Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors

Victor Chernozhukov (MIT), Denis Chetverikov (UCLA), and Kengo Kato (U. of Tokyo)

March 14, 2014

- This talk is based upon the paper:
  Chernozhukov, V., Chetverikov, D. and K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786-2819
- Applications to moment inequality models are based on an ongoing paper.

Introduction

- Let $x_1, \ldots, x_n$ be independent random vectors in $\mathbb{R}^p, p \geq 2$.
- $\mathbf{E}[x_i] = 0$ and $\mathbf{E}[x_i x_i']$ exists. $\mathbf{E}[x_i x_i']$ may be degenerate.
- (Important!) Possibly $p \gg n$. Keep in mind $p = p_n$.
- This paper is about approximating the distribution of

$$T_0 = \max_{1 \leq j \leq p} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{ij}.$$

- By making

$$x_{i,p+1} = -x_{i1}, \ldots, x_{i,2p} = -x_{ip},$$

we have

$$\max_{1 \leq j \leq p} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{ij} \right| = \max_{1 \leq j \leq 2p} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{ij}.$$

Introduction

- Let $y_1, \ldots, y_n$ be independent normal random vectors with

$$y_i \sim N(0, \mathbf{E}[x_i x_i']).$$

- Define

$$Z_0 = \max_{1 \le j \le p} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} y_{ij}.$$

- *When p is fixed*, (subject to the Lindeberg condition) the central limit theorem guarantees that

$$\sup_{t \in \mathbb{R}} |\mathbf{P}(T_0 \le t) - \mathbf{P}(Z_0 \le t)| \to 0.$$

Introduction

- Basic question: How large $p = p_n$ can be while having

$$\sup_{t \in \mathbb{R}} |\mathrm{P}(T_0 \le t) - \mathrm{P}(Z_0 \le t)| \to 0?$$

- Related to multivariate CLT with growing dimension (Portnoy, 1986, PTRF; Götze, 1991, AoP; Bentkus, 2003, JSPI, etc.).
- Write

$$X = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i, \; Y = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} y_i.$$

They are concerned with conditions under which

$$\sup_{A \in \mathcal{A}} |\mathrm{P}(X \in A) - \mathrm{P}(Y \in A)| \to 0,$$

while allowing for $p = p_n \to \infty$.

Introduction

- Bentkus (2003) proved that (in case of i.i.d. and $\mathbf{E}[x_i x_i'] = I$),

$$\sup_{A:\text{convex}} |\mathbf{P}(X \in A) - \mathbf{P}(Y \in A)| = O(p^{1/4}\mathbf{E}[|x_1|^3]n^{-1/2}).$$

  Typically $\mathbf{E}[|x_1|^3] = O(p^{3/2})$, so that the RHS$=o(1)$ provided that

$$p = o(n^{2/7}).$$

- The main message of the paper: to make

$$\sup_{t\in\mathbb{R}} |\mathbf{P}(T_0 \le t) - \mathbf{P}(Z_0 \le t)| \to 0,$$

  p *can be much larger*. Subject to some conditions,

$$\log p = o(n^{1/7})$$

  will suffice.

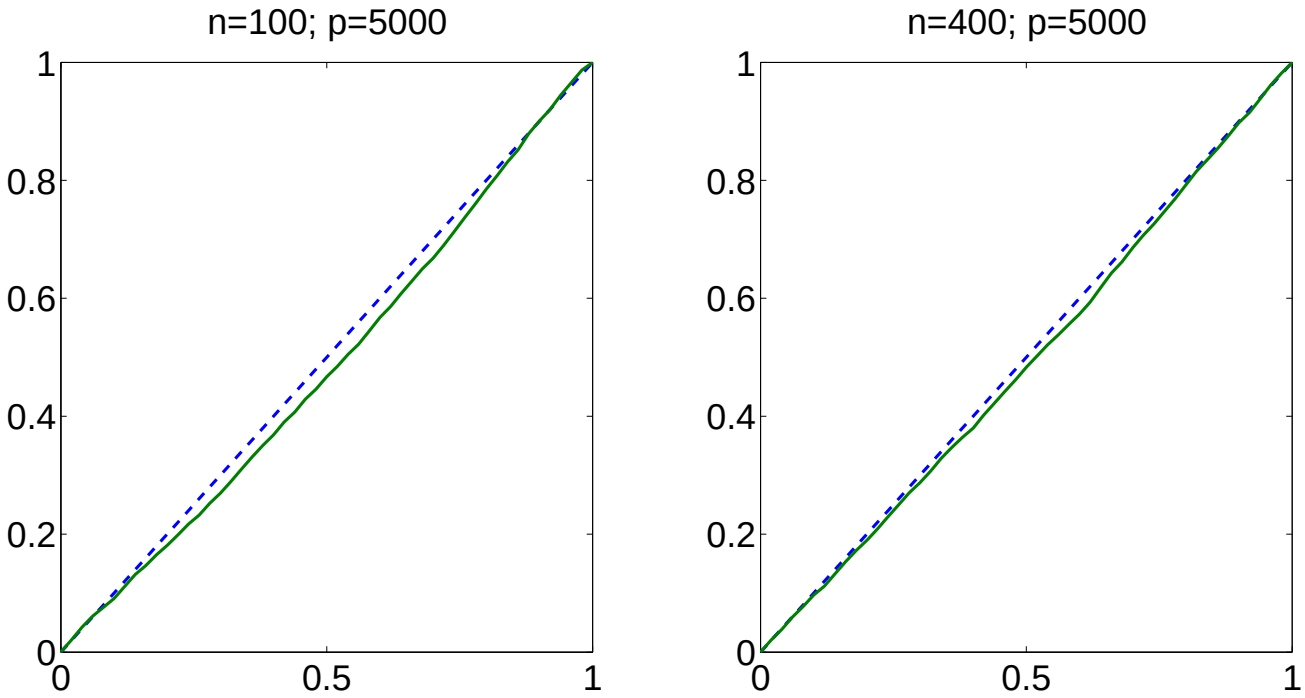

Figure : P-P plots comparing distributions of T_0 and Z_0 in the example motivated by the problem of selecting the penalty level of the Dantzig selector. Here x_{ij} are generated as $x_{ij} = z_{ij}\varepsilon_i$ with $\varepsilon_i \sim t(4)$, (a t-distribution with four degrees of freedom), and z_{ij} are non-stochastic (simulated once using $U[0,1]$ distribution independently across i and j). The dashed line is 45°. The distributions of T_0 and Z_0 are close, as (qualitatively) predicted by the GAR derived in the paper. The quality of the Gaussian approximation is particularly good for the tail probabilities, which is most relevant for practical applications.

Introduction

- Still the above approximation results are not directly usable unless the cov. structure between the coordinates in $\boldsymbol{X}$ is unknown.
- In some cases, we know the cov. structure. e.g. think of $\boldsymbol{x_i = \varepsilon_i z_i}$ where ε_i is a scalar (error) r.v. with mean zero and common variance, and $\boldsymbol{z_i}$ is the vector of non-stochastic covariates. Then $\boldsymbol{T_0}$ is the maximum of $\boldsymbol{t}$-statistics.
- But usually not. In such cases the dist. of $\boldsymbol{Z_0}$. is unknown.
- $\Rightarrow$ We propose a Gaussian multiplier bootstrap for approximating the dist. of $\boldsymbol{T_0}$ when the cov. structure between the coordinates of $\boldsymbol{X}$ is unknown. Its validity is established through the Gaussian approximation results. Still $\boldsymbol{p}$ can be much larger than $\boldsymbol{n}$.

Applications

- Selecting design-adaptive tuning parameters for Lasso (Tibshirani, 1996, JRSSB) and Dantzig selector (Candès and Tao, 2007, AoS).
- Multiple hypotheses testing (too many references).
- Adaptive specification testing. These three applications are examined in the arXiv paper.
- Testing *many* moment inequalities. Will be treated if time allowed.

Literature

- Classical CLTs with $p = p_n \to \infty$: Portnoy (1986, PTRF), Götze (1991, AoP), Bentkus (2003, JSPI), among many others.
- Modern approaches on multivariate CLTs: Chatterjee (2005, arXiv),Chatterjee and Meckes (2008, ALEA), Reinert and Röllin (2009, AoP), Röllin (2011,AIHP). Developing Stein's methods for normal approximation. Harsha, Klivans, and Meka (2012, J.ACM).
- Bootstrap in high dim.: Mammen (1993, AoS), Arlot, Blanchard, and Roquain (2010a,b, AoS).

Main Thm.

Theorem

Suppose that there exists const. $0 < c_1 < C_1$ *s.t.* $c_1 \le n^{-1}\sum_{i=1}^n \mathbf{E}[x_{ij}^2] \le C_1,\ 1 \le \forall j \le p$. *Then*

$$\sup_{t\in\mathbb{R}} |\mathbf{P}(T_0 \le t) - \mathbf{P}(Z_0 \le t)|$$
$$\le C \inf_{\gamma\in(0,1)} \Big[n^{-1/8}(M_3^{3/4} \vee M_4^{1/2}) \log^{7/8}(pn/\gamma) + n^{-1/2} Q(1-\gamma) \log^{3/2}(pn/\gamma) + \gamma \Big],$$

where $C = C(c_1, C_1) > 0$. *Here*

$Q(1-\gamma) = (1-\gamma)$-*quantile of* $\max_{i,j} |x_{ij}| \vee (1-\gamma)$-*quantile of* $\max_{i,j} |y_{ij}|$

and $M_k = \max_{1\le j\le p}(n^{-1}\sum_{i=1}^n \mathbf{E}[|x_{ij}|^k])^{1/k}$.

Comments

- No restriction on correlation structure.
- The extra parameter γ appears essentially to avoid the appearance of the term of the form

$$\mathbf{E}[\max_{1\le j\le p}|x_{ij}|^k]$$

  in the bound. Notice the difference from M_k.
- To avoid this, we use a suitable truncation, and γ controls the level of truncation.

Techniques

- There are a lot of techniques used to prove the main thm.
- Directly bounding the probability difference $(\mathbf{P}(T_0 \le t) - \mathbf{P}(Z_0 \le t))$ is difficult. Transform the problem into bounding

$$\mathbf{E}[g(X) - g(Y)], \ g\text{: smooth},$$

  where $X = n^{-1/2} \sum_{i=1}^n x_i, Y = n^{-1/2} \sum_{i=1}^n y_i$.
- How? Approximate $z = (z_1, \ldots, z_p)' \mapsto \max_{1 \le j \le p} z_j$ by

$$F_\beta(z) = \beta^{-1} \log(\sum_{j=1}^p e^{\beta z_j}).$$

  Then $0 \le F_\beta(z) - \max_{1 \le j \le p} z_j \le \beta^{-1} \log p$.

Techniques

- Approximate the indicator function $\mathbf{1}(\cdot \le t)$ by a smooth function h (standard). Then take $g = h \circ F_\beta$.
- Use a variant of Stein's method to bound

$$\mathbf{E}[g(X) - g(Y)]. \qquad (*)$$

  Truncation + some fine properties of F_β are used here.
- To obtain a bound on the probability difference from (*), we need an anti-concentration ineq. for maxima of normal random vectors.
- Intuition: from (*), we will have a bound on

$$\mathbf{P}(T_0 \le t) - \mathbf{P}(Z_0 \le t + \text{error}).$$

  Want to replace $\mathbf{P}(Z_0 \le t + \text{error})$ by $\mathbf{P}(Z_0 \le t)$.

Simplified anti-concentration ineq.

Lemma (Simplified form)

Let $(Y_1, \dots, Y_p)'$ *be a normal random vector with* $\mathbf{E}[Y_j] = 0$ *and* $\mathbf{E}[Y_j^2] = 1$ *for all* $1 \le j \le p$. *Then* $\forall \epsilon > 0$,

$$\sup_{t \in \mathbb{R}} \mathbf{P}(|\max_{1 \le j \le p} Y_j - t| \le \epsilon) \le 4\epsilon(\mathbf{E}[\max_{1 \le j \le p} Y_j] + 1).$$

This bound is universally tight (up to constant).

Note 1: $\mathbf{E}[\max_{1 \le j \le p} Y_j] \le \sqrt{2 \log p}$.
Note 2: The inequality is *dimension-free*: Easy to extend it to separable Gaussian processes.

Some consequences

Assumption: *either*

$$\text{(E.1)} \quad \mathbf{E}[\exp(|x_{ij}|/B_n)] \leq 2, \forall i, j; \text{ or}$$

$$\text{(E.2)} \quad (\mathbf{E}[\max_{1 \leq j \leq p} x_{ij}^4])^{1/4} \leq B_n, \ \forall i.$$

Moreover, assume *both*

$$\text{(M.1)} \quad c_1 \leq n^{-1} \sum_{i=1}^n \mathbf{E}[x_{ij}^2] \leq C_1, \ \forall j; \text{ and}$$

$$\text{(M.2)} \quad n^{-1} \sum_{i=1}^n \mathbf{E}[|x_{ij}|^{2+k}] \leq B_n^k, \ k = 1, 2, \forall j.$$

Here $B_n \to \infty$ is allowed. e.g. consider the case where $x_i = \varepsilon_i z_i$ with ε_i mean zero scalar error and z_i vector of non-stochastic covariates normalized s.t. $n^{-1} \sum_{i=1}^n z_{ij}^2 = 1, \ \forall j$. Then (E.2),(M.1),(M.2) are satisfied if

$$\mathbf{E}[\varepsilon_i^2] \geq c_1, \ \mathbf{E}[\varepsilon_i^4] \leq C_1, \ |z_{ij}| \leq B_n, \ \forall i, j,$$

after adjusting constants.

Corollary

Corollary

Suppose that one of the following conditions is satisfied:

(i) *(E.1) and* $B_n^2 \log^7(pn) \le C_1 n^{1-c_1}$*; or*

(ii) *(E.2) and* $B_n^4 \log^7(pn) \le C_1 n^{1-c_1}$*.*

Moreover, suppose that (M.1) and (M.2) are satisfied. Then

$$\sup_{t \in \mathbb{R}} |\mathrm{P}(T_0 \le t) - \mathrm{P}(Z_0 \le t)| \le C n^{-c},$$

where c, C *depend only on* c_1, C_1*.*

Multiplier bootstrap

- Unless the cov. structure of X is known, the dist. of Z_0 is still unknown. Propose a multiplier bootstrap.
- Generate i.i.d. $N(0,1)$ r.v.'s $e_1, ..., e_n$ indep. of $x_1, ..., x_n$. Define

$$W_0 = \max_{1 \le j \le p} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_i x_{ij}.$$

- Note that cond. on $x_1, \ldots, x_n$,

$$n^{-1/2} \sum_{i=1}^{n} e_i x_i \sim N(0, n^{-1} \sum_{i=1}^{n} x_i x_i').$$

"Close" to $N(0, n^{-1} \sum_{i=1}^{n} \mathrm{E}[x_i x_i']) \stackrel{d}{=} Y$. Recall $Z_0 = \max_{1 \le j \le p} Y_j$.

- Bootstrap critical value:

$$c_{W_0}(\alpha) = \inf\{t \in \mathbb{R} : \mathrm{P}_e(W_0 \le t) \ge \alpha\}.$$

Theorem (Multiplier bootstrap theorem)

Suppose that one of the following conditions is satisfied:

(i) *(E.1) and* $B_n^2 \log^7(pn) \le C_1 n^{1-c_1}$*; or*

(ii) *(E.2) and* $B_n^4 \log^7(pn) \le C_1 n^{1-c_1}$*.*

Moreover, suppose that (M.1) and (M.2) are satisfied. Then

$$\sup_{\alpha \in (0,1)} |\mathrm{P}(T_0 \le c_{W_0}(\alpha)) - \alpha| \le C n^{-c},$$

where c, C *depend only on* c_1, C_1*.*

Key fact

The key to the above theorem is the fact that

$$\sup_{t\in\mathbb{R}} |\mathbf{P}_e(W_0 \le t) - \mathbf{P}(Z_0 \le t)|$$

is essentially controlled by

$$\max_{1\le j,k\le p} |n^{-1}\textstyle\sum_{i=1}^n (x_{ij}x_{ik} - \mathbf{E}[x_{ij}x_{ik}])|,$$

which can be $o_P(1)$ even if $p \gg n$.

Testing many moment inequalities

- $x_1, \ldots, x_n \sim$ i.i.d. in $\mathbb{R}^p$ with $\mathbf{E}[x_i] = \mu$. Assume $\sigma_j^2 = \mathrm{Var}(x_{ij}) > 0, \ \forall j$.
- Possibly $p \gg n$. Think of $p = p_n$.
- We are interested in testing the null hypothesis

$$H_0 : \mu_j \leq 0, \ \forall j,$$

against the alternative

$$H_1 : \mu_j > 0, \ \exists j.$$

Literature on testing moment inequalities

- Testing unconditional moment inequalities: Chernozhukov, Hong, and Tamer (2007, ECMT), Romano and Shaikh (2008, JSPI), Andrews and Guggenberger (2009, ET), Andrews and Soares (2010, ECMT), Canay (2010, JoE), Bugni (2011, working), Andrews and Jia-Barwick (2012, ECMT), Romano, Shaikh, and Wolf (2012, working). # of moment ineq. is *fixed*.
- Testing conditional moment inequalities: Andrews and Shi (2013, ECMT), Chernozhukov, Lee, and Rosen (2013, ECMT), Armstrong (2011, working), Chetverikov (2011, working), Armstrong and Chan (2012, working).
- When *many* moment inequalities?: Entry game example in Ciliberto and Tamer (2009, ECMT), testing conditional moment inequalities in Andrews and Shi (2013, ECMT).

Test statistic and MB critical value

- Def. $\hat{\mu}_j = n^{-1} \sum_{i=1}^n x_{ij}$ and $\hat{\sigma}_j^2 = n^{-1} \sum_{i=1}^n (x_{ij} - \hat{\mu}_j)^2$.
- Test stat.

$$T = \max_{1 \le j \le p} \sqrt{n} \hat{\mu}_j / \hat{\sigma}_j.$$

- Under H_0,

$$T \le \max_{1 \le j \le p} \sqrt{n} (\hat{\mu}_j - \mu_j) / \hat{\sigma}_j.$$

  Want to approximate the distribution of the RHS.

- Generate i.i.d. $N(0,1)$ r.v.'s $e_1, \ldots, e_n$ indep. of the data. Def.

$$W = \max_{1 \le j \le p} \frac{1}{\sqrt{n}} \sum_{i=1}^n e_i (x_{ij} - \hat{\mu}_j) / \hat{\sigma}_j,$$

$$c_W(1-\alpha) = \text{conditional } (1-\alpha)\text{-quantile of } W.$$

Refinement by moment selection

- Take $0 < \beta_n < \alpha/2$. $\beta_n \to 0$ is allowed but $\sup_{n \geq 1} \beta_n < \alpha/2$.
- Take

$$\hat{J} = \{j \in \{1, \ldots, p\} : \hat{\mu}_j \geq -2c_W(1 - \beta_n)/\sqrt{n}\}.$$

Def.

$$W_R = \max_{j \in \hat{J}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_i(x_{ij} - \hat{\mu}_j)/\hat{\sigma}_j,$$

$c_{W_R}(1 - \alpha) =$ conditional $(1 - \alpha + 2\beta_n)$-quantile of W_R.

Size control

Theorem

Define $z_{ij} = (x_{ij} - \mu_j)/\sigma_j$ *and* $z_i = (z_{i1}, \ldots, z_{ip})'$. *Suppose that (E.2) and (M.2) are satisfied with* $x_i = z_i$. *Then*

$$\mathrm{P}(T > c_W(1-\alpha)) \le \alpha + Cn^{-c},$$
$$\mathrm{P}(T > c_{W_R}(1-\alpha)) \le \alpha + Cn^{-c}, \ (\text{if } \log(1/\beta_n) \le C_1 \log n).$$

Moreover, if all the inequalities are binding and $\beta_n \le C_1 n^{-c_1}$, *then*

$$\mathrm{P}(T > c_W(1-\alpha)) \ge \alpha - Cn^{-c},$$
$$\mathrm{P}(T > c_{W_R}(1-\alpha)) \ge \alpha - Cn^{-c}.$$

Here c, C *depend only on* c_1, C_1.

Summary

- We derived a new Gaussian approximation result for the maximum of the sum of high-dimensional random vectors which is valid even when $p \gg n$ and without any restriction on correlation structure between coordinates of the random vectors.
- We proved validity of the Gaussian multiplier bootstrap.
- We demonstrated usefulness of the results for testing many moment inequalities.
- The results presented here have many other applications as well.

Thank you for your attention.