfies (8).

- To satisfy this criterion, we need c to approximate the $1-\alpha$ quantile of $\sup_{\theta \in \Theta_0} T(\theta)$. Romano and Shaikh (2010) use a *step-down* procedure, adapted from the multiple testing literature.

- Let us describe a version of this procedure based on *subsampling*, proposed in Romano and Shaikh (2010). Let $T(\theta) = S(\sqrt{n}\hat{g}(\theta), \hat{\Sigma}(\theta))$ denote a test statistic such as the max statistic or quasi-likelihood ratio statistic. To compute the critical value, we draw $B = B_n$ random subsets $\mathcal{I}_1, \ldots, \mathcal{I}_B$ of the indices $\{1, \ldots, n\}$ (without replacement), with each of the subsets having $b = b_n$ elements, where $b_n \to \infty$ and $b_n/n \to 0$. Let

$$\hat{g}_j^*(\theta) = \frac{1}{b} \sum_{i \in \mathcal{I}_j} g(w_i, \theta)$$

  denote the sample moments computed with the jth subsample, and similarly for $\Sigma_j^*(\theta)$. Let

$$T_j^*(\theta) = S(\sqrt{b}\hat{g}_j^*(\theta), \hat{\Sigma}_j^*(\theta))$$

  denote the test statistic computed with the jth subsample (note that we scale by b, the subsample size, rather than n, the sample size).

  The step-down procedure is as follows. Let Θ be the parameter space for θ (we can take it to be some large set known to contain θ_0).

– Step 0: For each $j = 1, \ldots, B$, solve

$$\sup_{\theta} T_j^*(\theta) \quad \text{s.t.} \quad \theta \in \Theta.$$

Let c_0 be the $1-\alpha$ empirical quantile of $\sup_{\theta \in \Theta} T_j^*(\theta)$ over $j = 1, \ldots B$.

– Step k: For each $j = 1, \ldots, B$, solve

$$\sup_{\theta} T_j^*(\theta) \quad \text{s.t.} \quad T(\theta) \leq c_{k-1}$$

Let c_k be the $1-\alpha$ empirical quantile of $\sup_{\theta: T(\theta) \leq c_{k-1}} T_j^*(\theta)$ over $j = 1, \ldots B$.

– Final step: Choose a stopping rule and let $c_{k_{\text{final}}}$ be the critical value after iterating this procedure until stopping. To compute the projection CI for $h(\theta)$, solve

$$\sup_{\theta} h(\theta) \quad \text{s.t.} \quad T(\theta) \leq c_{k_{\text{final}}}.$$

Report $[\underline{h}, \overline{h}]$ where $\overline{h}$ is the value of the above maximization problem and $\underline{h}$ is the value of the analogous minimization problem.

- The step-down procedure involves $B \cdot (k_{\text{final}} + 1) + 1$ optimization problems.
- The confidence set for Θ_0 is given by $\mathcal{C} = \{\theta : T(\theta) \leq c_{k_{\text{final}}}\}$. The last step given above describes the procedure for getting the smallest interval $[\underline{h}, \overline{h}]$ that contains the projection confidence set $\mathcal{C}_h = \{h(\theta) : \theta \in \mathcal{C}\}$ for $h(\theta)$.

5 Issues with Interpretation of Confidence Sets

- In the settings we have considered here, one must be careful about interpreting the confidence set $\mathcal{C}$ or $\mathcal{C}_h$ as "summarizing uncertainty about the true value of θ."
- To understand these issues, let us contrast the confidence sets considered here with those that arise in "regular" settings. Consider the confidence set

$$\hat{\theta} \pm z_{1-\alpha/2} \text{se}(\hat{\theta}). \tag{10}$$

Suppose that $\hat{\theta}$ has an exact normal distribution, and $\text{se}(\hat{\theta})$ is the exact (known) standard deviation of $\hat{\theta}$.

  - Let $\chi = z_{1-\alpha/2}\text{se}(\hat{\theta})$. Then the fact that (10) constitutes a $1-\alpha$ confidence set and the fact that χ is fixed means that $\hat{\theta}$ has risk bounded by α when we use the zero-one loss function $L(\hat{\theta},\theta) = I(|\hat{\theta}-\theta| > \chi)$:

$$EI(|\hat{\theta}-\theta| > \chi) \leq \alpha.$$

  - Thus, the length of the confidence interval (10) gives us a bound on the worst-case (i.e. minimax) risk of the estimator $\hat{\theta}$ (or, more precisely, it tells us a loss function for which we can bound the risk). In this sense, we get a statement that does tell us something about our uncertainty about the true parameter value: it gives us an upper bound on risk of a particular estimator.

- This nice relationship between estimation risk and CI length happens because the length of the CI (10) is fixed. In contrast to this *fixed length* CI, the CIs we have considered here for weak IV and moment inequality settings are *random length* CIs.
  - While weak IV and moment inequality settings do not lead to fixed length CIs, fixed length CIs turn out to be close to optimal in some other "irregular" settings. See Armstrong and Kolesár (2016).

- In general, the ex-post length of a random length CI does not necessarily tell us anything useful about the risk of any estimator or procedure. However, a CI inverts tests that control type I error, so this tells us something about the risk of these tests (namely, that type I risk is less than α). One approach to summarizing uncertainty would be to try to summarize power function of these tests (i.e. the type II risk). This is the subject of a *statistical power analysis*: one tries to determine whether the power of a test is likely to be good based on a priori considerations, before computing the test.

- Another approach, based on asking whether one could profit by betting that the parameter is not in the CI after observing the CI, has been applied recently by Müller and Norets (2012) to weak IV, moment inequalities and other problems in econometrics.

References

Anderson, T. W. and H. Rubin (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *The Annals of Mathematical Statistics*, 20, 46–63.

Andrews, I. (2016): "Conditional Linear Combination Tests for Weakly Identified Models," *Econometrica*, 84, 2155–2182.

——— (2017): "Valid Two-Step Identification-Robust Confidence Sets for GMM," *The Review of Economics and Statistics*.

Andrews, I. and T. B. Armstrong (2017): "Unbiased instrumental variables estimation under known first-stage sign," *Quantitative Economics*, 8, 479–503.

Armstrong, T. (2014a): "On the Choice of Test Statistic for Conditional Moment Inequality Models," *Unpublished Manuscript*.

Armstrong, T. B. (2014b): "A Note on Minimax Testing and Confidence Intervals in Moment Inequality Models," .

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

Armstrong, T. B. and H. P. Chan (2016): "Multiscale adaptive inference on conditional moment inequalities," *Journal of Econometrics*, 194, 24–43.

Armstrong, T. B. and M. Kolesár (2016): "Optimal inference in a class of regression models," *arXiv:1511.06028 [math, stat]*.

Chaudhuri, S. and E. Zivot (2011): "A new method of projection-based inference in GMM with weakly identified nuisance parameters," *Journal of Econometrics*, 164, 239–251.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014): "Testing many moment inequalities," *arXiv:1312.7614 [math, stat]*, arXiv: 1312.7614.

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

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

Chetverikov, D. (2017): "Adaptive Tests of Conditional Moment Inequalities," *Econometric Theory*, 1–42.

Chioda, L. and M. Jansson (2005): "Optimal Conditional Inference for Instrumental Variables Regression," *Unpublished Manuscript*.

Cragg, J. G. and S. G. Donald (1993): "Testing Identifiability and Specification in Instrumental Variable Models," *Econometric Theory*, 9, 222–240.

Dubé, J.-P., J. T. Fox, and C.-L. Su (2012): "Improving the Numerical Performance of Static and Dynamic Aggregate Discrete Choice Random Coefficients Demand Estimation," *Econometrica*, 80, 2231–2267.

Dufour, J.-M. and M. Taamouti (2005): "Projection-Based Statistical Inference in Linear Structural Models with Possibly Weak Instruments," *Econometrica*, 73, 1351–1365.

Dumbgen, L. and V. G. Spokoiny (2001): "Multiscale Testing of Qualitative Hypotheses," *The Annals of Statistics*, 29, 124–152.

Fan, J., P. Hall, and Q. Yao (2007): "To How Many Simultaneous Hypothesis Tests Can Normal, Student's t or Bootstrap Calibration Be Applied?" *Journal of the American Statistical Association*, 102, 1282–1288.

Hansen, B. E. (2017): *Econometrics*, Online manuscript available at http://www.ssc.wisc.edu/~bhansen/econometrics/.

Hansen, P. R. (2005): "A Test for Superior Predictive Ability," *Journal of Business & Economic Statistics*, 23, 365–380.

Ichimura, H. and P. E. Todd (2007): "Chapter 74 Implementing Nonparametric and Semiparametric Estimators," Elsevier, vol. Volume 6, Part 2, 5369–5468.

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

Ingster, Y. and I. A. Suslina (2003): *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, Springer.

Kleibergen, F. (2005): "Testing Parameters in GMM without Assuming That They Are Identified," *Econometrica*, 73, 1103–1123.

Lehmann, E. L. and J. P. Romano (2005): *Testing statistical hypotheses*, Springer.

Mikusheva, A. (2010): "Robust confidence sets in the presence of weak instruments," *Journal of Econometrics*, 157, 236–247.

Montiel Olea, J. L. and C. Pflueger (2013): "A Robust Test for Weak Instruments," *Journal of Business & Economic Statistics*, 31, 358–369.

Moreira, M. J. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71, 1027–1048.

Müller, U. K. and A. Norets (2012): "Credibility of Confidence Sets in Nonstandard Econometric Problems," .

Romano, J. P. and A. M. Shaikh (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78, 169–211.

Rosen, A. M. (2008): "Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities," *Journal of Econometrics*, 146, 107–117.

Staiger, D. and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Stock, J. H. and J. H. Wright (2000): "GMM with Weak Identification," *Econometrica*, 68, 1055–1096.

Stock, J. H. and M. Yogo (2002): "Testing for Weak Instruments in Linear IV Regression," *National Bureau of Economic Research Technical Working Paper Series*, No. 284.

Su, C.-L. and K. L. Judd (2012): "Constrained Optimization Approaches to Estimation of Structural Models," *Econometrica*, 80, 2213–2230.

Tsybakov, A. B. (2009): *Introduction to Nonparametric Estimation*, New York: Springer.

Wright, J. H. (2003): "Detecting Lack of Identification in Gmm," *Econometric Theory*, 19, 322–330.