Inference on functions of parameters partially identified by the intersection of continuous linear inequalities

Zach Flynn*

University of Wisconsin — Madison

April 12, 2017

Abstract

I develop a consistent estimator of bounds on functions of parameters partially identified by the intersection of continuous linear inequalities. Aside from allowing for continuous constraints, an advantage of the estimator is that it can be used to compute a closed form confidence interval, without numerically inverting a hypothesis test. So it is easy to compute confidence intervals even if the number of parameters is very large, especially when we are interested in a linear function of parameters. I also consider the dual problem of bounding a linear function of a sequence, an infinite dimensional parameter, partially identified by finitely many linear restrictions on the sequence.

Keywords. Bounds estimation, consistent estimator of the value of semi-infinite programs

*E-mail: zflynn@wisc.edu. I thank Jack Porter, Xiaoxia Shi, Joachim Freyberger, Amit Gandhi, Alan Sorensen, and Ken Hendricks for their great feedback on the paper.

1. Introduction.

I estimate bounds on functions of parameters identified by the intersection of continuous linear-in-parameters inequalities where each coefficient in the inequalities is a smooth function of moments. I allow the sample moments to be computed using arbitrarily few observations.

Let $\mathscr{U} = [0,1]^P \cap \mathbb{Q}^P$ (so $\mathscr{U}$ is countable), $\Theta \subset \mathbb{R}^b$ be the parameter space, $\left\{h_j\right\}_{j=1}^J$ be a collection of random variables with bounded variance with $J < \infty$, and $\left\{x_j\right\}_{j=1}^J$ be a collection of random P-vectors each with support on $[0,1]^P$. The identified set is Θ_0,

$$\begin{aligned} v_j(u) &= \mathbb{E}\left[h_j \mathbf{1}\left(x_j \geq u\right)\right] \\ \Theta_0 &= \left\{\theta \in \Theta : d_\ell(v(u))^\top \theta \leq p_\ell(v(u)) \text{ for } u \in \mathscr{U}, \ell = 1, \ldots, L < \infty\right\}, \end{aligned}$$

Where $d_\ell(\cdot)$ and $p_\ell(\cdot)$ are continuously differentiable functions. I want to estimate bounds on a function, $t(\theta, \gamma)$, of θ and a vector of point-identified parameters, γ.

The problem is related to the literature on identified sets defined by a finite number of convex moment inequalities, see Kaido and Santos (2014), Kaido (2016), and Beresteanu and Molinari (2008). It is also related to the literature on inference on functions of partially identified parameters. Kaido, Molinari, and Stoye (2016) and Bugni, Canay, and Shi (2016) study inference on functions of parameters identified by a finite number of moment inequalities.

There are two main applications: to conditional restrictions (monotone instrumental variables as in Manski and Pepper 2000 and conditional moment inequalities) and to assumptions that hold for a certain class of functions (like the linear positive association assumption in Flynn 2017).

The most closely related paper is from Chernozhukov, Lee, and Rosen (2013) who study intersection bounds, where the lower bound on a parameter is the maximum of many statistics for which there are normally distributed estimators. In principle, the inference method developed in Chernozhukov, Lee, and Rosen (2013) can be applied to study the problems I consider in this paper, but I take advantage of the special structure of the problem to make the estimator more attractive in practice: the inference theory is standard asymptotic normality; when the statistic of interest is a linear function of the parameters, the estimator is the value of a linear program with a closed-form formula for inference; the confidence interval is theoretically non-conservative for a variety of data generating processes; and I do not need restrictions on the number of constraints.

The other advantage of my estimator, besides allowing for an infinite number of constraints, is that it is easy in practice to do inference using it; the confidence interval has a closed form. When we are interested in a linear function of parameters, to estimate the confidence interval only requires solving a linear program.

In Section 2, I put some restrictions on the identified set and give examples of economic assumptions that satisfy the restrictions.

In Section 3, I detail the estimator and the theory behind it.

In Section 4, I give practical advice on how to compute the estimator.

In Section 5, I show how to compute standard errors.

In Section 6, I suggest a practical method for implementing the estimator, choosing its tuning parameters.

In Section 7, I do a Monte Carlo study to learn the size of the test in finite samples.

2. Setting and examples.

Let $\{h_j\}_{j=1}^J$ be a collection of observable random variables with bounded variance. Let $u \in \mathscr{U}$ and $x_j \in \mathbb{R}^P$ be a random vector for each natural number j. Define:

$$v_j(u) = \mathbb{E}\left[h_j \mathbf{1}\left(x_j \geq u\right)\right].$$

Let $\mathscr{V}$ be the set with all typical elements $v_j(u)$.

The identified parameter set, $\Theta_0 \subset \mathbb{R}^b$ where $b < \infty$, is the intersection of countably many linear inequalities,

$$\Theta_0 = \left\{\theta \in \Theta : d_\ell(v(u))^\top \theta \leq p_\ell(v(u)) \text{ for } u \in \mathscr{U}, \ell = 1, \ldots, L < \infty\right\}, \tag{1}$$

Where (d_ℓ, p_ℓ) are continuously differentiable functions.

Additively-separable models where the observables are correlated with the unobservable residual in a certain way tend to generate identified sets that fit in this class.

Let y be the response variable, x be covariates, and e be the structural error,

$$y = x^\top \theta + e. \tag{2}$$

I give several examples of identified sets for θ where x is correlated with e in some way, and we have partial information on that relationship.

The following collection of sets will be useful:

$$\mathscr{C}_L = \left\{C \subset \mathbb{R}^L : C = \left\{z : \Phi_\ell(z_\ell) \in \left(\frac{o_\ell - 1}{m}, \frac{o_\ell}{m}\right)\right\} \text{ for some } o_\ell \in \{1, \ldots, m\}\right\}. \tag{3}$$

For some increasing functions Φ_ℓ with output in $[0,1]$. These are the countable cube instruments from Andrews and Shi (2013).

Example 1. *Monotone instrumental variable assumptions (see Manski and Pepper 2000),*

$$\frac{\partial}{\partial z}\mathbb{E}(e|z) \geq 0, \tag{4}$$

Are in the class of identified sets because they are equivalent to a countable number of covariance restrictions.

Theorem 1. *Say $z \in \mathbb{R}$ and that its distribution has no mass points. Monotone instrumental variable assumptions are equivalent to a countable number of restrictions on the covariance.*

Assume $\mathbb{E}(e|z)$ is continuously differentiable.

Then:

$$\frac{\partial}{\partial z}\mathbb{E}(e|z) \geq 0 \iff cov(e,z|z \in C) \geq 0 \quad \forall C \in \mathscr{C}_1. \tag{5}$$

Proof. $\Longrightarrow$: The conditional mean is increasing on C so the covariance is positive on C (a textbook result).

$$\begin{aligned} h(z) &= \mathbb{E}(e|z) \\ \operatorname{cov}(e,z) &= \mathbb{E}[z\mathbb{E}[e|z]] - \mathbb{E}[z]\mathbb{E}[\mathbb{E}[e|z]] = \operatorname{cov}(h(z),z) \end{aligned} \tag{6}$$

Let w have the same distribution as z but be independent of it. Because h is an increasing function:

$$\begin{aligned} & \mathbb{E}\{(w-z)(h(w)-h(z))\} \geq 0 \\ \iff & \mathbb{E}[wh(w)] - \mathbb{E}(w)\mathbb{E}(h(z)) - \mathbb{E}(z)\mathbb{E}(h(w)) + \mathbb{E}[zh(z)] \geq 0 \\ \iff & 2\operatorname{cov}(h(z),z) \geq 0 \end{aligned} \tag{7}$$

$\Longleftarrow$: $h(z)$ can not be decreasing for all z in any set $C \in \mathscr{C}_1$ or else the covariance of e and z conditional on $z \in C$ would be negative. Let z_0 be a point such that $h'(z_0) < 0$. Because $h'(z_0)$ is continuous, there exists an open ball containing z_0 such that $h'(z)$ is negative for all z in the open ball. Because there exists a $C \in \mathscr{C}_1$ that is a subset of the open ball, we have a contradiction. There is no such point z_0. □

Let φ be a strictly increasing function with output in $[0,1]$. The d and p are,

$$\begin{aligned} d_j(u_1,u_2) &= \mathbb{E}\left[x_j z \mathbf{1}(u_1 \leq \varphi(z) \leq u_2)\right]\mathbb{E}[\mathbf{1}(u_1 \leq \varphi(z) \leq u_2)] \\ &\quad - \mathbb{E}\left[x_j \mathbf{1}(u_1 \leq \varphi(z) \leq u_2)\right]\mathbb{E}[z\mathbf{1}(u_1 \leq \varphi(z) \leq u_2)] \\ p(u_1,u_2) &= \mathbb{E}[yz\mathbf{1}(u_1 \leq \varphi(z) \leq u_2)]\mathbb{E}[\mathbf{1}(u_1 \leq \varphi(z) \leq u_2)] \\ &\quad - \mathbb{E}[y\mathbf{1}(u_1 \leq \varphi(z) \leq u_2)]\mathbb{E}[z\mathbf{1}(u_1 \leq \varphi(z) \leq u_2)]. \end{aligned} \tag{8}$$

Example 2. *The linear positive association assumption which Flynn (2017) uses to partially identify the production function and productivity distribution. Let z be a vector of observables. For all increasing functions ϕ_1 and ϕ_2,*

$$cov[e\phi_1(z), \phi_2(z)] \geq 0, \tag{9}$$

After normalizing $e \geq 0$.

Let φ be a vector of functions with which are all strictly increasing and have output in $[0,1]$. The d and p are, for $(u_1,u_2) \in [0,1]^{2\delta_z}$, with δ_z being the dimension of z,

$$\begin{aligned} d_j(u_1,u_2) &= \mathbb{E}\left[x_j \mathbf{1}(\varphi(z) \geq u_1)\mathbf{1}(\varphi(z) \geq u_2)\right] - \mathbb{E}\left[x_j \mathbf{1}(\varphi(z) \geq u_1)\right]\mathbb{E}[\mathbf{1}(\varphi(z) \geq u_2)] \\ p(u_1,u_2) &= \mathbb{E}[y\mathbf{1}(\varphi(z) \geq u_1)\mathbf{1}(\varphi(z) \geq u_2)] - \mathbb{E}[y\mathbf{1}(\varphi(z) \geq u_1)]\mathbb{E}[\mathbf{1}(\varphi(z) \geq u_2)], \end{aligned}$$

Where φ is an L-vector of strictly increasing functions, each with range $[0,1]$.

Example 3. *Conditional moment inequalities,*

$$\mathbb{E}[e|z] \geq 0. \tag{10}$$

Which are a countable number of inequalities (see Andrews and Shi 2013),

$$\mathbb{E}[e\mathbf{1}(z \in C)] \geq 0 \quad \text{for all } C \in \mathscr{C}_L \tag{11}$$

Where $\mathscr{C}_L$ is as defined for the monotone instrumental variable assumptions and $z \in \mathbb{R}^L$.

The d and p are,

$$\begin{aligned} d_j(u_1,u_2) &= \mathbb{E}\left[x_j \mathbf{1}(u_1 \leq \varphi(z) \leq u_2)\right] \\ p(u_1,u_2) &= \mathbb{E}[y\mathbf{1}(u_1 \leq \varphi(z) \leq u_2)]. \end{aligned} \tag{12}$$

3. Bounds on functions of parameters.

Say we want to find a lower bound on $t(\theta,\gamma)$, a function of θ and point identified parameters, γ (moments of the data), given that $\theta \in \Theta_0$.

We can not simply minimize $t(\theta,\widehat{\gamma}_n)$ subject to θ belonging to the sample analogue of Θ_0 and be able to make standard normal inference. There are infinitely many constraints so there are badly-estimated constraints at all sample sizes if $\text{pr}\{z \geq u\}$ can be arbitrarily small. The probability that some moments, $v_j(u)$, are zero which ought to be non-zero does not decrease with sample size.

My solution is to make the bounds wider by relaxing the constraints which only use a few observations in estimation.

Throughout, I will assume $L = 1$ (there is only one set of functions for the coefficients) which is true in the examples in Section 2 and it is clear that all of these results hold without loss of generality, allowing L to be greater than one.

Define $n_j(u)$,

$$n_j(u) = \sum_{i=1}^{n} \mathbf{1}\left(x_{j,i} \geq u\right),$$

Where $x_{j,i}$ is the i'th observation of the random vector x_j and n is sample size.

Let $n(u) = \min_j n_j(u)$.

Let $\mu_j(u) = \mathbb{E}n_j(u)/n$ and $\mu(u) = \min_j \mu_j(u)$. Ideally, we would choose to use only constraints which have a high probability of drawing a high $n(u)$. We would only use constraints such that,

$$\mu(u) \geq \beta_n,$$

And we would send β_n to zero slowly. But we do not observe $\mu(u)$ and if we replace $\mu(u)$ with its sample estimate,

$$\frac{n(u)}{n} \geq \beta_n,$$

Then small errors in estimation will have discontinuous effects on the value of the program because a small error in estimation will delete a whole constraint.

We can think about not using a constraint as adding ∞ to the right hand side of the inequality. So using the constraint selection method above, we would write the constraints as,

$$d(u)^\top \theta \leq p(u) + \mathbf{1}(n(u) \geq n\beta_n) \times \infty.$$

Small changes in $(n(u)/n)$ will be the difference between adding ∞ or 0 to the right hand side. I replace the indicator function with a smooth version of the function,

$$d(u)^\top \theta \leq p(u) + \exp\left(\frac{\beta_n - \frac{n(u)}{n}}{\epsilon_n}\right).$$

To see that this is a smooth version of $\infty \times \mathbf{1}(n(u) \geq n\beta_n)$, take $\epsilon_n \to 0$ for a fixed β_n. If $n(u)/n > \beta_n$, then the penalty function goes to zero as $\epsilon_n \to 0$. If $n(u)/n < \beta_n$, then the penalty function goes to infinity.

My estimate of the lower bound on $t(\theta, \gamma)$ is the value of a finite linear program,

$$\begin{aligned} \widehat{t}_n = \min_\theta \quad & t(\theta, \widehat{\gamma}) \\ \text{ST:} \quad & \widehat{d}(u)^\top \theta \leq \widehat{p}(u) + \exp\left(\frac{\beta_n - (n(u)/n)}{\epsilon_n}\right) \quad \text{for } u \in \mathscr{U}_{K_n}, \end{aligned} \tag{13}$$

where,

$$\mathscr{U}_K = \left\{0, \frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}, 1\right\}^P. \tag{14}$$

Define the non-estimated, but penalized, program,

$$\begin{aligned} t_n = \min_\theta \quad & t(\theta, \gamma) \\ \text{ST:} \quad & d(u)^\top \theta \leq p(u) + \exp\left(\frac{\beta_n - \mu(u)}{\epsilon_n}\right) \quad \text{for } u \in \mathscr{U}_{K_n}. \end{aligned} \tag{15}$$

I make a few assumptions.

Assumption 1. *The program associated with t_n is continuously differentiable in all of its parameters (γ, d, p).*

Assumption 2. *γ is a finite vector of moments.*
Assumption 3. *Some subset of the constraints used in Θ_0 (the true identified set without the penalty terms) form a compact set. That is, there exists a $K < \infty$ such that the set of θ allowed by using only the $u \in \mathcal{U}_K$ is compact.*
Assumption 4. *The support of h_j is countable and $\underline{h} \leq h_j \leq \overline{h}$.*
Assumption 5. *Θ_0 has a non-empty interior (rules out point identification).*
Assumption 6. *The functions $(d(u), p(u))$ are bounded away from positive and negative infinity by fixed constants.*
Assumption 7. *$\beta_n \to 0$, $\epsilon_n \to 0$, $n\epsilon_n^2 \to \infty$, $\epsilon_n \log n \to 0$, $K_n \to \infty$, and $\sqrt{n}/K_n \to 0$.*

There exists some sequence ι_n such that: $\iota_n \to 0$, $\iota_n/\epsilon_n \to \infty$, $\iota_n < \beta_n$, $n(\beta_n - \iota_n)^3 \to \infty$.

Assumption 3 ensures a lower bound exists (because t is continuous in θ), and that it exists even if we use only a subset of the constraints.

Assumption 5 ensures that we are not trying to estimate a very small set so that, in large samples, the estimated set is non-empty.

Assumption 7 puts restrictions on the tuning parameters. Assumption 7 says the number of constraints we use is allowed to go to infinity at a rate faster than sample size, at least as fast as $o\left(n^{2-2\mu}\right)$ for $0 < \mu \leq 2/3$. This assumption can be removed if we have uniform convergence in the moments used to estimate the program (for example, in the linear positive association of Flynn 2017; this case will be discussed more later).

Assumption 1 is the most demanding assumption so I give sufficient conditions for it to be true in Lemma 1 (Theorem 6 from Georg Still's Parametric Optimization Lectures adapted to the current problem).

Lemma 1. *Let θ_n be a solution to the program, t_n. Let B be the collection of constraints that bind,*

$$k \in B \iff d(u_k)^\top \theta = p(u_k) + \exp\left(\frac{\beta_n - (n(u_k)/n)}{\epsilon_n}\right).$$

Assume the Linear Independence Constraint Qualification (LICQ) holds. That is, the vectors $\{d(u_k)\}_{k \in B}$ are linearly independent. This implies there are unique Lagrange multipliers which solve the Kuhn Tucker equations, λ. I assume $\lambda_k > 0$ if $k \in B$.

I assume either:

(1) The number of elements in B is b, the dimension of θ.

(2) Let $\nabla^2_\theta \mathcal{L}$ be the Hessian of the Lagrangian with respect to θ (which exists if $t(\theta, \gamma)$ is twice differentiable in θ). For all $m \neq 0$ such that $d_k^\top m = 0$ for all $k \in B$,

$$m^\top \nabla^2_\theta \mathcal{L} m > 0.$$

If either (1) or (2) is true, then for all $\left(\widehat{\gamma}, \widehat{d}, \widehat{p}\right)$ in an open neighborhood of (γ, d, p), there exists two functions, $\theta\left(\widehat{\gamma}, \widehat{d}, \widehat{p}\right)$ and $\lambda\left(\widehat{\gamma}, \widehat{d}, \widehat{p}\right)$ which give local minimizers and Lagrange multipliers

that solve the Kuhn-Tucker conditions of the $\widehat{t}_n$ program. Those functions are differentiable and so is the value function.

If $t(\theta,\gamma)$ is continuously differentiable in γ, then Assumption 1 holds.

Proof. Apply implicit function theorem to the Kuhn-Tucker conditions. □

Although the conditions required for Lemma 1 to hold are strong for a general optimization problem, we only need the *penalized* program to be differentiable and the penalty parameters are under our control. If we choose the penalty parameters carefully, we can ensure the program is differentiable (see Section 6).

3.1. Local normality.

I first prove a local normality result: the estimated program is normally distributed around the non-estimated program in large samples. I will then show the estimated bound is consistent and use both results to build a valid confidence interval for the lower bound.

The key is that the penalty term ensures badly-estimated constraints do not bind. Lemma 3 tells us that constraints where there is a probability less than ϵ_n of having all indicator functions non-zero will not bind in large samples.

I start by restating the standard Dvoretzky-Kiefer-Wolfowitz inequality in the context of the current problem.

Lemma 2. *For $\delta > 0$,*

$$pr\left\{\max_j \max_u |\frac{n_j(u)}{n} - \mu_j(u)| \geq \delta\right\} \leq 2J \times \exp\left(-2n\delta^2\right) \tag{16}$$

Proof. See Appendix A. □

Lemma 3. *Given Assumption 6 ($||(d,p)|| < \mathscr{M} < \infty$), in large samples, constraints with $\mu(u) < \beta_n - \iota_n$ do not bind if Assumption 3 and Assumption 7 are true.*

Proof. See Appendix A. □

Each linear inequality is constructed from a finite number of moments of the data with the form, $\mathbb{E}\left[h_j \mathbf{1}\left(x_j \geq u\right)\right]$. Stack up all the moments used for constraints with $u \in \mathscr{U}_{K_n}$ into a vector, V_n. Because the inequalities are linear, only b constraints can hold with equality (after removing redundant constraints). The maximum number of moments a constraint depends on is J so the maximum number of moments that are associated with a binding constraint is bJ.

The dimension of γ is J_g. Define $\kappa_b = bJ + J_g$.

I need to bound the second and third central moments of the sample moment estimators.

Assumption 8. *Let $W \subset V_n$ be such that W has at most κ_b elements. Let Σ_W be the variance-covariance matrix in,*

$$\sqrt{n}\left[\widehat{W} - W\right] \xrightarrow{d} N\left(0, \Sigma_W\right),$$

Where $\widehat{W}$ is the sample analogue of W (a vector of sample means).

Let $\mu(W)$ be the smallest $\mu_j(u) = pr\left(x_j \geq u\right)$ of any of the elements of W. Define:

$$\overline{\Sigma}_W = \frac{1}{\mu(W)}\Sigma_W. \tag{17}$$

Let η_W be the minimum eigenvalue of $\overline{\Sigma}_W$. I assume:

$$\min_W \eta_W \geq \underline{\eta} > 0. \tag{18}$$

Assumption 9. *Let $W \subset V_n$ be such that W has at most κ_b elements. Then, I estimate element W_ℓ by,*

$$\widehat{W_\ell} = \frac{1}{n}\sum_{i=1}^{n} h_{i,\ell} \mathbf{1}\left(x_{i,\ell} \geq u_\ell\right) = \frac{1}{n}\sum_{i=1}^{n} w_{i,\ell}. \tag{19}$$

I assume:

$$\mathbb{E}\left[||w_{i,\ell} - W_\ell||_2^3\right] < \mathscr{M} < \infty. \tag{20}$$

Assumption 8 can not be true for Σ_W because, without factoring out $\mu(W)$, the trace of Σ_W approaches zero as $\mu \to 0$ (and the trace is the sum of the eigenvalues, all of which are non-negative in a positive-definite matrix).

Lemma 4. *Let $W \subset V_n$ be such that W has at most κ_b elements. Given Assumption 8 and Assumption 9,*

$$\mathbb{E}\left[||\overline{\Sigma}_W^{-1/2}(w_i - W)||_2^3\right] < \mathscr{M} < \infty. \tag{21}$$

Proof. See Appendix A. □

I can write t_n as a function of V_n, the moments of the program. Because t_n is differentiable in its (γ, d, p) parameters and those parameters are differentiable in the moments V_n, we can take the derivative of t_n with respect to V_n. Define $\nabla t_n = \nabla t_n(V_n)$.

Let N be a fixed number, define Σ_N by,

$$\sqrt{n}\left(\widehat{V}_N - V_N\right) \xrightarrow{d} N\left(0, \Sigma_N\right),$$

Where $\widehat{V}_N$ is the sample analogue estimator of V_N, a vector of sample means.

Let N be a fixed number. Because t_N and $\widehat{t}_N$ are both smooth functions of only a finite number of moments, applying the Delta method gives,

$$\sqrt{n}\left(\widehat{t}_N - t_N\right) \xrightarrow{d} N\left(0, \nabla t_N^\top \Sigma_N \nabla t_N\right)$$

Define $\sigma_N = \sqrt{\nabla t_N^\top \Sigma_N \nabla t_N}$.

If we were to stop the complexity of the program at N, the bounds would be valid but inconsistent for the true bounds. To get a consistent estimator, I need to eventually use all the constraints.

Before we can state the main result, I first establish the uniform convergence of the moments $v \in \mathscr{V}$.

Lemma 5. *Let Assumption 4 hold. Estimates of the moments $v \in \mathscr{V}$ convergence uniformly in the sup-norm. Let P be the greatest dimension of any x_j. For any $\delta > 0$,*

$$pr\left\{\max_j \sup_u |\widehat{v}_j(u) - v_j(u)| \geq \epsilon\right\} \leq J \times C(\delta, P+1) \exp\left(-(2-\delta) n\epsilon^2\right), \tag{22}$$

Where $C(\delta, P+1)$ is a function of only δ and P.

Proof. See Appendix A. □

I combine the bounds on the moment from Lemma 4 and the fact that only constraints with $\mu(u) \geq \beta_n - \iota_n$ can bind with a Berry-Essen-like inequality from Gotze (1991) to get a "local normality" result.

Theorem 2. *Given Assumptions 1, 2, 3, 5, 6, 7, 8, and 9,*

$$\sup_u |pr\left\{\sqrt{n}\left[\widehat{t}_n - t_n\right]/\widehat{\sigma}_n \leq u\right\} - \Phi(u)| = O\left(n^{-1/2}(\beta_n - \iota_n)^{-3/2}\right), \tag{23}$$

Where Φ is the standard normal CDF.

Proof. See Appendix A. □

3.2. Consistency.

Having established that $\sqrt{n}\left[\widehat{t}_n - t_n\right]/\widehat{\sigma}_n$ converges to a normal distribution, I now show that $\widehat{t}_n \xrightarrow{p} t$, the true value of the lower bound which uses all the (non-penalized) constraints.

$$t = \min_\theta t(\theta, \gamma) \quad \text{ST:} \quad d(u)^\top \theta \leq p(u) \quad \text{for } u \in \mathscr{U}.$$

I use the fact that t is the value of a semi-infinite programming problem, a problem with a finite number of controls and an infinite number of constraints, to derive two consistency results.

Theorem 3. *Say $t(\theta, \gamma)$ is a linear function of θ. If the premises of Theorem 2 are true, then $t_n \to t$. Which, given Theorem 2, implies $\widehat{t}_n \xrightarrow{p} t$.*

Proof. See Appendix A. □

The linearity of t is not important. What is important is that the program can be "discretized". That is, the value of the program using a finite selection of the constraints approaches the value of the program with all the constraints as we add more and more constraints.

If the objective function is not linear, the program is discretizable if the magnitude of the coefficients becomes small in large samples, or, more generally, if eventually the constraints we add are not so different from the previous constraints.

Theorem 4. *If the assumptions of Theorem 2 hold and either:*

(1) $\lim_K \sup_{u \notin \mathscr{U}_K} ||(d(u), p(u))||_\infty = 0$.

(2) Or, more generally,

$$\lim_K \sup_u \inf_{u' \in \mathscr{U}_K} ||(d(u), p(u)) - \left(d\left(u'\right), p\left(u'\right)\right)||_\infty \to 0$$

Then, $t_n \to t$ *and so, from Theorem 2,* $\widehat{t}_n \xrightarrow{p} t$.

Proof. See Appendix A. □

3.3. One-sided confidence intervals.

I can use these results to form a one-sided confidence interval for the lower bound on $t(\theta, \gamma)$ for $\theta \in \Theta_0$.

I am looking for a random variable c_n such that,

$$\lim_n \text{pr}\{t \geq c_n\} \geq \alpha. \tag{24}$$

Where "pr" denotes the probability measure under the true data generating process.

First, $t_n \leq t$ because it uses fewer constraints and adds positive numbers to the right hand side of the constraints it does use.

Write,

$$\frac{\sqrt{n}}{\widehat{\sigma}_n}\left[t - \widehat{t}_n\right] = \frac{\sqrt{n}}{\widehat{\sigma}_n}\left[t_n - \widehat{t}_n\right] + \frac{\sqrt{n}}{\widehat{\sigma}_n}\left[t - t_n\right] = \frac{\sqrt{n}}{\widehat{\sigma}_n}\left[t_n - \widehat{t}_n\right] + o_+\left(\sqrt{n}\right),$$

Where $o_+(a_n)$ denotes a positive sequence such that $b_n \in o_+(a_n)$ implies $b_n \geq 0$ and $b_n/a_n \to 0$.

From Theorem 2,

$$\sup_{u, o_+\left(\sqrt{n}\right)} \left|\text{pr}\left\{\frac{\sqrt{n}}{\sigma_n}\left[\widehat{t}_n - t_n\right] \leq u - o_+\left(\sqrt{n}\right)\right\} - \Phi\left(u - o_+\left(\sqrt{n}\right)\right)\right| = O\left(n^{-1/2}\left(\beta_n - \iota_n\right)^{-3/2}\right),$$

Uniformly across sequences in $o_+\left(\sqrt{n}\right)$. So, we have:

$$\begin{aligned}\text{pr}\left\{\frac{\sqrt{n}}{\widehat{\sigma}_n}\left[t-\widehat{t}_n\right]\geq c_n\right\} &= \alpha + O\left(n^{-1/2}\left(\beta_n-\iota_n\right)^{-3/2}\right)\\ c_n &= \Phi^{-1}\left(1-\alpha\right)+o_+\left(\sqrt{n}\right).\end{aligned}$$

But we can not choose the sequence c_n because we do not know the sequence in $o_+\left(\sqrt{n}\right)$. For the lower confidence interval, what matters is the smallest sequence in $o_+\left(\sqrt{n}\right)$, which is the zero sequence. Because $c_n \geq \Phi^{-1}\left(1-\alpha\right)$ we have,

$$\alpha + O\left(n^{-1/2}\left(\beta_n-\iota_n\right)^{-3/2}\right) = \text{pr}\left\{\frac{\sqrt{n}}{\widehat{\sigma}_n}\left[t-\widehat{t}_n\right]\geq c_n\right\} \leq \text{pr}\left\{\frac{\sqrt{n}}{\widehat{\sigma}_n}\left[t-\widehat{t}_n\right]\geq \Phi^{-1}\left(1-\alpha\right)\right\}.$$

So the one-sided confidence interval for t is,

$$\text{pr}\left\{t \geq \widehat{t}_n + \frac{\widehat{\sigma}_n}{\sqrt{n}}\Phi^{-1}\left(1-\alpha\right)\right\}\geq \alpha.$$

The advantage of the confidence interval is that I do not need to invert a hypothesis test or search the parameter space to compute it. I only need to compute $\widehat{\sigma}_n$ and $\widehat{t}_n$, enabling me to use flexible specifications without hitting computational limits, especially when t is linear (in the linear case, $\widehat{\sigma}_n$ is a known function of the primal and dual values of the linear program and the variance-covariance matrix of the moments of the program). See Section 5 for details.

The confidence interval is not conservative because the confidence interval has size α if only a finite number of constraints are actually needed (but which constraints are needed is unknown to the econometrician). The property in the optimization literature is called, "reducibility", see Goberna and Lopez (1987) for a set of conditions that imply a program is reducible if $t\left(\theta,\gamma\right)$ is a linear function of θ.

Theorem 5. *If the program associated with t, the true lower bound, is reducible—that is, if there exists a finite subprogram which gives the same value as t,*

$$t = \min_{\theta} t\left(\theta,\gamma\right) \quad \text{ST:} \quad d\left(u\right)^{\top}\theta \leq p\left(u\right) \quad \text{for } u \in \mathscr{U}_K. \tag{25}$$

If $\epsilon_n \log n \to 0$, then, if the assumptions from Theorem 2 hold,

$$\frac{\sqrt{n}}{\widehat{\sigma}_n}\left(\widehat{t}_n - t\right) \xrightarrow{d} N\left(0,1\right). \tag{26}$$

Proof. See Appendix A. □

3.4. Two-sided confidence intervals for $t(\theta,\gamma)$.

But we are often interested in two-sided confidence intervals for $t(\theta,\gamma)$. If we are willing to assume that the difference between the upper and lower bound is large relative to the error in estimating the lower and upper bound, then we can use the one-sided confidence intervals for the lower and upper bound to make a two-sided confidence interval. But Imbens and Manski (2004) point out that there are data-generating processes that can make these confidence intervals too small at all sample sizes. They propose a uniform two-sided confidence interval which can be adapted to my problem.

I start with a uniformity result.

Theorem 6. *Let $\mathscr{M}$ be a fixed, real number. Let $\mathscr{G}$ be a set of data generating processes such that all the assumptions of Theorem 2 hold and, in addition, if for all $G \in \mathscr{G}$,*

$$\sup_W \mathbb{E}_G\left[||\overline{\Sigma}_W^{-1/2}\left(w_{i,\ell} - W\right)||_2^3\right] \leq \mathscr{M}, \tag{27}$$

I have:

$$\lim_n \inf_{G\in\mathscr{G}} pr_G\left\{t \geq \widehat{t}_n + \frac{\widehat{\sigma}_n}{\sqrt{n}}\Phi^{-1}(\alpha)\right\} \geq \alpha. \tag{28}$$

Proof. See Appendix A. □

Define t_n^u and the t_n^ℓ to be the value of the programs for the upper and lower bounds (the upper bound minimizes the negative of $t(\theta,\gamma)$).

The data-generating processes that give the tightest bounds are those where both the upper and lower bounds on $t(\theta,\gamma)$ are reducible, where t^ℓ and t^u depend on only an unknown, but finite subset of the constraints.

For the reducible data generating processes (by a simple generalization of Theorem 5 to the two-dimensional case),

$$\sqrt{n}\begin{pmatrix} \widehat{t}_n^u - t^u \\ \widehat{t}_n^\ell - t^\ell \end{pmatrix} \xrightarrow{d} N(0,S). \tag{29}$$

Assume for all data generating processes with $t^\ell \leq t \leq t^u$ (whether they are reducible or not),

$$\sup_W \mathbb{E}_G\left[||\overline{\Sigma}_W^{-1/2}(w_i - W)||_2^3\right] \leq \mathscr{M}, \tag{30}$$

And that the diagonal elements of S are bounded from above and below, uniformly, across data generating processes with $t^\ell \leq t \leq t^u$.

Define Δ to be the distance between the upper and lower bound,

$$t^u - t^\ell = \Delta.$$

Lemma 6. *If for all data generating processes that satisfy the assumptions of Theorem 2 and which have* $t^\ell \leq t \leq t^u$,

$$\sup_W \mathbb{E}_G \left[||\overline{\Sigma}_W^{-1/2}(w_i - W)||_2^3 \right] \leq \mathscr{M}, \tag{31}$$

Let W_n *be the stacked vector of moments that belong to a binding constraint in either* $\widehat{t}_n^u$ *or* $\widehat{t}_n^\ell$. *Define* $\widehat{S}_n = \nabla_W \widehat{t}_n^\top \widehat{\Sigma}_{W_n} \nabla_W \widehat{t}_n$.

For the class of data generating processes such that $t_n^\ell - t^\ell = o\left(n^{-1/2}\right)$ *and* $t_n^u - t^u = o\left(n^{-1/2}\right)$, $\mathscr{G}_1$,

$$\sup_{G \in \mathscr{G}_1} \sup_v \left| pr\left\{ \sqrt{n} \widehat{S}_{G,n}^{-1/2} \left[\begin{pmatrix} \widehat{t}_n^u - t^u \\ \widehat{t}_n^\ell - t^\ell \end{pmatrix} \right] \leq v \right\} - \Phi(v) \right| = O\left(n^{-1/2} (\beta_n - \iota_n)^{-3/2} \right). \tag{32}$$

Proof. See Appendix A. □

If, in addition to the conditions of Lemma 6, for all data generating processes in $\mathscr{G}_1$, $\underline{\sigma} \leq \sigma_\ell \leq \overline{\sigma}$ and $\underline{\sigma} \leq \sigma_u \leq \overline{\sigma}$ for fixed $(\underline{\sigma}, \overline{\sigma})$ and if there exists an $\delta > 0$ such that $t^u - t^\ell \geq \delta$ for all data generating processes in $\mathscr{G}_1$, then Assumption 1 of Imbens and Manski (2004) holds for data generating process in $\mathscr{G}_1$ and so we can use their two-sided confidence interval:

$$\inf_{G \in \mathscr{G}_1} \lim_n \text{pr} \left\{ \widehat{t}_n^\ell - c_n \frac{\widehat{\sigma}^\ell}{\sqrt{n}} \leq t \leq \widehat{t}_n^u + c_n \frac{\widehat{\sigma}^u}{\sqrt{n}} \right\} \geq \alpha, \tag{33}$$

Where c_n solves:

$$\Phi\left(c_n + \sqrt{n} \times \frac{\widehat{t}_n^u - \widehat{t}_n^\ell}{\max\left(\widehat{\sigma}_n^\ell, \widehat{\sigma}_n^u\right)} \right) - \Phi(-c_n) = \alpha.$$

Because the probability of the interval covering non-reducible programs is greater than the probability of the interval covering reducible programs, the two-sided confidence interval is valid for data generating processes that are not reducible.

3.5. Linear functions of infinite-dimensional parameters partially identified by a finite number of linear constraints.

The estimator and confidence interval can also be used for problems when θ is a real sequence (infinite-dimensional parameter) and there are a finite number of constraints. Say, after some transformation, the problem is in the standard form:

$$\begin{aligned} \max_\theta \quad & \sum_{j=1}^\infty \theta_j t_j(\gamma) \\ \text{ST:} \quad & \sum_j d_{k,j} \theta_j = p_k \quad \text{for } k = 1, \ldots, K \\ & \theta_j \geq 0 \quad \sup\left\{ j : \theta_j > 0 \right\} < \infty \end{aligned} \tag{34}$$

This is the dual program to the minimization problem I presented earlier in this section. There is no duality gap under a variety of conditions (see Karney 1981), including the broad condition that the dual constraint set is bounded and non-empty.

I can do inference on problems with this form by transforming them into their dual and using the estimator I presented above.

The dual is:

$$\begin{aligned} \min_{\lambda} \quad & -p^{\top}\lambda \\ \text{ST:} \quad & d_j^{\top}\lambda \leq -t_j(\gamma) \text{ for } j = 1, 2, \ldots \end{aligned} \tag{35}$$

3.6. Two-sided inference on the value of a semi-infinite linear program.

Although the motivating cases for this paper deal with partial identification of $t(\theta, \gamma)$, there may be independent interest in making inference on the value of the optimization problem and, so far, I have only established a lower confidence interval for the value of the optimization problem.

To establish a two-sided confidence interval for the value of the optimization problem, all we need to do is ensure that,

$$\sqrt{n}[t_n - t] \to 0. \tag{36}$$

To establish the lower confidence interval, I used the fact that $t_n \leq t$. To establish an upper confidence interval, I need a lower bound on $t_n - t$.

Say $t(\theta, \gamma) = t(\gamma)^{\top}\theta$. Then, from Still (2001), we know,

$$\sqrt{n}[t_n - t] \geq \text{constant}\sqrt{n} \times d_H\left(\mathscr{U}, \mathscr{U}_{K_n}\right) = \text{constant}\sqrt{n} \times \frac{1}{2K_n} \times \sqrt{P}, \tag{37}$$

Where the last equality holds via geometry (the distance from the center of a P-cube to its vertex). So, as long as $\sqrt{n}/K_n \to 0$, we can use the standard two-sided confidence interval based on the normal distribution as a confidence interval for the value of the linear program,

$$\text{pr}\left\{\widehat{t}_n - \frac{\widehat{\sigma}_n}{\sqrt{n}} \times \Phi^{-1}(\alpha) \leq t \leq \widehat{t}_n + \frac{\widehat{\sigma}_n}{\sqrt{n}}\Phi^{-1}(\alpha)\right\} \geq \alpha. \tag{38}$$

4. Computation.

When $t(\theta, \gamma)$ is linear in θ, $\widehat{t}_n$ can be computed using standard linear programming methods. But when $t(\theta, \gamma)$ is nonlinear, the estimator is a constrained optimization problem with a very large number of linear constraints. How should we compute the estimator in the nonlinear case?

The best solutions I have found are the successive linear approximation algorithms which approximate the objective function by a linear function, solve the approximate linear program, and update the parameter using a rule until it converges.

The COBYLA method of Powell (1994) works well for a reasonable number of parameters (it is a derivative-free method). It uses linear interpolations of the objective function and solves linear programming problems to improve the interpolations, it was designed to work with problems with many constraints, and the initial value we use does not need to be feasible. Implementations of the method exist in many languages, including Fortran (Powell's original code), R (library `nloptr`), and in the C library nlopt.

If we know the derivative of the objective function, then Powell (1989)'s TOLMIN algorithm can be used as well.

5. Computing $\widehat{\sigma}_n$.

Computing $\widehat{\sigma}_n$ is straightforward. The first order conditions of the estimated program are,

$$\nabla_\theta t\left(\theta, \widehat{\gamma}\right) = -\widehat{D}_{n,B}^\top \widehat{\lambda}_{n,B},$$

Where $D_{n,B}$ denote the stacked $d(u)$ coefficients that belong to binding constraints ($k \in B$ where B is defined as in Lemma 1) and λ_B the corresponding Lagrange multipliers.

I assume $\widehat{D}_{n,B}^\top$ is left-invertible (the Linear Independence Constraint Qualification from Lemma 1 implies this). Then,

$$\widehat{\lambda}_{n,B} = -\left(\widehat{D}_{n,B}\widehat{D}_{n,B}^\top\right)^{-1} \widehat{D}_{n,B} \nabla_\theta t\left(\widehat{\theta}_n, \widehat{\gamma}\right).$$

Let Ω be the variance-covariance matrix in,

$$\sqrt{n}\left[\left(\text{vec}\left(\widehat{D}_{n,B}\right), \widehat{P}_{n,B}, \widehat{\gamma}, n_B/n\right) - \left(\text{vec}\left(D_B\right), P_B, \gamma, \mu_B\right)\right] \xrightarrow{d} N\left(0, \Omega\right).$$

Then the standard error, $\widehat{\sigma}_n$, is[1],

$$\widehat{\kappa}_n = \left(\left\{\widehat{\lambda}_{n,k}\widehat{\theta}_n\right\}_{k\in B}, \left\{-\widehat{\lambda}_{n,k}\right\}_{k\in B}, \nabla_\gamma t\left(\widehat{\theta}_n, \widehat{\gamma}_n\right), \left\{\frac{\widehat{\lambda}_{n,k}}{\epsilon_n}\exp\left(\frac{\beta_n - \left(n\left(u_k\right)/n\right)}{\epsilon_n}\right)\right\}_{k\in B}\right)$$

$$\widehat{\sigma}_n = \sqrt{\widehat{\kappa}_n^\top \widehat{\Omega}_n \widehat{\kappa}_n}.$$

[1] Note that if you scale the coefficients or penalty function in some way (as I will in the next section), you need to adjust the derivative ($\widehat{\kappa}_n$). In the next section, I multiply the penalty function by the largest absolute value of any coefficient so I would need to multiply the part of $\widehat{\kappa}_n$ corresponding to n_B/n by that value.

6. Choosing tuning parameters.

But we still need to choose the tuning parameters, β_n and ϵ_n.

I write the penalty function,

$$\exp\left(\frac{\beta_n}{\epsilon_n}\right)\exp\left(-\frac{(n(u)/n)}{\epsilon_n}\right),$$

I choose the ratio of β_n/ϵ_n separately from choosing ϵ_n. β_n/ϵ_n changes the scale of the penalty function and so to make the choice applicable across data generating processes we need to scale the constraints. I do so by normalizing the (d,p)'s to be between -1 and 1 by multiplying the penalty function by the max-norm of the coefficients,

$$\begin{aligned} z(u) &= ||(d(u), p(u))||_\infty \\ \frac{d(u)^\top}{z(u)}\theta &\leq \frac{p(u)}{z(u)} + \exp\left(\frac{\beta_n - (n(u)/n)}{\epsilon_n}\right) \\ \iff d(u)^\top\theta &\leq p(u) + z(u)\exp\left(\frac{\beta_n - (n(u)/n)}{\epsilon_n}\right). \end{aligned}$$

I found choosing β_n/ϵ_n such that,

$$\frac{\beta_n}{\epsilon_n} = \frac{1}{r}\log\log\log n,$$

For some parameter $r > 0$ worked well. Then the penalty function is,

$$d(u)^\top\theta \leq p(u) + z(u)(\log\log n)^{1/r} \times \exp\left(-\frac{n(u)}{n\epsilon_n}\right).$$

ϵ_n chooses which constraints bind (the other factors are constant across constraints) so it is the most important parameter. I choose ϵ_n by minimizing an approximation to the mean squared error of the estimator,

$$\epsilon_n = \arg\min_\epsilon n\left[\sum_{u\in\mathscr{U}_{K_n}} \widehat{\lambda}_n(u,\epsilon)\widehat{z}(u)(\log\log n)^{1/r}\exp\left(-\frac{n(u)}{n\epsilon}\right)\right]^2 + \widehat{\sigma}_n(\epsilon)^2,$$

Where $\widehat{\sigma}_n$ is the standard error of the lower bound and $\widehat{\lambda}_n$ are the Lagrange multipliers (both for a given $t(\theta,\gamma)$ function).

The second term is the variance of the estimator and the first term is a first order approximation to the bias from using the penalty function. Let $\widehat{t}_n$ be the estimated, penalized function and

$\widehat{t}_{n,0}$ be the estimated, non-penalized program (but still truncated, only using constraints with $u \in \mathscr{U}_{K_n}$). Then, the first order Taylor expansion gives,

$$\widehat{t}_{n,0} \approx \widehat{t}_n + \sum_{u \in \mathscr{U}_{K_n}} \widehat{\lambda}_n(u,\epsilon)\widehat{z}(u)(\log\log n)^{1/r} \exp\left(-\frac{n(u)}{n\epsilon}\right),$$

Which is the source of the first term.

The two terms capture the basic trade-off with choosing ϵ_n: low ϵ_n makes bias fall but allows lower probability constraints to bind, increasing variance.

Choose a grid for ϵ_n and compute the objective function over each point on the grid. Plot ϵ versus the value of the objective function. You will see the objective is discontinuous in ϵ around the points where the basis of the program switches. The idea is to choose the region of ϵ with the smallest value of the objective function but to choose ϵ a safe distance from the switching point (where the program would be non-differentiable). The easiest way to do this is to choose the ϵ that minimizes the objective function subject to the constraint that it is a reasonable distance from a point where the basis switches. Because the objective function (and the estimate $\widehat{t}_n$) will be continuous in ϵ in this neighborhood, it is not so important what distance you pick.

Choosing ϵ_n in this way ensures that the $\widehat{t}_n$ program is differentiable which, in large samples, will ensure the t_n program is, as well because $||\widehat{V}_n - V_n|| \xrightarrow{p} 0$.

If we want to ensure the theory holds, we can require the choice of ϵ_n to be between two fixed sequences that satisfy Assumption 7, but I do not do so. One theoretical advantage of not doing so is that if the true semi-infinite program is reducible, lowering ϵ_n will eventually not change which constraints bind so this method of choosing ϵ_n will send ϵ_n to zero very fast—which is exactly what we want in that case.

7. Monte Carlo.

Let $y = x^\top\theta + e$ and assume $\nabla\mathbb{E}(e|x_\ell) \geq 0$ for each x_ℓ and $\theta \geq 0$.

The data-generating process is,

$$\begin{aligned} a &\sim N(0,1) \\ x_\ell &= a + u_\ell, \quad u_\ell \sim N(0,1) \\ e &= 0.1 \times a^3 + \epsilon, \quad \epsilon \sim N(0,1) \\ y &= \sum_{\ell=1}^{L} x_\ell + e \end{aligned} \tag{39}$$

I want bounds on $\theta^\top 1$, which, in fact, is equal to L (the dimension of x), using the information that each x_ℓ is a monotone instrumental variable. The true upper bound is the solution to the

linear program,

$$\begin{aligned}
\max_{\theta} \quad & \sum_{\ell} \theta_\ell \\
\text{ST:} \quad & \theta_\ell + \sum_{k \neq \ell} \frac{1}{2}\theta_k \leq \min_{x_\ell} \frac{d}{dx_\ell}\mathbb{E}(y|x_\ell) = 1 + \frac{1}{2}(L-1) + \frac{3}{40} \quad \text{for } \ell = 1, \ldots, L \\
& \theta \geq 0.
\end{aligned} \tag{40}$$

I use the covariance restrictions from Example 1 to estimate the bounds,

$$\begin{aligned}
\text{cov}\left(y, x_\ell | \frac{k-1}{m} \leq \varphi(x_\ell) \leq \frac{k}{m}\right) \geq \sum_{k=1}^{L} \theta_k \text{cov}\left(x_k, x_\ell | \frac{k-1}{m} \leq \varphi(x_\ell) \leq \frac{k}{m}\right) \\
\forall k \in 1, \ldots, m \quad \forall m \leq m_n \quad \forall \ell,
\end{aligned} \tag{41}$$

I choose φ to be the normal CDF with the same mean and standard deviation as the observed data. It does not matter whether we estimate φ or not because φ just needs to be a strictly increasing function mapping the real line to $[0,1]$; it does not need to be any specific function. The true identified set is the same regardless of which φ we use (so if the φ changes with sample size, that is fine because we can write the true identified set using any φ). The only restriction is that φ must be *strictly* increasing in every sample (for large enough n), ruling out using the empirical CDF.

Where $m_n \to \infty$.

While it asymptotically will not matter what sequence m_n I use so long as it goes to infinity fast enough, I try using different m_n in order to learn whether it matters how many I use in finite samples. I choose $m_n = m_{1000} \times \left(n^{1/3}/1000\right)$ and try $m_{1000} = 5, 10$[2].

The main parameter to choose is r which sets the rate at which β_n/ϵ_n goes to infinity. I try $r = \{4, 5, 6\}$.

I do 10000 simulations of the data generating process. The results of the Monte Carlo are in Table 1. The size of the confidence interval is the probability that the estimated (one-sided) confidence interval contained the true value, which would be 90% if the confidence interval were non-conservative for this data generating process. The size of the confidence interval is always valid and not sensitive to how many constraints used or how fast β_n/ϵ_n goes to infinity.

I also compute the average distance of the confidence interval from the true upper bound as a percentage of the true upper bound (CI-Bound). The point of doing this is to learn how much of an error the confidence interval makes (if the confidence interval has the wrong size but is roughly the correct upper bound, then the size distortion does not matter to the economist). The average percentage difference of the confidence interval from the bound is sensitive to how many constraints are used (but the size is not).

[2]This rate is not fast enough for two-sided inference, but it works for one-sided inference. With some experimentation, I found that setting rates slower than $\sqrt{n}$ worked well for one-sided confidence intervals.

Table 1: Monte Carlo Results (for $\alpha = 0.90$ confidence intervals)

L	m_{1000}	r	n	Size	(CI - Bound)%	Bias (%)	Root-MSE
3	5	4	1000	0.95	5.2%	2.5%	0.23
3	5	4	5000	0.99	4.4%	3.5%	0.15
3	5	4	10000	0.99	4.1%	3.4%	0.14
3	10	4	1000	0.95	53.6%	−33.5%	1.24
3	10	4	5000	0.97	65.0%	−43.2%	1.49
3	10	4	10000	0.98	58.9%	−44.5%	1.51
3	5	5	1000	0.95	5.1%	2.3%	0.24
3	5	5	5000	0.99	4.4%	3.4%	0.16
3	5	5	10000	0.99	4.1%	3.3%	0.15
3	10	5	1000	0.95	54.4%	−34.5%	1.26
3	10	5	5000	0.98	96.0%	−47.1%	1.62
3	10	5	10000	0.98	79.6%	−48.4%	1.63
3	5	6	1000	0.94	5.1%	2.2%	0.24
3	5	6	5000	0.99	4.4%	3.3%	0.17
3	5	6	10000	0.99	4.1%	3.3%	0.15
3	10	6	1000	0.95	56.7%	−35.2%	1.27
3	10	6	5000	0.98	130.4%	−47.6%	1.63
3	10	6	10000	0.98	109.4%	−51.6%	1.74

The bias of the estimator is the average point estimate ($\widehat{t}_n$) minus the true upper bound (as a percentage of the true upper bound), and the root mean square is the square root of the average squared distance between the estimated bound and the true upper bound. The bias and mean squared error are also sensitive to the number of constraints used. The mean square error is increasing in sample size from $n = 1000$ to $n = 10000$ for $m_{1000} = 10$. Eventually, the mean squared error will fall but it does not need to be monotonic in sample size (for this data generating process, when $m_{1000} = 10$, the bias is increasing from $n = 1000$ to $n = 10000$, but the variance is falling).

The Monte Carlo results show that the size of the bounds is not sensitive to tuning parameter choices even though the point estimates are sensitive to how many constraints are used. In practical empirical problems, the bounds are often tight enough for us to learn something and for us to use a flexible model, see Flynn (2017) for an example of the estimator used in practice.

8. Conclusion.

I develop a consistent estimator for bounds on functions of parameters identified by the intersection of continuous linear inequalities and a confidence interval. The confidence interval can be computed in closed form, and its size is not so sensitive to tuning parameter choices.

When the function of interest is linear in parameters, the estimator is the value of a linear program and the confidence interval can be computed using only the dual and primal values of the program and a covariance matrix of sample moments.

The principles behind the estimator could be used to estimate identified sets defined by other forms of inequality constraints. Linearity of the constraints (in parameters) is mostly important for ensuring that only so many constraints can bind simultaneously, but it is not difficult to relax the assumption of linearity if we ensure this assumption holds for the nonlinear constraints.

Infinite-dimensional parameter spaces with an infinite number of constraints are more difficult to deal with because we then have two infinities that work in opposite directions. Truncating the number of constraints used make the value of a minimization problem smaller, but truncating the number of parameters used makes the value of a minimization problem larger. The "local normality" result still holds under restrictions on how fast the parameter space is allowed to grow, but consistency and confidence intervals are more difficult to establish.

References.

Andrews, D. and X. Shi (2013). Inference based on conditional moment inequalities. *Econometrica 81*(2), 609–666.

Beresteanu, A. and F. Molinari (2008). Asymptotic properties for a class of partially identified models. *Econometrica 76*(4), 763–814.

Bugni, F., I. Canay, and X. Shi (2016). Inference for subvectors and other functions of partially identified parameters in moment inequality models. *Quantitative Economics*.

Chernozhukov, V., S. Lee, and A. Rosen (2013). Intersection bounds: estimation and inference. *Econometrica 81*(2), 667–737.

Flynn, Z. (2017). Identifying productivity when it is a choice. *WORKING PAPER*.

Goberna, M. and M. Lopez (1987). Reduction and discrete approximation in linear semi-infinite programming. *Optimization 18*(5), 643–658.

Gotze, F. (1991). On the rate of convergence in the multivariate clt. *The Annals of Probability 19*(2), 724–739.

Imbens, G. and C. Manski (2004). Confidence intervals for partially identified parameters. *Econometrica 72*(6), 1845–1857.

Kaido, H. (2016). A dual approach to inference for partially identified econometric models. *Journal of Econometrics 192*(1), 269–290.

Kaido, H., F. Molinari, and J. Stoye (2016). Confidence intervals for projects of partially identified parameters. *Working paper*.

Kaido, H. and A. Santos (2014). Asymptotically efficient estimation of models defined by convex moment inequalities. *Econometrica 82*(1), 387–413.

Karney, D. (1981). Duality gaps in semi-infinite linear programming—an approximation problem. *Mathematical Programming 20*, 129–143.

Kiefer, J. (1961). On large deviations of the empiric d. f. of vector chance variables and a law of the iterated logarithm. *Pacific Journal of Mathematics 11*(2), 649–660.

Manski, C. and J. Pepper (2000). Monotone instrumental variables, with an application to the returns to schooling. *Econometrica 68*(4), 997–1012.

Powell, M. (1989). A tolerant algorithm for linearly constrained optimization calculations. *Mathematical Programming 45*, 547–566.

Powell, M. (1994). A direct search optimization method that models the objective and constraint functions by linear interpolation. *Advances in Optimization and Numerical Analysis*, 51–67.

Reemtsen, R. (1991). Discretization methods for the solution of semi-infinite programming problems. *Journal of Optimization Theory and Applications 71*(1).

Still, G. (2001). Discretization in semi-infinite programming: the rate of approximation. *Mathematical Programming 91*(1), 53–69.

A. Proofs.

A.1. Proof of Lemma 2.

Proof of Lemma 2.

From the union bounds,

$$\text{pr}\left\{\max_j \max_u |\frac{n_j(u)}{n} - \mu_j(u)| \geq \delta\right\} \leq \sum_{j=1}^{J} \text{pr}\left\{\max_u |\frac{n_j(u)}{n} - \mu_j(u)| \geq \delta\right\}. \tag{42}$$

Because $\mu_j(u) = \text{pr}\{x_j \geq u\}$ is a CDF and $n_j(u)/n$ is an empirical CDF estimating it, we can apply the Dvoretzky-Kiefer-Wolfowitz inequality to obtain,

$$\sum_{j=1}^{J} \text{pr}\left\{\max_u |\frac{n_j(u)}{n} - \mu_j(u)| \geq \delta\right\} \leq J \times 2\exp\left(-2n\delta^2\right). \tag{43}$$

□

A.2. Proof of Lemma 3.

Proof of Lemma 3.

Let ι_n be any sequence such that $\iota_n/\epsilon_n \to \infty$ and $\iota_n \to 0$ and $\iota_n < \beta_n$ (as in Assumption 7). For $\mu(u) \leq \beta_n - \iota_n$,

$$\exp\left(\frac{\beta_n - \mu(u)}{\epsilon_n} + \frac{\mu(u) - (n(u)/n)}{\epsilon_n}\right) \geq \exp\left(\frac{\iota_n}{\epsilon_n} + \frac{\mu(u) - (n(u)/n)}{\epsilon_n}\right) \tag{44}$$

For any sequence $\delta_n > 0$ such that $n\delta_n^2 \to \infty$, with probability approaching one $-\delta_n \leq n(u)/n - \mu(u) \leq \delta_n$.

So, with probability approaching one, we have:

$$\exp\left(\frac{\iota_n}{\epsilon_n} + \frac{\mu(u) - (n(u)/n)}{\epsilon_n}\right) \geq \exp\left(\frac{\iota_n}{\epsilon_n} - \frac{\delta_n}{\epsilon_n}\right). \tag{45}$$

If $n\epsilon_n^2 \to \infty$ (which is true by Assumption 7), then choose $\delta_n = \epsilon_n$ and then,

$$\exp\left(\frac{\iota_n}{\epsilon_n} - 1\right) \to \infty, \tag{46}$$

So the penalty function approaches infinity for u such that $\mu(u) \leq \beta_n - \iota_n$.

But for constraints with $\mu(u) \geq \beta_n + \iota_n$, the penalty function approaches zero.

$$\exp\left(\frac{\beta_n - \mu(u)}{\epsilon_n} + \frac{\mu(u) - (n(u)/n)}{\epsilon_n}\right) \leq \exp\left(-\frac{\iota_n}{\epsilon_n} + \frac{\mu_k - (n_k/n)}{\epsilon_n}\right) \leq \exp\left(-\frac{\iota_n}{\epsilon_n} + 1\right) \to 0. \tag{47}$$

Because some finite subset of the constraints ($u \in \mathscr{U}_K$ for some K) form a compact set (Assumption 3) and $\beta_n + \iota_n \to 0$, the finite subset of constraints with $\mu(u) \geq \beta_n + \iota_n$ and $u \in \mathscr{U}_K$ is compact in large samples so it is bounded, $||(\theta, -1)||_2 \leq \mathscr{K}/\mathscr{M}$.

Let θ be in the set using constraints with $\mu(u) \geq \beta_n + \iota_n$. I will show it is in the interior of the sets with $\mu(u) \leq \beta_n - \iota_n$ in large samples.

I write the constraint k,

$$d_k^\top \theta - p_k \leq \exp\left(\frac{\beta_n - \mu_k}{\epsilon_n} + \frac{\mu_k - (n_k/n)}{\epsilon_n}\right). \tag{48}$$

The non-penalty difference, $d_k^\top \theta - p_k$, is bounded.

$$(d_k, p_k)^\top (\theta, -1) \leq |(d_k, p_k)^\top (\theta, -1)| \leq ||(d_k, p_k)||_2 ||(\theta, -1)||_2 \leq \mathscr{K} \tag{49}$$

But the penalty term, goes to infinity, is unbounded. So for some n,

$$\mathcal{K} < \exp\left(\frac{\beta_n - \mu_k}{\epsilon_n} + \frac{\mu_k - (n_k/n)}{\epsilon_n}\right) \tag{50}$$

For all constraints k with $\mu_k \leq \beta_n - \iota_n$. Therefore, θ in the constraints with $\mu_k \geq \beta_n + \iota_n$, a compact set, are in the interior of the constraints with $\mu_k \leq \beta_n - \iota_n$. So the constraints with $\mu_k \leq \beta_n - \iota_n$ can not bind (if they did, the constraints with $\mu_k \geq \beta_n + \iota_n$ would be violated). □

A.3. Proof of Lemma 4.

Proof of Lemma 4.

Recall $w_{i,\ell} = h_{i,\ell}\mathbf{1}\left(x_{i,\ell} \geq u_\ell\right)$. By the Cauchy-Schwartz inequality,

$$|\overline{\Sigma}_{W,k}^{-1/2}(w_i - W)| < ||\overline{\Sigma}_{W,k}^{-1/2}||_2 ||w_i - W||_2. \tag{51}$$

So:

$$\begin{aligned}\mathbb{E}\left[||\overline{\Sigma}_W^{-1/2}(w_i - W)||_2^3\right] &= \mathbb{E}\left[\left\{\sum_{k=1}^{\kappa}\left[\overline{\Sigma}_{W,k}^{-1/2}(w_i - W)\right]^2\right\}^{3/2}\right] \leq \mathbb{E}\left[\left\{\sum_{k=1}^{\kappa}||\overline{\Sigma}_{W,k}^{-1/2}||_2^2 ||w_i - W||_2^2\right\}^{3/2}\right] \\ &= \mathbb{E}\left[||w_i - W||_2^3\left\{\sum_{k=1}^{\kappa}||\overline{\Sigma}_{W,k}^{-1/2}||_2^2\right\}^{3/2}\right] = \mathbb{E}\left[||w_i - W||_2^3 ||\overline{\Sigma}_{W,k}^{-1/2}||_F^3\right],\end{aligned} \tag{52}$$

Where $||\overline{\Sigma}_W^{-1/2}||_F$ is the Frobenius norm of $\overline{\Sigma}_W^{-1/2}$. Let $\eta_1, \ldots, \eta_\kappa$ be the eigenvalues of $\overline{\Sigma}_W$. The Frobenius norm is,

$$||\overline{\Sigma}_W^{-1/2}||_F = \sqrt{\sum_{k=1}^{\kappa}\eta_k^{-1}} \leq \sqrt{\kappa}\underline{\eta}^{-1/2}. \tag{53}$$

Where $\underline{\eta}$ is from Assumption 8. So the inequality becomes:

$$\mathbb{E}\left[||\overline{\Sigma}_W^{-1/2}(w_i - W)||_2^3\right] \leq \mathbb{E}\left[||w_i - W||_2^3 ||\overline{\Sigma}_{W,k}^{-1/2}||_F^3\right] \leq \mathbb{E}\left[||w_i - W||_2^3 \kappa^{3/2}\underline{\eta}^{-3/2}\right]. \tag{54}$$

Which is bounded because $\underline{\eta} > 0$ and $\mathbb{E}\left[||w_i - W||_2^3||\right]$ is bounded by Assumption 9. □

A.4. Proof of Lemma 5.

Proof of Lemma 5.

Because of the assumptions on the form of v,

$$||\widehat{v}-v||_{\infty} = \max_{j=1,\ldots,J} \sup_{u} |\frac{1}{n}\sum_{i=1}^{n} h_{j,i}\mathbf{1}\left(x_{j,i}\geq u\right) - \mathbb{E}\left[h_j \mathbf{1}\left(x_j \geq u\right)\right]|. \tag{55}$$

By way of approximation, let $h_{j,i}$ have $M < \infty$ points of support (I drop j for conciseness in the following). Let $q_i = \ell \implies h_i = \sum_{p=1}^{\ell} s_p$ and write,

$$h_i = s_1 \times \mathbf{1}\left(q_i \geq 1\right) + \cdots + s_M \mathbf{1}\left(q_i \geq M\right). \tag{56}$$

Without loss of generality, I assume $0 \leq h \leq 1$.

The following sequence of implications is then true,

$$\sup_{u} |\sum_{p=1}^{\ell} s_p \left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(q_i \geq p, x_i \geq u\right) - \mathbb{E}\left[\mathbf{1}\left(q \geq p, x \geq u\right)\right]\right]| \geq \epsilon \implies \tag{57}$$

$$\sum_{p=1}^{\ell} s_p \times \sup_{u} |\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(q_i \geq p, x_i \geq u\right) - \mathbb{E}\left[\mathbf{1}\left(q \geq p, x \geq u\right)\right]| \geq \epsilon \implies \tag{58}$$

$$\sup_{u,u_q} |\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(q_i \geq u_q, x_i \geq u\right) - \mathbb{E}\left[\mathbf{1}\left(q \geq u_q, x \geq u\right)\right]| \geq \epsilon \tag{59}$$

For any $\epsilon > 0$.

From results on the convergence of the empirical distribution function (see Kiefer 1961) (for any $\delta > 0$; of course, choose $\delta < 2$),

$$\text{pr}\left\{\sup_{u,u_q} |\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(q_i \geq u_q, x_i \geq u\right) - \mathbb{E}\left[\mathbf{1}\left(q \geq u_q, x \geq u\right)\right]| \geq \epsilon\right\} \tag{60}$$

$$\leq C\left(\delta, P+1\right)\exp\left(-(2-\delta)n\epsilon^2\right), \tag{61}$$

Where P is the dimension of x (and the set, $\mathscr{U}$) and $C\left(\delta, P+1\right)$ is a universal constant depending only on P and δ.

So when h_j has M points of support,

$$\text{pr}\left\{\max_{j=1,\ldots,\mathscr{J}} \sup_u \left|\frac{1}{n}\sum_{i=1}^{n} h_{j,i}\mathbf{1}\left(x_{j,i}\geq u\right)-\mathbb{E}\left[h_j\mathbf{1}\left(x_j\geq u\right)\right]\right|\geq\epsilon\right\}\leq \tag{62}$$

$$\sum_{j=1}^{\mathscr{J}}\text{pr}\left\{\sup_u \left|\frac{1}{n}\sum_{i=1}^{n} h_{j,i}\mathbf{1}\left(x_{j,i}\geq u\right)-\mathbb{E}\left[h_j\mathbf{1}\left(x_j\geq u\right)\right]\right|\geq\epsilon\right\}\leq \tag{63}$$

$$J\times C\left(\delta,P+1\right)\exp\left(-(2-\delta)n\epsilon^2\right) \tag{64}$$

The bound does not depend on the number of points of support so we can increase the points of support to infinity and have the same result hold. Because the support of h_j are countable, we have shown the uniform convergence of $||\widehat{\nu}-\nu||_\infty$.

□

A.5. Proof of Theorem 2.

Proof of Theorem 2.

Let $\widehat{V}_n$ be the moments used in the $\widehat{t}_n$ program and V_n be the moments used in the t_n program. Let $||\cdot||_\infty$ be the max-norm. Because $||\widehat{V}_n-V_n||_\infty \xrightarrow{p} 0$, $\widehat{V}_n$ and V_n are close in large samples.

Given that $t_n(V_n)$ is differentiable in a neighborhood of V_n and $\widehat{V}_n$ is in a neighborhood of V_n for large n with probability approaching one,

$$\frac{\sqrt{n}}{\widehat{\sigma}_n}\left[\widehat{t}_n-t_n\right]=\frac{\sqrt{n}}{\widehat{\sigma}_n}\nabla t_n\left(\widetilde{V}_n\right)^\top\left[\widehat{V}_n-V_n\right]. \tag{65}$$

For some $\widetilde{V}_n=V_n+\widehat{s}_n\left(\widehat{V}_n-V_n\right)$ where $\widehat{s}_n\in[0,1]$.

The derivative of $t_n\left(\widetilde{V}_n\right)$ is only non-zero if the constraint the moment belongs to binds. Because the constraint set is a finite number of linear inequalities, there are a maximum number of constraints that can bind, and so, moments that can have a non-zero derivative (maximum number of moments with non-zero derivative is κ_b).

Let $W\subset V$ be a possible basis (the subset of the moments associated with the binding constraints plus the γ moments). Define Σ_W to be the variance-covariance matrix of these moments.

Define the following class of random variables:

$$Y_n(W)=\sqrt{n}\Sigma_W^{-1/2}\left[\widehat{W}_n-W_n\right] \tag{66}$$

Let Φ be the multivariate normal probability measure with zero mean and with a covariance matrix equal to the identity matrix.

Let $\mathscr{E}$ be the set of all convex sets of $\mathbb{R}^{\kappa_b}$. Let $\mathscr{K}_b$ be a constant that depends on κ_b. Gotze (1991) shows the Berry-Essen like inequality,

$$\sup_{E\in\mathscr{E}} |\text{pr}\{Y_n(W)\in E\} - \Phi(E)| \leq \frac{\mathscr{K}_b}{\sqrt{n}} \mathbb{E}\left[||\Sigma_W^{-1/2}(w_i - W_n)||_2^3\right] \tag{67}$$

Define $\mu(W)$ to be the smallest μ_k associated with a moment in W. $\Sigma_W^{-1/2} = \mu(W)^{-1/2}\overline{\Sigma}_W^{-1/2}$. From Lemma 4, there is a $\mathscr{M}$ such that:

$$\mathbb{E}\left[||\overline{\Sigma}_W^{-1/2}[w_i - W_n]||_2^3\right] \leq \mathscr{M} < \infty, \tag{68}$$

Where $\mathscr{M}$ does not depend on W.

Then, we have:

$$\sup_{E\in\mathscr{E}} |\text{pr}\{Y_n(W)\in E\} - \Phi(E)| \leq \frac{\mathscr{K}_b}{\sqrt{n}} \mathbb{E}\left[||\Sigma_W^{-1/2}(w_i - W_n)||_2^3\right] \leq \frac{\mathscr{K}_b}{\sqrt{n}\mu(W)^{3/2}}\mathscr{M} \tag{69}$$

Taking the max of both sides over all possible W tells us nothing because some $\mu(W)$'s are small. But I only need to consider basis that could actually be the true basis. What is in question is whether any of the W's with small $\mu(W)$ can be the true basis of the program.

From Lemma 3, W's with $\mu_k < \beta_n - \iota_n$ can not bind in large samples. So take $\mu(W) \geq \beta_n - \iota_n$ when taking the maximum over possible W's:

$$\sup_{W:\mu(W)\geq\epsilon_n} \sup_{E\in\mathscr{E}} |\text{pr}\{Y_n(W)\in E\} - \Phi(E)| \leq \frac{\mathscr{K}_b}{\sqrt{n}(\beta_n - \iota_n)^{3/2}}\mathscr{M} \tag{70}$$

By Assumption 7: $n(\beta_n - \iota_n)^3 \to \infty$. Then, I have convergence in how the distribution of $Y_n(W)$ measures convex sets versus how Φ measures convex sets.

Let $W_n^\star$ be an optimal basis for the $t_n\left(\widetilde{V}_n\right)$ program. Then,

$$\widetilde{\sigma}_n = \sqrt{\nabla_{W_n^\star} t_n\left(\widetilde{V}_n\right)^\top \Sigma\left(W_n^\star\right) \nabla_{W_n^\star} t_n\left(\widetilde{V}_n\right)}. \tag{71}$$

Where $\nabla_W t_n$ is the derivative of t_n with respect to the W moments.

A class of convex sets:

$$C_{W,u} = \left\{Y : \frac{1}{\widetilde{\sigma}_n}\nabla_W t_n\left(\widetilde{V}_n\right)^\top \Sigma(W)^{1/2} Y \leq u\right\} \tag{72}$$

Let $\mathscr{Y}$ be a random vector with distribution given by Φ,

$$\sup_{W:\mu(W)\geq\epsilon_n} \sup_u |\text{pr}\left\{\frac{\sqrt{n}}{\widetilde{\sigma}_n}\nabla_W t_n\left(\widetilde{V}_n\right)^\top\left(\widehat{W}_n - W_n\right) \leq u\right\} - \text{pr}\left\{\frac{1}{\widetilde{\sigma}_n}\nabla_W t_n\left(\widetilde{V}_n\right)^\top \Sigma(W)^{1/2}\mathscr{Y} \leq u\right\}| \to 0. \tag{73}$$

Let $\mathcal{N}$ be the CDF of the standard normal distribution.

$$\begin{aligned} \nabla t_n\left(\widetilde{V}_n\right)^\top \Sigma(W)^{1/2}\mathscr{Y} \sim N\left(0, \nabla t_n\left(\widetilde{V}_n\right)^\top \Sigma(W)\nabla t_n\left(\widetilde{V}_n\right)\right) \implies \\ \sup_u \left|\text{pr}\left\{\frac{\sqrt{n}}{\widetilde{\sigma}_n}\nabla_{W^\star} t_n\left(\widetilde{V}_n\right)^\top\left(\widehat{W}_n^\star - W_n^\star\right) \le u\right\} - \mathcal{N}(u)\right| = O\left(n^{-1/2}(\beta_n - \iota_n)^{-3/2}\right). \end{aligned} \tag{74}$$

Because $\nabla_{W^\star} t_n\left(\widetilde{V}_n\right)^\top\left(\widehat{W}_n^\star - W_n^\star\right) = \nabla t_n\left(\widetilde{V}_n\right)^\top\left(\widehat{V}_n - V_n\right)$ (derivative of $\widetilde{t}_n$ is zero for constraints that don't bind in the $\widetilde{t}_n$ program) and because $n\epsilon_n^3 \to \infty$, I have shown,

$$\frac{\sqrt{n}}{\widetilde{\sigma}_n}\left[\widehat{t}_n - t_n\right] \xrightarrow{d} N(0,1) \tag{75}$$

$\widetilde{\sigma}_n/\widehat{\sigma}_n \xrightarrow{p} 1$ because $\widehat{\sigma}_n$ is a continuous function of $\widehat{V}_n$ (by the continuity of the derivatives in V_n) such that $\widehat{\sigma}_n(V_n) = \sigma_n$ and $||\widehat{V}_n - V_n||_\infty \xrightarrow{p} 0$. So I have:

$$\frac{\sqrt{n}}{\widehat{\sigma}_n}\left[\widehat{t}_n - t_n\right] \xrightarrow{d} N(0,1) \tag{76}$$

□

A.6. Proof of Theorem 3.

Proof of Theorem 3.

Consider the finite linear program which only uses constraints with $\mu(u)$ greater than $\beta_n + \iota_n$ and does not use the penalty function. Call the value of this program, $\underline{t}_n$.

The penalty function for sets with $\mu(u) \ge \beta_n + \iota_n$ goes to zero (from Lemma 3).

Call the program which only uses constraints with μ_k greater than $\beta_n + \iota_n$ *and* uses the penalty function, $\overline{t}_n$. $\overline{t}_n$ is smaller than t_n because fewer constraints are used (in large samples).

The difference between $\overline{t}_n$ and $\underline{t}_n$ becomes small because the penalty function goes to zero and the value of the program is continuous in the right hand side coefficients (because it is differentiable in them).

$$\overline{t}_n = \underbrace{\overline{t}_n - \underline{t}_n}_{\text{Small in large samples}} + \underline{t}_n \tag{77}$$

From Karney (1981), the value of any discretization (a selection of a finite number of the constraints) of a countable semi-infinite linear program (a linear program with a finite number of choices and a countable number of constraints) approaches the value of the semi-infinite program if the feasible set of the semi-infinite program is non-empty and bounded.

Because $\underline{t}_n$ is a non-random discretization of t, we have:

$$\underline{t}_n \to t. \tag{78}$$

I know:

$$\overline{t}_n \leq t_n \leq t. \tag{79}$$

Where $\overline{t}_n \leq t_n$ is true because we are using fewer constraints in the $\overline{t}_n$ program by requiring $\mu(u) \geq \beta_n + \iota_n$.

Because $\overline{t}_n - \underline{t}_n + \underline{t}_n \approx \underline{t}_n$ in large samples and $\underline{t}_n \to t$, I have $\overline{t}_n \to t$, which implies $t_n \to t$.

□

A.7. Proof of Theorem 4.

Proof of Theorem 4.

First, I establish the semi-infinite program is discretizable using a theorem from Reemtsen (1991).

Given Assumption 3, the feasible set in program t_n is compact for large enough n. Given Assumption 5, the feasible set of the semi-infinite program t is non-empty. Given Assumption 6, the set of points $B = \cup_u (d(u), p(u))$ is compact and so is each truncation of the coefficients, $B_K = \cup_{u \in \mathscr{U}_K} (d(u), p(u))$.

The Hausdorff distance (with the max-norm) between B_K and B goes to zero by premise (2) for which premise (1) is sufficient.

Proof that (1) is sufficient: without loss of generality, assume $0 \in B_K$ (a redundant constraint). Then,

$$\lim_K d_H(B_K, B) = \lim_K \sup_{x \in B} \inf_{y \in B_K} ||x - y||_\infty \leq \lim_K \sup_{x \in B \setminus B_K} ||x||_\infty = \lim_K \sup_{u \notin \mathscr{U}_K} ||(d(u), p(u))||_\infty = 0.$$

The set defined by $\cup_{k=1}^K \left\{\theta : d_k^\top \theta \leq p_k\right\}$ is compact for large enough K (Assumption 3).

Therefore, the premises of Reemtsen (1991)'s theorem are satisfied and the program is discretizable.

Let $\overline{t}_n$ (only uses constraints with $\mu \geq \beta_n + \iota_n$ and includes the penalty function) and $\underline{t}_n$ (only uses constraints with $\mu \geq \beta_n + \iota_n$ but with no penalty function) be as in Theorem 3.

Because the program is discretizable,

$$\underline{t}_n \to t. \tag{80}$$

Following the proof of Theorem 3, we have:

$$\overline{t}_n \leq t_n \leq t \implies \overline{t}_n - \underline{t}_n + \underline{t}_n \leq t_n \leq t. \tag{81}$$

Because $\left(\overline{t}_n - \underline{t}_n\right) \to 0$, the left hand side converges to t. So:

$$t_n \to t. \tag{82}$$

□

A.8. Proof of Theorem 5.

Proof of Theorem 5.

Let μ be the smallest $\mu(u)$ of any binding constraint in the program,

$$t = \min_{\theta} t(\theta, \gamma) \quad \text{ST:} \quad d(u)^\top \theta \leq p(u) \quad \text{for } u \in \mathscr{U}_K. \tag{83}$$

For large n, $K_n > K$, $\epsilon_n < \mu$, and $\beta_n < \mu$. Write, the difference $\widehat{t}_n - t$ as,

$$\frac{\sqrt{n}}{\widehat{\sigma}_n} \times \left[\widehat{t}_n - t\right] = \frac{\sqrt{n}}{\widehat{\sigma}_n} \times \left[\widehat{t}_n - t_n\right] + \frac{\sqrt{n}}{\widehat{\sigma}_n} \times \left[t_n - t\right]. \tag{84}$$

Let ϕ_n be the vector of penalty terms and let $\nabla_\phi \widetilde{t}_n$ be the derivative of a program with respect to ϕ. The second term converges in probability to zero because, by the mean value theorem,

$$\frac{\sqrt{n}}{\widehat{\sigma}_n} \times \left[t_n - t\right] = \frac{\sqrt{n}}{\widehat{\sigma}_n} \nabla_\phi \widetilde{t}_n^\top \phi_n \tag{85}$$

The derivative is only non-zero for at most κ_b constraints. The penalty term is eventually largest for the constraint with $\mu(u) = \mu$. If $n\epsilon_n \to \infty$, then the largest penalty term of a binding constraint is at most, in large samples,

$$\exp\left(\frac{\beta_n}{\epsilon_n} - \frac{1}{\epsilon_n}\mu + o_p(1)\right). \tag{86}$$

The result holds because then $\left(n_k/n - \mu_k\right)/\epsilon_n \to 0$ for each k and we do not need to worry about errors in estimating non-binding (in the true program) constraints because, even if their penalty term was zero, in large samples, they would not bind.

Scaling the penalty term by $\sqrt{n}/\widehat{\sigma}_n$ makes it,

$$\exp\left(\frac{\beta_n}{\epsilon_n} + \frac{1}{2}\log n - \log \widehat{\sigma}_n - \frac{1}{\epsilon_n}\mu\right) \tag{87}$$

$\widehat{\sigma}_n$ does not go to zero because the covariance matrix is bounded away from zero. What is important is that,

$$\frac{\beta_n}{\epsilon_n} + \frac{1}{2}\log n - \frac{1}{\epsilon_n}\mu \to -\infty. \tag{88}$$

If $\epsilon_n \log n \to 0$, then the above statement is true because,

$$\frac{\frac{\beta_n}{\epsilon_n} + \frac{1}{2}\log n}{\frac{\mu}{\epsilon_n}} = \frac{1}{\mu} \times \left[\beta_n + \frac{1}{2}\epsilon_n \log n\right] \to 0, \tag{89}$$

So, $\mu/\epsilon_n \to \infty$ faster than $\beta_n/\epsilon_n + \frac{1}{2}\log n$.

So, by Theorem 2,

$$\frac{\sqrt{n}}{\widehat{\sigma}_n} \times \left[\widehat{t}_n - t\right] = \frac{\sqrt{n}}{\widehat{\sigma}_n} \times \left[\widehat{t}_n - t_n\right] + \frac{\sqrt{n}}{\widehat{\sigma}_n} \times \left[t_n - t\right] = \frac{\sqrt{n}}{\widehat{\sigma}_n} \times \left[\widehat{t}_n - t_n\right] + o_p(1) \xrightarrow{d} N(0,1). \tag{90}$$

□

A.9. Proof of Theorem 6.

Proof of Theorem 6.

Starting from the application of Gotze (1991) inequality in the proof of Theorem 2, for each $G \in \mathscr{G}$,

$$\sup_{W:\mu(W)\geq\epsilon_n} \sup_u \sup_{G'} |\text{pr}_G\left\{\frac{\sqrt{n}}{\widetilde{\sigma}_n^{G'}}\left(\nabla \widetilde{t}_n^{G'}\right)^\top \left(w_{i,n}^G - W_n^G\right) \leq u\right\} - \mathscr{N}(u)| \leq \mathscr{K}_b \mathscr{M}_G \times \frac{1}{\sqrt{n}(\beta_n - \iota_n)^{3/2}}. \tag{91}$$

Where G' is another data generating process in $\mathscr{G}$ and where $\mathscr{M}_G$ is the upper bound on the moment $\mathbb{E}_G\left[||\overline{\Sigma}_W^{-1/2}\left(\widehat{W} - W\right)^3||\right]$. Taking the sup over $G \in \mathscr{G}$ on both sides of the inequality gives,

$$\sup_{G\in\mathscr{G}} \sup_{W:\mu(W)\geq\epsilon_n} \sup_u \sup_{G'} |\text{pr}_G\left\{\frac{\sqrt{n}}{\widetilde{\sigma}_n^{G'}}\nabla \widetilde{t}_n^{G'\top}\left(\widehat{W}_n - W_n\right) \leq u\right\} - \mathscr{N}(u)| \leq \mathscr{K}_b \mathscr{M} \times \frac{1}{\sqrt{n}(\beta_n - \iota_n)^{3/2}}. \tag{92}$$

So we have convergence in distribution uniformly over all data generating processes in $\mathscr{G}$, implying:

$$\begin{aligned} &\sup_G |\text{pr}_G\left\{t_n \geq \widehat{t}_n - \frac{\widehat{\sigma}_n}{\sqrt{n}}\Phi^{-1}(\alpha)\right\} - \alpha| \to 0 \\ \implies &\inf_G \text{pr}\left\{t_n \geq \widehat{t}_n - \frac{\widehat{\sigma}_n}{\sqrt{n}}\Phi^{-1}(\alpha)\right\} \geq \alpha + j_n \end{aligned} \tag{93}$$

Where $j_n \to 0$.

□

A.10. Proof of Lemma 6.

Proof of Lemma 6.

Define $t^u = -\min_\theta -t(\theta,\gamma) \quad \text{ST:} \quad d_k^\top \theta \leq p_k \quad \text{for } k = 1,2,\ldots.$

From Theorem 2, we know that for any potentially binding basis of t^u, W^u, and for any potentially binding basis of t^ℓ, W^ℓ,

$$\begin{aligned} \sup_{E\in\mathscr{E}} |\text{pr}\left\{\sqrt{n}\Sigma_{W^\ell}^{-1/2}\left[\widehat{W}_n^\ell - W^\ell\right]\in E\right\} - \Phi(E)| &\leq \frac{\mathscr{K}_b}{\sqrt{n}(\beta_n-\iota_n)^{3/2}}\mathscr{M} \\ \sup_{E\in\mathscr{E}} |\text{pr}\left\{\sqrt{n}\Sigma_{W^u}^{-1/2}\left[\widehat{W}_n^u - W^u\right]\in E\right\} - \Phi(E)| &\leq \frac{\mathscr{K}_b}{\sqrt{n}(\beta_n-\iota_n)^{3/2}}\mathscr{M}, \end{aligned} \tag{94}$$

Define $W = \left(W^\ell, W^u\right)$ and let Σ_W be the covariance matrix of $\sqrt{n}\widehat{W}$. Let $\mathscr{E}$ be the set of all convex sets. Applying the Gotze (1991) inequality gives,

$$\sup_{E\in\mathscr{E}} |\text{pr}\left\{\sqrt{n}\Sigma_W^{-1/2}\left[\widehat{W}_n - W\right]\in E\right\} - \Phi(E)| \leq \frac{\mathscr{K}'_b}{\sqrt{n}(\beta_n-\iota_n)^{3/2}}\mathscr{M}. \tag{95}$$

Define $t = \left(t^u, t^\ell\right)$. Consider the class of convex sets:

$$C_u = \left\{y : \widehat{S}^{-1/2}\nabla_W \widetilde{t}^\top \Sigma_W^{1/2} y \leq u\right\}. \tag{96}$$

Where $\widetilde{t} = t\left(\widetilde{V}_n\right)$, a vector of moments between the true measure and the estimated measure, as in the proof of Theorem 2.

The Gotze (1991) inequality then gives:

$$\sup_u |\text{pr}\left\{\sqrt{n}\widehat{S}^{-1/2}\left[\widehat{t}_n - t_n\right]\leq u\right\} - \Phi(C_u)| \leq \frac{\mathscr{K}'_b}{\sqrt{n}(\beta_n-\iota_n)^{3/2}}\mathscr{M}. \tag{97}$$

So the difference between the distribution of $\sqrt{n}\left[\widehat{t}_n - t_n\right]$ and a bi-variate normal distribution with variance-covariance matrix $\left(\nabla_W \widehat{t}_n^\top \widehat{\Sigma}_W \nabla_W \widehat{t}_n\right)^{-1/2} \nabla_W \widetilde{t}^\top \Sigma_W \nabla_W \widetilde{t} \left(\nabla_W \widehat{t}_n^\top \widehat{\Sigma}_W \nabla_W \widehat{t}_n\right)^{-1/2}$.

By the continuous differentiability of $\widehat{t}_n$ and because $||\widehat{V}_n - V_n|| \xrightarrow{p} 0$ (from Theorem 2, the variance-covariance matrix approaches the identity matrix.

□