Tutorial and applications in econometrics: Self-normalization, Gaussian approximation, and inference with many moment inequalities

Kengo Kato (U. of Tokyo)

May 15, 2014@NUS

Schedule

- Part 1a: Introduction to partially identified models.
- Part 1b: Inference with many moment inequalities.
- Part 1c: Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors.

Most of the lectures are based upon the following joint works with Victor Chernozhukov (MIT) and Denis Chetverikov (UCLA):

- Testing many moment inequalities. (2013). arXiv:1312.7614.
- Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. arXiv:1212.6906. Ann. Statist. (2013).
- Comparison and anti-concentration bounds for maxima of Gaussian random vectors. arXiv:1301.4807. To appear in Probab. Theory Related Fields. (2014).

Preliminary

- In preparing the paper "Testing many moment inequalities", we found that the self-normalization theory is useful for inference with (many) moment inequalities, a class of partially identified models in which people in econometrics have had a lot of interests in the last decade.
- There we studied inference procedures based upon (i) a moderate deviation inequality for self-normalized sums, (ii) (a version of) high-dimensional central limit theorem (developed in the last two papers), and (iii) combination of (i) and (ii).

Preliminary (cont.)

- Part 1b will cover the content of "Testing" paper.
- Part 1a will cover a brief introduction to inference for partially identified models, which hopefully motivates Part 1b.
- Part 1c will cover the content of the remaining two papers, on which some results in Part 1b rely (the material in Part 1c has only small connection to self-normalization theory, but hopefully you will find it intriguing).

Part 1a: Introduction to partially identified models

What is partially identified model?

- Romano and Shaikh (Econometrica, 2010, p.169):

  > *A partially identified model is any model in which the parameter of interest is not uniquely defined by the distribution of the observed data.*

- The model only restricts the value of the parameter of interest to a (multi-element) set, called *identified set*.

- Partially identified models frequently appear in economic applications, where you typically encounter the following situation: you are interested in the parameter in the latent structure, but the observed data does not contain enough information to give you point identification of the parameter.

Example 1: interval data

- Suppose you have a r.v. Y which is *unobservable*, but there are observable r.v.'s Y_1, Y_2 that bracket Y in the sense that

$$Y_1 \leq Y \leq Y_2.$$

- For example:
  1. (interval censoring) Let Y be the income. Some surveys only record brackets of income, say, $(y_0, y_1], \dots, (y_{K-1}, y_K]$. Defining

  $$Y_1 = y_{k-1}, Y_2 = y_k, \text{ when } Y \in (y_{k-1}, y_k], \ k = 1, \dots, K,$$

  we have $Y_1 \leq Y \leq Y_2$.
  2. (missing data) Let $Y \in [0,1]$ and $D \in \{0,1\}$, and we only observe Y when $D = 1$, so the observe variable is the pair (D, DY). Then

  $$DY \leq Y \leq DY + (1 - D).$$

Example 1: interval data (cont.)

- The parameter of interest is $\theta = \mathbf{E}[Y]$.
- However, without additional information, you can only know from the data that θ satisfies the restriction:

$$\mathbf{E}[Y_1] \le \theta \le \mathbf{E}[Y_2].$$

- In this case, the identified set is the closed interval

$$\big[\mathbf{E}[Y_1], \mathbf{E}[Y_2]\big].$$

Example 2: regression with interval outcomes

- Keep the setting in Example 1, but suppose there exists a regressor $\boldsymbol{X}$ in $\mathbb{R}^{\boldsymbol{d}}$, and the conditional mean $\mathbf{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ is a linear function of $\boldsymbol{X}$, i.e., $\mathbf{E}[\boldsymbol{Y} \mid \boldsymbol{X}] = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is the parameter of interest.
- Consider the simple case where the distribution of $\boldsymbol{X}$ is discrete:

$$\mathbf{P}(\boldsymbol{X} \in \{\boldsymbol{x}(\mathbf{1}), \ldots, \boldsymbol{x}(\boldsymbol{J})\}) = \mathbf{1}.$$

- Then the identified set is

$$\{\boldsymbol{\theta} : \mathbf{E}[\boldsymbol{Y_1} \mid \boldsymbol{X} = \boldsymbol{x}(\boldsymbol{j})] \leq \boldsymbol{x}(\boldsymbol{j})^{\boldsymbol{T}}\boldsymbol{\theta} \leq \mathbf{E}[\boldsymbol{Y_2} \mid \boldsymbol{X} = \boldsymbol{x}(\boldsymbol{j})], \forall \boldsymbol{j}\}.$$

Example 3: entry model

- Based on Ciliberto and Tamer (2009, Econometrica).
- Let m denote the number of firms that could potentially enter the market; let m-tuple $D = (D_1, \ldots, D_m)$ denote the observed entry decisions of these firms; that is, $D_j = 1$ if the firm j enters the market and $D_j = 0$ otherwise. Let $\mathcal{D} = \{0, 1\}^m$.
- Let X and ε denote (exogenous) characteristics of the market as well as characteristics of the firms that are observed and not observed by the researchers, respectively.
- Profit of the firm j is given by

$$\pi_j(D, X, \varepsilon, \theta),$$

where π_j is known up to θ which is the parameter of interest.

- Assume that both $\boldsymbol{X}$ and ε are observed by the firms and that a Nash equilibrium is played, so that for each j,

$$\pi_j((D_j, D_{-j}), X, \varepsilon, \theta) \geq \pi_j((1 - D_j, D_{-j}), X, \varepsilon, \theta).$$

- Then there exist sets $R_1(d, X, \theta)$ and $R_2(d, X, \theta)$ for ε such that 1) $D = d$ is the unique equilibrium whenever $\varepsilon \in R_1(d, X, \theta)$; 2) $D = d$ is one of several equilibria whenever $\varepsilon \in R_2(d, X, \theta)$.

- When $\varepsilon \in R_1(d, X, \theta)$ for some $d \in \mathcal{D}$, we know for sure that $D = d$ but when $\varepsilon \in R_2(d, X, \theta)$, the probability that $D = d$ depends on the equilibrium selection mechanism, and, without further information, can be anything in $[0, 1]$.
- Hence

$$
\begin{aligned}
\mathrm{P}(\varepsilon \in R_1(d, X, \theta) \mid X) &\leq \mathrm{P}(D = d \mid X) \\
&\leq \mathrm{P}(\varepsilon \in R_1(d, X, \theta) \mid X) + \mathrm{P}(\varepsilon \in R_2(d, X, \theta) \mid X).
\end{aligned}
$$

- $\geq 2^{m+1}$ inequalities.

Example 4: CRS (2013, Quant. Econ.)

- Based on Chesher, Rosen, Smolinski (2013, Quant. Econ.).
- An individual is choosing an alternative out of options in $\mathcal{D}$.
- Let D denote his choice; let X denote characteristics of the individual that are observed by the researcher; and let V denote characteristics of the individual that are *not* observed by the researcher.
- Choosing an alternative $d \in \mathcal{D}$ yields the utility

$$u(d, X, V).$$

- The individual is maximizing his utility, so that

$$u(D, X, V) \geq u(d, X, V), \ \forall d \in \mathcal{D}.$$

The object of interest is the pair (u, P_V) where P_V denotes the distribution of the vector V.

- A complication arises because in many applications, X may be endogenous (not independent of V); hence assume that there exists a vector Z of instruments that is related with X but independent of V.
- To generate moment inequalities, let $\tau(d, X, u)$ denote the set for V such that $D = d$ whenever $V \in \tau(d, X, u)$, so that

$$V \in \tau(D, X, u).$$

- Since $V \in \tau(D, X, u)$, we have that for any set S,

$$\mathrm{P}(V \in S) = \mathrm{P}(V \in S \mid Z) \geq \mathrm{P}(\tau(D, X, u) \subset S \mid Z),$$

  so that for each S, we have a conditional moment inequality.
- Then the question is "how to choose a class of sets S to sharply identify (u, P_V)?"
- CRS proved that it suffices to consider all unions of sets on the support of $\tau(D, X, u)$. When X is discrete with the support consisting of m points, this gives $|\mathcal{D}| \cdot 2^m$ sets.
- Chesher and Rosen (2013) provide a more general framework called *Generalized Instrumental Variable* model.

- In many examples, partially identified models can be represented as moment inequality models.
- Let ξ be a r.v. taking values in a measurable space $(S, \mathcal{S})$ with distribution P, let Θ be an ambient parameter space which is B-measurable subset of a metric space (usually subset of a Euclidean space), and let $g = (g_1, \dots, g_p)^T : S \times \Theta \to \mathbb{R}^p$ be a B-measurable map.
- Then the identified set is assumed to be

$$\Theta_I = \Theta_I(P) = \{\theta \in \Theta : \mathrm{E}_P[g_j(\xi, \theta)] \leq 0, \ 1 \leq \forall j \leq p\}.$$

- i.i.d. data $\xi_1, \dots, \xi_n \sim P$ are available.
- We will keep this setting in what follows.

Inference on what?

There may be two possibilities.

- The entire identified set Θ_I — we want to construct a stochastic subset $\mathcal{C}_n(\alpha) \subset \Theta$ based on the data $\xi_1, \dots, \xi_n$ such that

$$\mathrm{P}(\Theta_I \subset \mathcal{C}_n(\alpha)) \geq 1 - \alpha. \quad \text{(or approximately)}$$

- Any particular $\theta \in \Theta_I$ — we want to construct $\mathcal{C}_n(\alpha)$ such that

$$\inf_{\theta \in \Theta_I} \mathrm{P}(\theta \in \mathcal{C}_n(\alpha)) \geq 1 - \alpha. \quad \text{(or approximately)}$$

  The CR for the latter is generally smaller than the former. Probably more suitable when there is a "true parameter".

- We will focus on the latter problem in the next lecture when p is possibly large ($p = p_n \to \infty$); in this lecture we assume p *is fixed.*

Inference on Θ_I

CHT approach

- Based on Chernozhukov, Hong, Tamer (2007,Ecoometrica).
- For a given $p \times p$ positive definite matrix $W(\theta)$, consider

$$Q(\theta) = Q(\theta, P) = (\mathrm{E}_P[g(\xi, \theta)])_+^T W(\theta) (\mathrm{E}_P[g(\xi, \theta)])_+,$$

  where $((x_1, \ldots, x_p)^T)_+ = (\max\{x_1, 0\}, \ldots, \max\{x_p, 0\})^T$.
- Then

$$\theta \in \Theta_I \Leftrightarrow Q(\theta) = 0.$$

- Define the sample analogue of $Q(\theta)$ by

$$\hat{Q}(\theta) = \left(\frac{1}{n} \sum_{i=1}^{n} g(\xi_i, \theta) \right)_+^T W(\theta) \left(\frac{1}{n} \sum_{i=1}^{n} g(\xi_i, \theta) \right)_+.$$

Consistent estimation of Θ_I

- A lower contour set $C_n(c)$ of level c of $\hat{Q}$ is defined by

$$C_n(c) = \{\theta \in \Theta : \hat{Q}(\theta) \leq c/n\}.$$

- The estimator for Θ_I will take of the form

$$\hat{\Theta}_I = C_n(c_n),$$

where $c_n \uparrow \infty$ slowly; CHT suggested $c_n = \log n$ (c_n could be 0 for some examples but not generally so).

Rates of convergence of $\hat{\Theta}_I$ in Hausdorff distance

- Denote by $d(\cdot,\cdot)$ the metric on Θ; then the Hausdorff distance between subsets in Θ is defined by

$$d_H(A,B) = \max\left\{\sup_{a\in A} d(a,B), \sup_{b\in B} d(b,A)\right\},$$

where $d(a,B) = \inf_{b\in B} d(a,b)$.

- CHT proved that, (when Θ is a subset of a Euclidean space),

$$d_H(\hat{\Theta}_I, \Theta_I) = O_P(\sqrt{\max(c_n,1)/n}),$$

(of course) subject to suitable regularity conditions.

Inference on Θ_I

- Idea:

$$\Theta_I \subset C_n(c) \Leftrightarrow \sup_{\theta \in \Theta_I} n\hat{Q}(\theta) \le c.$$

- Hence by taking

$$c_{1-\alpha} = (1-\alpha)\text{-quantile of } \sup_{\theta \in \Theta_I} n\hat{Q}(\theta),$$

we have

$$\mathrm{P}(\Theta_I \subset C_n(c_{1-\alpha})) \ge 1-\alpha.$$

- Critical value $c_{1-\alpha}$ can be approximated by
  1. subsampling applied with Θ_I replaced by $\hat{\Theta}_I$; or
  2. simulating the limit distribution of $\sup_{\theta \in \Theta_I} n\hat{Q}(\theta)$.

Some other references

Beresteanu and Molinari (2008), Bugni (2010, Econometrica), Romano and Shaikh (2010), and Kaido (2012)...

Inference on $\theta \in \Theta_I$

Duality

- We may exploit duality between construction of confidence sets for any fixed $\theta \in \Theta_I$ and testing the hypothesis

$$H_\theta : \mathrm{E}_P[g_j(\xi, \theta)] \leq 0, 1 \leq \forall j \leq p,$$

against

$$H'_\theta : \mathrm{E}_P[g_j(\xi, \theta)] > 0, 1 \leq \exists j \leq p.$$

- To fix idea: suppose there is a test statistic $T_n(\theta)$ for testing H_θ v.s. H'_θ, and denote by $R_{n,\alpha}(\theta)$ any rejection region with size α, i.e.,

$$\mathrm{P}(T_n(\theta) \in R_{n,\alpha}(\theta)) \leq \alpha,$$

whenever H_θ is true (i.e., $\theta \in \Theta_I$). Then the CR

$$\mathcal{C}_n(\alpha) = \{\theta : T_n(\theta) \notin R_{n,\alpha}(\theta)\}$$

contains θ with probability at least $1 - \alpha$ whenever $\theta \in \Theta_I$.

- Hence the problem boils down to testing the following multivariate one-sided problem (with composite null hypothesis): let $X_1, \dots, X_n$ be i.i.d. random vectors in $\mathbb{R}^p$ with mean $\mu = (\mu_1, \dots, \mu_p)^T = \mathrm{E}[X_1]$, and consider testing

$$H_0 : \mu_j \le 0, 1 \le \forall j \le p, \text{ v.s. } H_1 : \mu_j > 0, 1 \le \exists j \le p.$$

- Closely related to classical multivariate one-sided tests where the null is simple — Kudo (1963, Biometrika), Perlman (1969, Ann. Math. Statist.) etc.

- Suppose $\Sigma = \mathbf{E}[(X_1 - \mu)(X_1 - \mu)^T]$ is non-singular.
- Consider the test of the form

$$T_n := \min_{t \in \mathbb{R}^p_-} n(\bar{X} - t)^T \Sigma^{-1} (\bar{X} - t) > c \Rightarrow \text{ reject } H_0,$$

  where $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ and $\mathbb{R}^p_- = \{t \in \mathbb{R}^p : t_j \leq 0, \ \forall j\}$.
- When $\Sigma = I$, $T_n = n|(\bar{X})_+|^2$.
- We need to choose c such that

$$\sup_{\mu_j \leq 0, \forall j} \mathrm{P}(T_n > c) \leq \alpha + o(1).$$

- By simple algebra,

$$T_n = \min_{t \in K} |\sqrt{n}\Sigma^{-1/2}\bar{X} - t|^2,$$

where $K = \Sigma^{-1/2}\mathbb{R}^p_-$ (polyhedral cone).

- Denote by K° its polar cone:

$$K^\circ = \{t \in \mathbb{R}^p : t^T s \le 0, \forall s \in K\}.$$

Then

$$T_n = |\mathbf{Proj}_{K^\circ}\sqrt{n}\Sigma^{-1/2}\bar{X}|^2.$$

Close look at K°

- Since, for $e_j = (0, \dots, \underbrace{1}_{j\text{th}}, \dots, 0)^T$,

$$K = \{t \in \mathbb{R}^p : (\Sigma^{1/2} e_j)^T t \le 0, \ \forall j\},$$

  the polar cone K° is expressed as

$$K^\circ = \left\{ \sum_{j=1}^p \lambda_j \Sigma^{1/2} e_j : \lambda_j \ge 0 \right\}.$$

- Proof: Use $(K^\circ)^\circ = K$.
- When Σ is diagonal, $K^\circ = \{t \in \mathbb{R}^p : t_j \ge 0, \forall j\}$.

- Observe that, whenever $\mu_j \leq 0, \forall j$,

$$T_n \leq \min_{t \in \mathbb{R}^p_-} n(\bar{X} - \mu - t)^T \Sigma^{-1} (\bar{X} - \mu - t)$$

$$= |\mathbf{Proj}_{K^\circ} \sqrt{n} \Sigma^{-1/2} (\bar{X} - \mu)|^2 =: T'_n$$

and the equality takes place when $\mu_j = 0, \forall j$, so that

$$\sup_{\mu_j \leq 0, \forall j} \mathrm{P}(T_n > c) = \mathrm{P}(T'_n > c).$$

- Recall that the projection onto a closed convex set is a contraction.
- Hence by CLT and the continuous mapping theorem,

$$\boldsymbol{T_n' \xrightarrow{d} |\mathbf{Proj}_{K^\circ} Z|^2,}$$

  where $\boldsymbol{Z \sim N_p(0, I_p)}$.
- Possible to simulate the limit distribution.
- Bootstrap may be used to approximate the distribution of $\boldsymbol{T_n'}$ (but not $\boldsymbol{T_n}$).

Failure of bootstrap to approximate the distribution of T_n

- Due to Andrews (2000, Econometrica).
- Consider $p = 1, \Sigma = 1$, so that under H_0,

$$T_n = n(\bar{X})_+^2 \xrightarrow{d} \begin{cases} 0 & \mu < 0 \\ (N(0,1))_+^2 & \mu = 0. \end{cases}$$

- Consider $\mu = 0$. Let $X_1^*, \dots, X_n^*$ be i.i.d. draws from the e.d. of $\{X_1, \dots, X_n\}$. Then

$$T_n^* = n(\bar{X}^*)_+ = \left(\sqrt{n}(\bar{X}^* - \bar{X}) + \sqrt{n}\bar{X}\right)_+^2.$$

- As $\sqrt{n}\bar{X} \xrightarrow{d} N(0,1)$, $\mathrm{P}(\sqrt{n}\bar{X} > 1) = \mathrm{P}(N(0,1) > 1) + o(1)$; on the event $\sqrt{n}\bar{X} > 1$,

$$T_n^* \geq \left(\sqrt{n}(\bar{X}^* - \bar{X}) + 1\right)_+^2.$$

  Moreover, conditional on $X_1, X_2, \dots$, for a.e. realizations of $X_1, X_2, \dots$,

$$\text{right side} \xrightarrow{d} (N(0,1) + 1)_+^2.$$

- Hence, with probability $\mathrm{P}(N(0,1) > 1) + o(1)$,

$$\begin{aligned} &\text{conditional } 0.95\text{-quantile of } T_n^* \\ &\qquad \geq 0.95\text{-quantile of } (N(0,1) + 1)_+^2 - o(1). \end{aligned}$$

Comments

- Replace $\boldsymbol{\Sigma}$ by $\hat{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T$ in practice. Validity follows immediately(?).
- Rosen actually proposed to bounding quantiles of the limit distribution as

$$\mathbf{P}(|\mathbf{Proj}_{K^\circ} Z|^2 > c) \leq \frac{1}{2}\mathbf{P}(\chi_p^2 > c) + \frac{1}{2}\mathbf{P}(\chi_{p-1}^2 > c),$$

but this will lead to more conservative CRs.

Alternative approaches

- Subsampling applied to T_n. See Romano and Shaikh (2008, J. Stat. Plan. Infer.) and Andrews and Guggenberger (2009, Econometric Theory).
- Incorporating moment selection. Exclude j such that $\bar{X}_j$ is negatively small when calculating critical values. See Andrews and Soares (2010, Econometrica), Andrews and Jia Barwick (2012, Econometrica), Romano, Shaikh, Wolf (2014, Econometrica) etc.
- Other test statistics: (in addition to already mentioned references) Canay (2010, J. Econometrics), Chernozhukov, Chetverikov, K. (2013), etc.

Multiple hypothesis testing

- The problem of testing moment inequalities discussed so far is related but different from the multiple hypothesis testing problem: $H_{0j} : \mu_j \le 0$ v.s. $H_{1j} : \mu_j > 0, \ j = 1, \dots, p$.
- In testing moment inequalities, we try to control

$$\sup_{H_{0j}, 1 \le j \le p} \mathrm{P}(\text{at least one of } H_{0j}, 1 \le j \le p, \text{ is rejected}) \le \alpha,$$

  and improve the power when some of inequalities are not binding ($\mu_j < 0$) by moment selection.
- In multiple hypothesis testing, we typically try to control

$$\max_{J \subset \{1,\dots,p\}} \sup_{H_{0j}, j \in J} \mathrm{P}(\text{at least one of } H_{0j}, j \in J, \text{ is rejected}) \le \alpha.$$