Gaussian Approximation, Moderate Deviation and High Dimensional Hypothesis Testing

Wen-Xin Zhou

Department of Mathematics and Statistics
University of Melbourne

ASC-IMS 2014
Sydney, Australia

Outline

1. Univariate Gaussian approximation: Berry-Esseen bound and Cramér-type moderate deviation
2. Cramér-type moderate deviations for Student's t-statistic (one- and two-sample)
3. Multivariate Gaussian approximation
4. Application: High dimensional test of mean vectors under general covariance heterogeneity

1. Univariate Gaussian approximation: Berry-Esseen bound and Cramér-type moderate deviation

Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variable with zero mean and unit variance. Denote by

$$W_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$$

the normalized partial sum. Let $Z \sim N(0,1)$, the standard normal distribution.

1. Univariate Gaussian approximation: Berry-Esseen bound and Cramér-type moderate deviation

Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variable with zero mean and unit variance. Denote by

$$W_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$$

the normalized partial sum. Let $Z \sim N(0,1)$, the standard normal distribution.
It is well-known that under the Lindeberg condition, the Central Limit Theorem holds:

$$P(W_n \leq x) \to P(Z \leq x) := \Phi(x) \quad \text{for every } x \in \mathbb{R} \quad \text{as } n \to \infty.$$

Moreover,

$$\sup_{x \in \mathbb{R}} \left| P(W_n \leq x) - P(Z \leq x) \right| \to 0 \quad \text{as } n \to \infty.$$

- Question: What is the error of this approximation?

- Two types of errors spark general interest:
  - Absolute error: Berry-Esseen-type bound

$$\left|P(W_n \leq x) - \Phi(x)\right| = \text{ error}$$

  - Relative error

$$\frac{P(W_n \geq x)}{1 - \Phi(x)} = 1 + \text{ error}$$

For the relative error, the Cramér-type moderate deviation addresses the following question: what is the largest possible a_n, such that

$$\frac{P(W_n \geq x)}{1 - \Phi(x)} \to 1$$

holds uniformly in $x \in [0, a_n]$.

- Uniform Berry-Esseen bounds
  - Optimal rate of convergence

$$\sup_{x\in\mathbb{R}} \left|P(W_n \le x) - \Phi(x)\right|$$
$$\lesssim \frac{1}{n}\sum_{i=1}^{n} EX_i^2 I\{|X_i| > \sqrt{n}\} + \frac{1}{n^{3/2}}\sum_{i=1}^{n} E|X_i|^3 I\{|X_i| \le \sqrt{n}\}.$$

  - If $E|X_i|^3 < \infty$,

$$\sup_{x\in\mathbb{R}} \left|P(W_n \le x) - \Phi(x)\right| \lesssim \frac{1}{n^{3/2}}\sum_{i=1}^{n} E|X_i|^3.$$

Proof strategies: Lindeberg's method (swapping technique); Analytical method (Characteristic function); Stein's method (Chen, Goldstein and Shao, 2010), etc.

From sample mean to non-linear statistics

Write $\xi_i = X_i/\sqrt{n}$, such that $W_n = \sum_{i=1}^n \xi_i$. Note that

$$EW_n = 0, \quad EW_n^2 = 1.$$

Let $T_n := T(X_1, \ldots, X_n)$ be a general sampling statistic, which in many cases, can be written as a linear statistic plus an error term, say

$$T_n = W_n + \Delta_n.$$

Typical examples include U-statistics (one- and multi-sample), L-statistics, random sums, etc.

- Berry-Esseen bound for T_n: A rough estimate

  Suppose that $E|\Delta_n|^r < \infty$ for some $r \geq 1$. Let

  $$\mathcal{E}_{n,\delta} = \{|\Delta_n| \leq \delta\}$$

  for some $\delta > 0$.

For every $z \in \mathbb{R}$,

$$
\begin{aligned}
&P(W_n + \Delta_n \le z) - \Phi(z) \\
&= P(W_n + \Delta_n \le z,\, \mathcal{E}_{n,\delta}) - \Phi(z) + P(W_n + \Delta_n \le z,\, \mathcal{E}_{n,\delta}^c) \\
&\le P(W_n \le z + \delta) - \Phi(z + \delta) + \Phi(z + \delta) - \Phi(z) + P(|\Delta_n| > \delta) \\
&\le P(W_n \le z + \delta) - \Phi(z + \delta) + \delta + \delta^{-r} E|\Delta_n|^r.
\end{aligned}
$$

Letting $\delta = \left(E|\Delta_n|^r\right)^{1/r}$ gives us

$$
\begin{aligned}
&\sup_{z \in \mathbb{R}} \left| P(W_n + \Delta_n \le z) - \Phi(z) \right| \\
&\le \sup_{z \in \mathbb{R}} \left| P(W_n \le z) - \Phi(z) \right| + C \left(E|\Delta_n|^r\right)^{1/(r+1)}.
\end{aligned}
$$

Usually the above argument will not lead to an optimal rate.

- Concentration inequality approach (anti-concentration inequality)
  1. Lévy concentration function (Rudelson and Vershynin, 2009) Let ξ be a real-valued r.v. and $\varepsilon > 0$. Define

$$\mathcal{L}(\xi, \varepsilon) = \sup_{x \in \mathbb{R}} P\big(|\xi - x| \leq \varepsilon\big).$$

  2. Randomized concentration inequality (Chen and Shao, 2007) Let $\gamma = n^{-3/2} \sum_{i=1}^{n} E|X_i|^3$. Then

$$\mathcal{L}(W_n, \varepsilon) \leq 2\sqrt{2}\,\varepsilon + 2(\sqrt{2}+1)\gamma.$$

Let $\eta_n = \eta(X_1, \ldots, X_n)$ be a positive random variable. Then

$$\mathcal{L}(W_n, \eta_n) \lesssim \gamma + E|W_n \cdot \eta_n| + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E|X_i(\eta_n - \eta_n^{(i)})|,$$

whenever X_i is independent of $(W - \xi_i, \eta_n^{(i)})$. For example, we may take

$$\eta_n^{(i)} = \eta(X_1, \ldots, X_{i-1}, 0, X_{i+1}, \ldots, X_n).$$

▶ Studentized non-linear statistics

Nonlinear statistics are building blocks in various statistical inference problems. Many of them can be written as a partial sum plus a negligible term, for example, U-statistics (one- and multi-sample), L-statistics, random sums and functions of linear statistics.

As the standardized statistics often involve some unknown nuisance parameters, the Studentized analogues are most commonly used in practice.

Let $\xi_1, \ldots, \xi_n$ be independent random variables with $E\xi_i = 0$, $0 < E\xi_i^2 < \infty$ satisfying $\sum_{i=1}^n E\xi_i^2 = 1$. Write

$$W_n = \sum_{i=1}^n \xi_i, \quad V_n^2 = \sum_{i=1}^n \xi_i^2,$$

and let $\Delta_{n,1}$ and $\Delta_{n,2}$ be measurable functions of $\xi_1, \ldots, \xi_n$.

Recall that the the non-linear statistic of interest is $T_n = W_n + \Delta_{n,1}$, while its Studentized version can be written as

$$\widehat{T}_n = \frac{W_n + \Delta_{n,1}}{V_n(1 + \Delta_{n,2})^{1/2}}.$$

In particular, when $\Delta_{n,1} = \Delta_{n,2} = 0$, $\widehat{T}_n = W_n/V_n$ becomes the self-normalized sum, which is closely related to Student's t-statistic.

- Examples: Studentized U-statistics (one- and two-sample), Studentized L-statistics
- Uniform Berry-Esseen bounds
  1. Wang, Q., Jing, B.-Y. and Zhao, L. The Berry-Esseen bound for Studentized statistics. *Ann. Probab.* **28** 511–535.
  2. Shao, Q.-M., Zhang, K. and Zhou, W.-X. (2014). Stein's method for nonlinear statistics: A brief survey and recent progress. *Technical report.*
- Cramér-type moderate deviations
  1. Lai, T. L., Shao, Q.-M. and Wang, Q. (2011). Cramér type moderate deviations for Studentized U-statistics. *ESAIM: P& S* **15** 168–179.
  2. Shao, Q.-M. and Zhou, W.-X. (2014). Cramér type moderate deviation theorems for self-normalized processes. *arXiv:1405.1218.*
  3. Chang, J., Shao, Q.-M. and Zhou, W.-X. (2014). Cramér-type moderate deviations for Studentized two-sample U-statistics with applications. *Technical report*.

2. Cramér-type moderate deviations for Student's t-statistic

- One-sample case

Let $X_1, \ldots, X_n$ be independent real-valued random variables with mean μ and variance $\sigma^2 > 0$. Let

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2$$

be the sample mean and sample variance, respectively. Moreover, write

$$S_n = \sum_{i=1}^{n} Z_i, \quad V_n^2 = \sum_{i=1}^{n} Z_i^2 \text{ with } Z_i = (X_i - \mu)/\sigma.$$

It is known that Student's t-statistic and the self-normalized sum are closely related; that is,

$$\sqrt{n}(\widehat{\mu}_n - \mu)/\widehat{\sigma}_n = T_n/\sqrt{1 - T_n^2/n},$$

where $T_n = S_n/V_n$.

- Self-normalized moderate deviations for independent r.v.'s:

  Let $X_1, X_2, \ldots$ be independent random variables with $EX_i = 0$. Write

$$S_n = \sum_{i=1}^{n} X_i, \quad V_n^2 = \sum_{i=1}^{n} X_i^2.$$

  The self-normalized moderate deviations describe the rate of convergence of the relative error of $P(S_n/V_n \geq x)$ to $1 - \Phi(x)$.

  The corresponding results with first and second order accuracies were established in Jing, Shao and Wang (2003) (first order accuracy under finite third moments) and Wang (2005, 2011) (second order accuracy under finite fourth moments).

Write

$$B_n^2 = \sum_{i=1}^{n} EX_i^2, \quad L_{kn} = B_n^{-k} \sum_{i=1}^{n} E|X_i|^k, \; k \geq 3.$$

Theorem (Jing, Shao and Wang, 2003)

If $X_1, X_2, \ldots$ are independent r.v.'s with $EX_i = 0$ and $0 < E|X_i|^3 < \infty$, then

$$P(S_n/V_n \geq x)/\{1 - \Phi(x)\} = 1 + O(1)(1+x)^3 L_{3n}$$

for $0 \leq x \leq L_{3n}^{-1/3}$, where $|O(1)|$ is bounded by an absolute constant.

Theorem (Wang, 2011)

If $X_1, X_2, \ldots$ are independent r.v.'s with $EX_i = 0$ and $0 < EX_i^4 < \infty$, then

$$P(S_n/V_n \geq x)/\{1 - \Phi(x)\}$$
$$= \exp\left(-\frac{x^3}{3} \sum_{i=1}^{n} EX_i^3 \right) \left[1 + O(1)\{(1+x)L_{3n} + (1+x)^4 L_{4n}\} \right]$$

for $0 \leq x \leq C^{-1} L_{4n}^{-1/4}$, where $|O(1)|$ is bounded by an absolute constant.

Theorem (Wang, 2011)

If $X_1, X_2, \ldots$ are independent r.v.'s with $EX_i = 0$ and $0 < EX_i^4 < \infty$, then

$$P(S_n/V_n \geq x)/\{1 - \Phi(x)\}$$
$$= \exp\left(-\frac{x^3}{3} \sum_{i=1}^{n} EX_i^3 \right) \left[1 + O(1)\{(1+x)L_{3n} + (1+x)^4 L_{4n}\} \right]$$

for $0 \leq x \leq C^{-1} L_{4n}^{-1/4}$, where $|O(1)|$ is bounded by an absolute constant.

- Question (Shao and Zhou, 2012): Whether a similar expansion holds for more general Studentized non-linear statistics, such as Hoeffding's class of U-statistics after suitably Studentized.

- Two-sample case:

  As a prototypical example of the two-sample U-statistics, the two-sample t-statistic is of significant interest due to its wide applicability.

  1. The robustness of the t-statistic is useful in high dimensional data analysis under the sparsity assumption on the signal of interest.
     - Delaigle, A., Hall, P. and Jin, J. (2011). Robustness and accuracy of methods for high dimensional data analysis based on Student's t-statistic. *J. R. Statist. Soc. B* **73** 283–301.
  2. When dealing with two experimental groups, typically assumed to be independent, in scientifically controlled experiments, the two-sample t-statistic is one of the most commonly used statistics for hypothesis testing and constructing confidence intervals for the difference between the means of the two groups.

Let $X_1, \ldots, X_m$ be a random sample from a population with mean μ_1 and variance σ_1^2, and let $Y_1, \ldots, Y_n$ be a random sample from another population with mean μ_2 and variance σ_2^2. Assume that the two random samples are drawn independently.

The two-sample t-statistic is defined as

$$\widehat{T}_{m,n} = \frac{\bar{X}_m - \bar{Y}_n}{\sqrt{\widehat{\sigma}_1^2/m + \widehat{\sigma}_2^2/n}},$$

where

$$\widehat{\sigma}_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X}_m)^2, \quad \widehat{\sigma}_2^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y}_n)^2.$$

Theorem (Chang, Shao and Zhou, 2014)

Assume that $\mu_1 = \mu_2$, *and* $E|X_1|^{2+\delta} < \infty$, $E|Y_1|^{2+\delta} < \infty$ *for some* $0 < \delta \leq 1$. *Then*

$$\frac{P(\widehat{T}_{m,n} \geq x)}{1 - \Phi(x)} = 1 + O(1)(1+x)^{2+\delta}\left\{\frac{(v_{1,2+\delta}/\sigma_1)^{2+\delta}}{m^{\delta/2}} + \frac{(v_{2,2+\delta}/\sigma_2)^{2+\delta}}{n^{\delta/2}}\right\}$$

holds uniformly for

$$0 \leq x \leq C^{-1} \min\left\{(\sigma_1/v_{1,2+\delta})m^{\delta/(4+2\delta)}, (\sigma_2/v_{2,2+\delta})n^{\delta/(4+2\delta)}\right\},$$

where $v_{1,s} = \left(E|X_1 - \mu_1|^s\right)^{1/s}$, $v_{2,s} = \left(E|Y_1 - \mu_2|^s\right)^{1/s}$ *for all* $s > 2$ *and* $|O(1)| \leq C$.

Theorem (Chang, Shao and Zhou, 2014)

Assume that $\mu_1 = \mu_2$, $E|X_1|^{2+\delta}, E|Y_1|^{2+\delta} < \infty$ *for some* $1 < \delta \leq 2$ *and write* $\gamma_1 = E(X_1 - \mu_1)^3$, $\gamma_2 = E(Y_1 - \mu_2)^3$. *Then*

$$P(\widehat{T}_{m,n} \geq x)/\{1 - \Phi(x)\}$$
$$= \exp\left\{ - \frac{x^3 \left(\gamma_1/m^2 + \gamma_2/n^2\right)}{3\left(\sigma_1^2/m + \sigma_2^2/n\right)^{3/2}} \right\}\{1 + O(1)\mathcal{R}_{m,n}(x)\}$$

holds uniformly for $0 \leq x \leq C^{-1}A_{m,n}$, *where*

$$A_{m,n} = \min\left\{(\sigma_1/v_{1,2+\delta})m^{\delta/(4+2\delta)}, (\sigma_2/v_{2,2+\delta})n^{\delta/(4+2\delta)}\right\},$$
$$\mathcal{R}_{m,n}(x)$$
$$= (v_{1,3}/\sigma_1)^3(1+x)m^{-1/2} + (v_{1,2+\delta}/\sigma_1)^{2+\delta}(1+x)^{2+\delta}m^{-\delta/2}$$
$$+ (v_{2,3}/\sigma_2)^3(1+x)n^{-1/2} + (v_{2,2+\delta}/\sigma_2)^{2+\delta}(1+x)^{2+\delta}n^{-\delta/2},$$

where $|O(1)| \leq C$.

▶ Key tool: A new randomized concentration inequality

Let $\xi_1, , ..., \xi_n$ be independent random variables and $W_n = \sum_{i=1}^n \xi_i$. Let $\eta_n = \eta_n(\xi_1, ..., \xi_n)$ be a non-negative measurable function of $\{\xi_1, ..., \xi_n\}$. Assume that $E\xi_i = 0$ and $\sum_{i=1}^n E\xi_i^2 = 1$ and define

$$\beta_2 = \sum_{i=1}^n E\xi_i^2 I(|\xi_i| > 1), \quad \beta_3 = \sum_{i=1}^n E|\xi_i|^3 I(|\xi_i| \leq 1).$$

Theorem (Shao and Zhou, 2012)

For each $1 \leq i \leq n$, let $\eta_n^{(i)}$ be a random variable such that ξ_i and $(\eta_n^{(i)}, W_n - \xi_i)$ are independent. Then

$$\sup_x P\big(|W_n - x| \leq \eta_n\big)$$

$$\lesssim \beta_2 + \beta_3 + E\eta_n + \sum_{i=1}^n E|\xi_i(\eta_n - \eta_n^{(i)})|.$$

3. Multivariate Gaussian approximation

Let $X, X_1, \ldots, X_n$ be i.i.d. p-variate random vectors with mean zero and covariance matrix Σ, and let $Z \sim N(0, \Sigma)$. As before, write

$$W_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i.$$

- Multivariate Berry-Esseen bound

Let $\mathcal{C}$ be the class of convex sets in $\mathbb{R}^p$.

(i) (Bentkus, 2003) Assume that $\mathrm{Cov}(X) = \Sigma = I$. Then

$$\sup_{A \in \mathcal{C}} \left| P(W_n \in A) - P(Z \in A) \right| \leq 400 \, \frac{p^{1/4} \, \beta}{\sqrt{n}},$$

where $\beta = E|X|_2^3$ and $|\cdot|_2$ denotes the Euclidean norm in $\mathbb{R}^p$.

In general, $\beta = O(p^{3/2})$ and the convergence rate is $O(p^{7/4} \, n^{-1/2})$.

(ii) (Bentkus, 2005) Let $X_1, \ldots, X_n$ be independent (not necessarily identically distributed) random vectors taking values in $\mathbb{R}^p$ such that $EX_i = 0$. Write

$$W_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i, \quad \Sigma = \mathrm{Cov}(W_n),$$

and let $Z \sim N(0, \Sigma)$. Then

$$\sup_{A \in \mathcal{C}} \left| P(W_n \in A) - P(Z \in A) \right| \lesssim \frac{p^{1/4}}{n^{3/2}} \sum_{i=1}^{n} E|\Sigma^{-1/2} X_i|_2^3.$$

Question: Why $p^{1/4}$?

Let $\Phi(\cdot)$ be the standard p-dimensional normal distribution with the density

$$\phi(x) = (2\pi)^{-p/2} \exp\left(-\frac{|x|_2^2}{2}\right), \quad x \in \mathbb{R}^p.$$

For every subset $A \subseteq \mathbb{R}^p$ and $\varepsilon > 0$, write

$$A^{\varepsilon} = \left\{x \in \mathbb{R}^p : d_A(x) \leq \varepsilon\right\}, \quad A^{-\varepsilon} = \left\{x \in A : B(x;\varepsilon) \subset A\right\},$$

where $d_A(x) = \inf_{y \in A} |x - y|_2$ and $B(x;\varepsilon) = \{y \in \mathbb{R}^p : |x - y|_2 \leq \varepsilon\}$. Then there exists a constant $a_p = a_p(\mathcal{C})$ depending only on p and $\mathcal{C}$, such that for all $A \in \mathcal{C}$ and $\varepsilon > 0$,

$$\Phi(A^{\varepsilon} \setminus A), \Phi(A \setminus A^{-\varepsilon}) \leq a_p\, \varepsilon.$$

In particular,

$$c\, p^{1/4} \leq a_p(\mathcal{C}) \leq 4\, p^{1/4},$$

where $c > 0$ is an absolute constant.

- Ball, K. (1993). The reverse isoperimetric problem for Gaussian measure. *Discrete Comput. Geom.* **10** 411–420.

In many statistical applications, a class of convex sets which is of particular interest is the rectangles; that is,

$$A_t = \left\{x = (x_1, \ldots, x_p)^T \in \mathbb{R}^p : x_j \leq t,\ 1 \leq j \leq p\right\},\quad t \in \mathbb{R}.$$

- Gaussian approximation for maxima of sums of random vectors (Chernozhukov, Chetverikov and Kato, 2013)

  Let $X_1, \ldots, X_n$ be i.i.d. p-variate random vectors with mean zero and covariance matrix $\Sigma = (\sigma_{jk})$, $Z = (Z_1, \ldots, Z_p)^T \sim N(0, \Sigma)$ and $W_n = (W_{n1}, \ldots, W_{np})^T = n^{-1/2} \sum_{i=1}^n X_i$. Moreover, write

$$m_r = \max_{1 \leq j \leq p} \left(E|X_{ij}/\sqrt{\sigma_{jj}}|^r\right)^{1/r}.$$

  (i) Polynomial tail: If m_r is bounded for some $r \geq 4$, then

$$\sup_{t \in \mathbb{R}} \left| P\left(\max_j W_{nj} \leq t\right) - P\left(\max_j Z_j \leq t\right)\right|$$
$$\lesssim \frac{\log^{7/8}(pn)}{n^{1/8}} + \log^{3/2}(pn)\left(\frac{p}{n^{r/2-1}}\right)^{1/(r+1)}.$$

(ii) Exponential tail: If $\max_j E \exp(c\,|X_{ij}/\sqrt{\sigma_{jj}}|^\gamma)$ is bounded for some $\gamma \geq \frac{1}{2}$, then

$$\sup_{t\in\mathbb{R}} \left| P\Big(\max_j W_{nj} \leq t \Big) - P\Big(\max_j Z_j \leq t \Big) \right| \lesssim \frac{\log^{7/8}(pn)}{n^{1/8}}.$$

- Anti-Concentration inequality

(Chernozhukov, Chetverikov and Kato, 2014)

Let $(Z_1, \ldots, Z_p)^T$ be a centered Gaussian random vector in $\mathbb{R}^p$ with $\sigma^2 = EZ_j^2$ for all j. Then for every $\varepsilon > 0$,

$$\mathcal{L}\Big(\max_j Z_j, \varepsilon \Big) \leq 4\,\varepsilon(a_p + 1)/\sigma,$$

where $a_p := E \max_{1\leq j\leq p} X_j/\sigma$.

Attention: The covariance structure can be arbitrary.

1. Chernozhukov, V., Chetverikov, D. and Kato, K. (2013). Gaussian approximations and multiplier bootstrap fro maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819.

2. Chernozhukov, V., Chetverikov, D. and Kato, K. (2014). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Forthcoming in Probab. Theory Relat. Fields.*

▶ Extension to U-statistics:

Let $\mathbf{X}_1, \ldots, \mathbf{X}_i = (X_{i1}, \ldots, X_{ip})', \ldots, \mathbf{X}_n$ be i.i.d. p-variate random vectors. Let $h(u, v)$ be a real-valued Borel measurable symmetric function of 2 variables. For each $j = 1, \ldots, p$, consider the Hoeffding's U-statistic

$$U_{nj} = \frac{1}{\binom{n}{2}} \sum_{1 \le i_1 < i_2 \le n} h(X_{i_1 j}, X_{i_2 j}),$$

which is an unbiased estimate of $\theta_j = Eh(X_{1j}, X_{2j})$. Let

$$h_{1j}(x) = E\big\{h(X_{1j}, X_{2j}) | X_{1j} = x\big\},\ x \in \mathbb{R}$$

and assume that $\sigma_j^2 = \mathrm{Var}\{h_{1j}(X_{1j})\} > 0$. Then the standardized non-degenerate U-statistic is given by

$$U_{nj}^* = \frac{\sqrt{n}}{2\sigma_j}\big(U_{nj} - \theta_j\big).$$

Let $\mathbf{Z} = (Z_1, \ldots, Z_p)'$ be a p-dimensional centered Gaussian random vector with covariance matrix $\boldsymbol{S} = (\sigma_{jk})_{1 \le j,k \le p}$, given by

$$\begin{aligned}\sigma_{jk} &= \mathrm{Corr}\big\{h_{1j}(X_{1j}), h_{1k}(X_{1k})\big\} \\ &= \mathrm{Cov}\big\{h_{1j}(X_{1j}), h_{1k}(X_{1k})\big\}/(\sigma_j\,\sigma_k),\ 1 \le j,k \le p.\end{aligned}$$

In particular, $\sigma_{jj} = 1$ for $j = 1, \ldots, p$.

Theorem (Shao and Zhou, 2014)

Suppose that there are positive constants c_0 and c_1 such that $\|h(\cdot,\cdot)\|_\infty \le c_0$ and $\min_{1 \le j \le p} \sigma_j \ge c_1$. Then there exist a constant $C > 0$ depending only on c_0 and c_1 such that

$$\sup_{x \in \mathbb{R}} \left| P\Big(\max_{1 \le j \le p} U_{nj}^* \le x \Big) - P\Big(\max_{1 \le j \le p} Z_j \le x \Big) \right| \le C \left\{ n^{-1} \log^7(pn) \right\}^{1/8}.$$

4. Application: High dimensional test of mean vectors under general covariance heterogeneity

Let $\mathbf{X}$ and $\mathbf{Y}$ be two random vectors in $\mathbb{R}^p$ with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, positive semi-definite covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. Consider two independent random samples of i.i.d. p-dimensional random vectors, $\mathcal{X}_n = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\mathcal{Y}_m = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$, drawn from the distributions of $\mathbf{X}$ and $\mathbf{Y}$, respectively. It is of general interest in testing the hypotheses

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \qquad \textit{versus} \qquad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2.$$

With fixed $p \leq n + m - 2$, Hotelling's T^2-statistic provides one of the most common procedures for testing $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

The two-sample Hotelling's T^2-statistic is defined by

$$T^2 = \frac{nm}{n+m}(\bar{\mathbf{X}} - \bar{\mathbf{Y}})' \boldsymbol{S}_{n,m}^{-1}(\bar{\mathbf{X}} - \bar{\mathbf{Y}}),$$

where $\bar{\mathbf{X}} = n^{-1}\sum_{i=1}^{n} \mathbf{X}_i$ and $\bar{\mathbf{Y}} = m^{-1}\sum_{j=1}^{m} \mathbf{Y}_j$ are the sample averages, and with $N = n + m$,

$$\boldsymbol{S}_{n,m} = \frac{1}{N-2}\left\{\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})' + \sum_{j=1}^{m}(\mathbf{Y}_j - \bar{\mathbf{Y}})(\mathbf{Y}_j - \bar{\mathbf{Y}})'\right\}$$

denotes the pooled sample covariance matrix.

Caution: In general $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$.

Two types of statistics, the sum-of-squares-type (L_2-type) and the maximum-type (L_∞-type) are used frequently for testing $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

- L_2-type statistics mimic the weighted Euclidean norm $\|\mathbf{A}^{1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\|_2^2$ for a given linear transformation $\mathbf{A}$. Tests based on the L_2-type statistics are powerful for detecting relatively dense signals; that is, deviations that $\boldsymbol{\mu}_1$ differs from $\boldsymbol{\mu}_2$ are spread in a large number of coordinates.

1. Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statist. Sinica* **6** 311–329.
2. Srivastava, M. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *J. Multivariate Anal.* **99** 386–402.
3. Srivastava, M. (2009). A test for the mean vector with fewer observations than the dimension under non-normality. *J. Multivariate Anal.* **100** 518–532.
4. Chen, S. X. and Qin, Y. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* **38** 808–835.
5. Zhong, P.-S., Chen, S. X. and Xu, M. (2013). Tests alternative to higher criticism for high-dimensional means under sparsity and column-wise dependence. *Ann. Statist.* **41** 2703–3110.

- Statistics of the L_∞-type are preferable for detecting relatively sparse signals. It has been used in a variety of applications including the medical image problem, anomaly detections and gene selections. Tests based on the L_∞-type statistics are reasonably more powerful than those based on the L_2-type when the alternatives are sparse.

1. Cai, T. T., Liu, W. and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *J. R. Statist. Soc. B* **76** 349–372.

2. Liu, W. and Shao, Q.-M. (2013). A Cramér moderate deviation theorem for Hotelling's T^2-statistic with applications to global tests. *Ann. Statist.* **41** 296–322.

A primary approach in the statistical hypothesis testing is to compute the critical value based on the asymptotic distribution of the test statistic. Existing research on testing high-dimensional means has so far focused on such the limiting distribution calibration approach.

The asymptotic distributions are usually established upon certain dependence conditions, while many of which can hardly be verified in practice.

- Cai, Liu and Xia (2013) assumed $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ and derived asymptotic null distributions of their test statistics under the following technical assumptions among other moment conditions:

  (1). $C^{-1} \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq C$ for some $C > 1$;

  (2). $\max_{1 \leq k < \ell \leq p} |\omega_{k\ell} / \sqrt{\omega_{kk}\omega_{\ell\ell}}| \leq c$ for some $c \in (0, 1)$, where $\boldsymbol{\Omega} = (\omega_{k\ell})_{1 \leq k, \ell \leq p} = \boldsymbol{\Sigma}^{-1}$ denotes the precision matrix.

- In the context of one-sample tests, Zhong, Chen and Xu (2013) proposed L_2-type thresholding test statistics and derived both limiting null and alternative distributions under weak dependence assumptions that

$$\mathbf{X}_i = \mathbf{W}_i + \boldsymbol{\mu}, \quad i = 1, \ldots, n,$$

where $\{\mathbf{W}_i = (W_{i,1}, \ldots, W_{i,p})^T\}_{i=1}^n$ forms a weakly stationary sequence and for each i,

$$\sum_k |\mathrm{Cov}(W_{i,1}, W_{i,k+1})| < \infty.$$

Motivated by applications in the image analysis and abnormality detection, we are particularly interested in detecting discrepancies when μ_1 and μ_2 are distinguishable to a certain extent in at least one coordinate.

- When dealing with a large number of automatically collected features in practice, we may not want to rule out the possibility that the seemingly high-dimensional problem has an intrinsic low dimensional structure and that a fraction of features are highly correlated.

  This raises a challenging question: are the existing methods still applicable if the underlying dependence structure is of a very general form?

- Another concern is that the convergence rate to the extreme value distribution of empirical processes is usually slow. Taking the extreme distribution of type I as an example, it was shown in Liu, Lin and Shao (2008) that the convergence rate is of order $O\left(\log(\log n)/\log n\right)$.

  Although the convergence rate may be improved by using proper intermediate approximations, still its validity relies on the dependence structure of the underlying distribution.

  The above two concerns inspire our work.

  - Chang, J., Zhou, W. and Zhou, W.-X. (2014). Simulation-based hypothesis testing of high dimensional means under covariance heterogeneity. Available at *arXiv:1406.1939*.

▶ Methodology

Let

$$T_{n,m} = \max_{1 \le k \le p} \frac{|\bar{X}_k - \bar{Y}_k|}{\left(\widehat{\sigma}_{1k}^2/n + \widehat{\sigma}_{2k}^2/m\right)^{1/2}},$$

where $\bar{X}_k = n^{-1} \sum_{i=1}^{n} X_{ik}$, $\bar{Y}_k = m^{-1} \sum_{j=1}^{m} Y_{jk}$ and

$$\widehat{\sigma}_{1k}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_{ik} - \bar{X}_k)^2, \quad \widehat{\sigma}_{2k}^2 = \frac{1}{m} \sum_{j=1}^{m} (Y_{jk} - \bar{Y}_k)^2.$$

Intuitively, large values of $T_{n,m}$ lead to a rejection of the null hypothesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

For a given level α, we wish to find appropriate critical values q_α to construct asymptotically α-level test; that is, reject H_0 if

$$T_{n,m} > q_\alpha.$$

Let $\widehat{\boldsymbol{\Sigma}}_1$ and $\widehat{\boldsymbol{\Sigma}}_2$ be estimates of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. Define

$$\widehat{\mathbf{R}}_{1,2} = \widehat{\mathbf{D}}_{1,2}^{-1/2}(\widehat{\boldsymbol{\Sigma}}_1 + \rho\,\widehat{\boldsymbol{\Sigma}}_2)\widehat{\mathbf{D}}_{1,2}^{-1/2}, \quad \widehat{\mathbf{D}}_{1,2} = \mathrm{diag}(\widehat{\boldsymbol{\Sigma}}_1 + \rho\,\widehat{\boldsymbol{\Sigma}}_2),$$

where $\rho = \rho_{n,m} = n/m$. Let $\mathbf{G} \sim \mathrm{N}(\mathbf{0}, \widehat{\mathbf{R}}_{1,2})$. We take the critical values as the following conditional $(1-\alpha)$-quantiles

$$q_\alpha = \inf\left\{t \in \mathbb{R} : P\left(|\mathbf{G}|_\infty > t \,|\, \widehat{\mathbf{R}}_{1,2}\right) \le \alpha\right\},$$

which can be computed via Monte Carlo simulations:

(i) Let $\{\mathbf{G}_1, \ldots, \mathbf{G}_M\}$ be a random sample of size $M \ge 1$ independently generated from underlying distributions of $\mathbf{G}$.

(ii) Define the following sample $(1-\alpha)$-quantiles:

$$\widehat{q}_\alpha = \widehat{F}_M^{-1}(1-\alpha) = \inf\left\{t \in \mathbb{R} : \widehat{F}_M(t) \ge 1-\alpha\right\},$$

where $\widehat{F}_M(t) = M^{-1}\sum_{\ell=1}^{M} I\{|\mathbf{G}_\ell|_\infty \le t\}$.

Define the empirical version of the near-oracle test $\Phi_\alpha = I\{T_{n,m} > q_\alpha\}$ by

$$\widehat{\Phi}_\alpha(M) = I\{T_{n,m} > \widehat{q}_\alpha\},$$

such that the null hypothesis H_0 is rejected as long as $\widehat{\Phi}_\alpha(M) = 1$.

The simulation-based test $\widehat{\Phi}_\alpha(M)$ approximates the near-oracle test Φ_α with increasing precision as M grows in the sense that

$$\lim_{M\to\infty} \widehat{\Phi}_\alpha(M) = \Phi_\alpha, \quad \text{almost surely.}$$

▶ Theoretical properties

For the p-dimensional random vectors $\mathbf{X} = (X_1, \ldots, X_p)^T$ (resp., $\mathbf{Y} = (Y_1, \ldots, Y_p)^T$) with mean $\boldsymbol{\mu}_1$ (resp., $\boldsymbol{\mu}_2$) and covariance matrix $\boldsymbol{\Sigma}_1$ (resp., $\boldsymbol{\Sigma}_2$), its marginally Studentized version is given by

$$\widetilde{\mathbf{X}} = (\widetilde{X}_1, \ldots, \widetilde{X}_p)' = \mathbf{D}_1^{-1/2}\mathbf{X}$$

(resp., $\widetilde{\mathbf{Y}} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_p)' = \mathbf{D}_2^{-1/2}\mathbf{Y}$), where $\mathbf{D}_1 = \mathrm{diag}(\boldsymbol{\Sigma}_1)$ (resp., $\mathbf{D}_2 = \mathrm{diag}(\boldsymbol{\Sigma}_2)$).

No regularity assumption on $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are required.

(M1) (Lower order moments) There exist $r \geq 4$ and $K_0 > 0$ such that the rth moments of all the components of $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ are uniformly bounded; that is,

$$\max_{1\leq k\leq p} \left(\mathbb{E}|\widetilde{X}_k|^r\right)^{1/r} \leq K_0 \quad \text{and} \quad \max_{1\leq k\leq p} \left(\mathbb{E}|\widetilde{Y}_k|^r\right)^{1/r} \leq K_0.$$

(M2) (Sub-exponential tails) There exist constants $K_1 > 0$, $K_2 > 1$ and $0 < \gamma \leq 2$ such that

$$\max_{1\leq k\leq p} \mathbb{E}\exp\left(K_1|\widetilde{X}_k|^\gamma\right) \leq K_2 \quad \text{and} \quad \max_{1\leq k\leq p} \mathbb{E}\exp\left(K_1|\widetilde{Y}_k|^\gamma\right) \leq K_2.$$

Throughout, we assume that $n, p \geq 2$, $n \asymp m$ ($n \leq m$) and that $\{\sigma_{\nu,kk} : \nu = 1, 2,\ k \geq 1\}$ is bounded away from 0 and ∞. For $r \geq 4$ as in condition (M1), we write

$$\alpha_{n,p} = p\, n^{1-r/2}.$$

Theorem (Chang, Zhou and Zhou, 2014)

Let $\mathbf{G} \sim N(\mathbf{0}, \widehat{\mathbf{R}}_{1,2})$ *and*

$$b = \min\left(\|\widehat{\mathbf{R}}_{1,2} - \mathbf{R}_{1,2}\|_\infty, 1\right).$$

(i) *Assume that (M1) holds. Then under* H_0,

$$\sup_{z>0} \left|P\left(T_{n,m} > z\right) - P\left(|\mathbf{G}|_\infty > z | \widehat{\mathbf{R}}_{1,2}\right)\right| \lesssim \{\log(pn)\}^{3/2}\left\{b^{1/3} + n^{-1/8} + \alpha_{n,p}^{1/(r+1)}\right\}.$$

(ii) *Assume that (M2) holds. Then under* H_0,

$$\sup_{z>0} \left|P\left(T_{n,m} > z\right) - P\left(|\mathbf{G}|_\infty > z | \widehat{\mathbf{R}}_{1,2}\right)\right| \lesssim \left[b^{1/3}\{\log(pn)\}^{2/3} + n^{-1/8}\{\log(pn)\}^{7/8} + n^{-1/2}\{\log(pn)\}^{3/2+1/\gamma}\right].$$

Corollary (Asymptotic level)

(i) *Assume that (M1) holds with* $p = O(n^{r/2-1-\delta})$ *for some* $\delta > 0$. *Then as* $n, p \to \infty$,

$$P_{H_0}\{\Phi_\alpha = 1\} \to \alpha.$$

(ii) *Assume that (M2) holds with* $\log(p) = o\left[n^{\min\{\gamma/(3\gamma+2),\, 1/7\}}\right]$. *Then as* $n, p \to \infty$,

$$P_{H_0}\{\Phi_\alpha = 1\} \to \alpha.$$

This implies that, under appropriate moment conditions, the proposed two-sample tests have size α asymptotically, while allowing for either a polynomial or an exponential rate of growth for the dimension p in the sample size.

The following results show the asymptotic power of the two-sample tests under conditions on the separation between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

Theorem (Asymptotic power)

Assume that either condition (M1) holds with $p = O(n^{r/2-1-\delta})$ *for some* $\delta > 0$, *or condition (M2) holds with* $\log(p) = o(n^{\gamma/2})$. *Fix some* $0 < \alpha < 1$, *let*

$$\lambda(p, \alpha) = \sqrt{2\log(p)} + \sqrt{2\log(1/\alpha)}$$

and let $\{\varepsilon_n\}_{n \geq 1}$ *be an arbitrary sequence of positive numbers satisfying* $\varepsilon_n \to 0$ *and* $\varepsilon_n \sqrt{\log(p)} \to \infty$ *as* $n \to \infty$. *Under the alternative* H_1 *with*

$$\max_{1 \leq k \leq p} \frac{|\mu_{1k} - \mu_{2k}|}{(\sigma_{1,kk}/n + \sigma_{2,kk}/m)^{1/2}} \geq (1 + \varepsilon_n)\, \lambda(p, \alpha),$$

we have as $n, p \to \infty$,

$$P_{H_1}\{\Phi_\alpha = 1\} \to 1.$$

The next result shows the power of the test Φ_α under sparse alternatives, i.e.

$$H_{s,1} : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \in \mathcal{S}(d) \text{ for } d = p^\tau \text{ and } 0 \le \tau < 1,$$

$$\min_{k:\mu_{1k} \neq \mu_{2k}} \frac{|\mu_{1k} - \mu_{2k}|}{(\sigma_{1,kk}/n + \sigma_{2,kk}/m)^{1/2}} \ge \sqrt{2\beta \log(p)} \text{ for } 0 < \beta < 1,$$

and the non-zero locations of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ are randomly uniformly drawn from $\{1, \ldots, p\}$.

We also need the following regularity condition on the covariance structure. For $\nu = 1, 2$ and $0 < \rho < 1$, let

$$\Lambda_\nu(\rho) = \left\{1 \le k \le p : |r_{\nu,k\ell}| \ge \rho \text{ for some } 1 \le \ell \le p,\ \ell \neq k\right\}.$$

(Dν). Suppose that $|\Lambda_\nu(\rho)| = o(p)$ for some $0 < \rho < 1$, and

$$\max_{1 \le k \le p} \sum_{\ell=1}^{p} r_{\nu,k\ell}^2 \le C_1$$

for some $C_1 > 0$.

Theorem (Asymptotic power under sparse alternatives)

Suppose both conditions (D1) and (D2) hold and assume that either condition (M1) holds with $p = O(n^{r/2-1-\delta})$ *for some* $\delta > 0$*, or condition (M2) holds with* $\log(p) = o(n^{\gamma/2})$*. Then under* $H_{s,1}$ *with* $\beta \geq (1-\sqrt{\tau})^2 + \varepsilon$ *for some* $\varepsilon > 0$*,*

$$P_{H_{s,1}}\left\{\Phi_{2,\alpha} = 1\right\} \to 1 \quad \text{as } n, p \to \infty.$$