Tutorial Part 1c: Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors

Kengo Kato (U. of Tokyo)

May 16, 2014@NUS

This part is based upon the papers:

- Chernozhukov, V., Chetverikov, D. and K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. arXiv:1212.6906. Ann. Statist.
- Chernozhukov, V., Chetverikov, D. and K. (2014). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. arXiv:1301.4807. To appear in Probab. Theory Related Fields

Some other papers related to this lecture:

- Chernozhukov, V., Chetverikov, D., and K. (2014). Gaussian approximation of suprema of empirical processes. arXiv:1212.6885. To appear in Ann. Statist.
- Chernozhukov, V., Chetverikov, D., and K. (2013). Anti-concentration and honest, adaptive confidence bands. arXiv:1303:7152. Revised and resubmitted to Ann. Statist.

Introduction

- Let $x_1, \ldots, x_n$ be independent random vectors in $\mathbb{R}^p, p \geq 2$.
- $\mathbf{E}[x_i] = 0$ and $\mathbf{E}[x_i x_i^T]$ exists. $\mathbf{E}[x_i x_i^T]$ may be degenerate.
- (Important!) Possibly $p \gg n$. Keep in mind $p = p_n$.
- Consider 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

$$y_i \sim N(0, \mathrm{E}[x_i x_i^T]).$$

- 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

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

Introduction

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

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

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

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

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 $\mathbf{E}[x_i x_i^T] = I$),

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

  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.

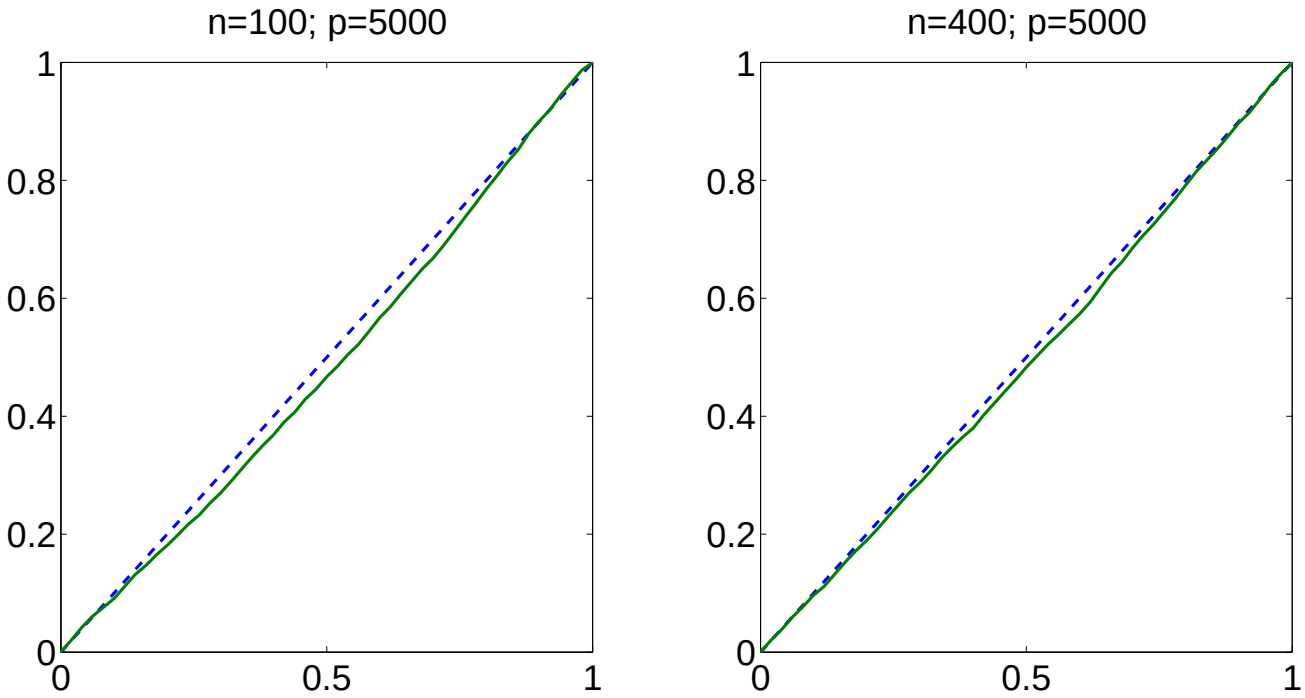

Figure: P-P plots comparing distributions of $\boldsymbol{T_0}$ and $\boldsymbol{Z_0}$ in the example motivated by the problem of selecting the penalty level of the Dantzig selector. Here $\boldsymbol{x_{ij}}$ are generated as $\boldsymbol{x_{ij} = z_{ij}\epsilon_i}$ with $\boldsymbol{\epsilon_i \sim t(4)}$, (a $\boldsymbol{t}$-distribution with four degrees of freedom), and $\boldsymbol{z_{ij}}$ are non-stochastic (simulated once using $\boldsymbol{U[0,1]}$ distribution independently across $\boldsymbol{i}$ and $\boldsymbol{j}$). The dashed line is 45°. The distributions of $\boldsymbol{T_0}$ and $\boldsymbol{Z_0}$ 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 applicable unless the covariance structure between the coordinates in $\boldsymbol{X}$ is unknown.
- In some cases, we know the covariance structure. E.g. think of $\boldsymbol{x_i} = \varepsilon_i \boldsymbol{z_i}$ 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 distribution of $\boldsymbol{Z_0}$ is unknown.
- $\Rightarrow$ We propose a Gaussian multiplier bootstrap for approximating the distribution of $\boldsymbol{T_0}$ when the covariance 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 (Candes 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 theorem

Theorem

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

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

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

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

Comments

- The constant C depends only on c_1, C_1, lower and upper bounds on *coordinate-wise* variances. 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

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

- To avoid this, we use a suitable truncation (which actually replies on self-normalization), 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 $z = (z_1, \ldots, z_p)^T \mapsto \max_{1 \le j \le p} z_j$ by

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

  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 AC inequality)

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

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

This bound is universally tight (up to constant).

Note 1: $\mathbf{E}[\max_{1 \leq j \leq p} Y_j] \leq \sqrt{2 \log p}$.
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.

Example

Consider e.g. 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 covariance structure of $\boldsymbol{X}$ is known, the distribution of $\boldsymbol{Z_0}$ is still unknown.
- To handle this case, consider a multiplier bootstrap.
- Generate i.i.d. $\boldsymbol{N(0,1)}$ r.v.'s $\boldsymbol{e_1, ..., e_n}$ independent of $\boldsymbol{x_1, ..., x_n}$.
- Define

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

- Want to approximate the distribution of $\boldsymbol{T_0}$ by the *conditional* distribution of $\boldsymbol{W_0}$ given $\boldsymbol{x_1, \ldots, x_n}$.

Multiplier bootstrap (cont.)

- Note that conditional on $x_1, \ldots, x_n$,

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

  "Close" to $N(0, n^{-1} \sum_{i=1}^n \mathrm{E}[x_i x_i^T]) \stackrel{d}{=} Y$.
- 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 $p \gg n$.

Comparison of Gaussian maxima

More formally, the above theorem is deduced from the following comparison result for Gaussian maxima.

Lemma

Let $\boldsymbol{V \sim N_p(0, \Sigma^V)}$ *and* $\boldsymbol{Y \sim N_p(0, \Sigma^Y)}$*. Suppose that there exists* $\boldsymbol{0 < c_1 < C_1}$ *such that* $\boldsymbol{c_1 \le \Sigma^Y_{jj} \le C_1, 1 \le \forall j \le p}$*. Then*

$$\sup_{t \in \mathbb{R}} |\mathbf{P}(\max_{1 \le j \le p} V_j \le t) - \mathbf{P}(\max_{1 \le j \le p} Y_j \le t)| \le C \Delta_0^{1/3} (1 \vee \log(p/\Delta_0))^{2/3},$$

where $\boldsymbol{C = C(c_1, C_1) > 0}$ *and*

$$\Delta_0 := \max_{1 \le j,k \le p} |\Sigma^V_{jk} - \Sigma^Y_{jk}|.$$

Maximal inequality

Lemma

Let $x_1, \ldots, x_n$ be independent random vectors in $\mathbb{R}^p$ with $p \geq 2$. Define $\sigma^2 := \max_{1 \leq j \leq p} \sum_{i=1}^n \mathrm{E}[x_{ij}^2]$. Then

$$\mathrm{E}[\max_{1 \leq j \leq p} |\textstyle\sum_{i=1}^n (x_{ij} - \mathrm{E}[x_{ij}])|] \lesssim \sigma \sqrt{\log p} + \sqrt{\mathrm{E}[\max_{i,j} |x_{ij}|^2]} \log p.$$

Proof sketch of anti-concentration inequality

Recall:

Lemma (Simplified AC inequality)

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

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

Proof sketch-1

- The problem boils down to bounding the density of $M := \max_{1 \le j \le p} Y_j$.
- The crucial observation is that the density of M can be written as

$$\phi(t)G(t),$$

where the map $t \mapsto G(t)$ is non-decreasing.

- Hence

$$\underbrace{\int_t^\infty \phi(r)dr}_{=1-\Phi(t)} G(t) \le \int_t^\infty \phi(r)G(r)dr = \mathbf{P}(M > t),$$

that is

$$G(t) \le \frac{\mathbf{P}(M > t)}{1 - \Phi(t)}.$$

Proof sketch-2

- The Gaussian concentration inequality (due to Borell-Sudakov-Tsirelson) states that

$$\mathbf{P}(M > \mathbf{E}[M] + t) \leq e^{-t^2/2}.$$

- Then the density of M is bounded by

$$\underbrace{\frac{\phi(t)}{1 - \Phi(t)}}_{\leq 2(t \vee 1) \because \text{Mill's ineq.}} e^{-(t-\mathbf{E}[M])_+^2/2} \leq 2(t \vee 1) e^{-(t-\mathbf{E}[M])_+^2/2}.$$

- The right side is

$$\leq 2(\mathbf{E}[M] + 1).$$

Digression-1

Lemma

Let $X(t), t \in T$ be a separable, centered Gaussian process indexed by a semimetric space T such that $\mathbf{E}[X^2(t)] = 1, \forall t \in T$. Suppose that $\sup_{t \in T} X(t)$ is finite a.s. (which ensures that $\mathbf{E}[\sup_{t \in T} X(t)]$ exists and is finite). Then

$$\sup_{x \in \mathbb{R}} \mathbf{P}(|\sup_{t \in T} X(t) - x| \leq \epsilon) \leq 4\epsilon(\mathbf{E}[\sup_{t \in T} X(t)] + 1).$$

Digression-2

- Jian Ding, Ronen Eldan, Alex Zhai (arXiv:1311.5592) made an interesting observation related to the AC inequality.
- That is,

$$(\mathbf{Var}(\max_{1\le j\le p} Y_j))^{1/2}(\mathbf{E}[\max_{1\le j\le p} Y_j]+1) \ge \frac{3}{32}.$$

  Proof: Markov's inequality + AC.
- In words,

  a good concentration for the supremum implies that the expected supremum has to be large