High-dimensional mean tests and extensions

1. Hotelling's T^2
2. Sum-of-squares type tests
3. Max (over dimension) type tests
4. Refinements and extensions to time series
5. Testing for (auto) covariances

Problem of interest

- Interested in one/two-sample mean test in the high dimension setting. For example, interested in identifying sets of genes which are significant with respect to certain treatments from microarray data, brain-connectivity detection using fMRI data, etc.
- Let $\{X_1, \ldots, X_n\}$ be IID $p \times 1$ vectors with

$$\mu := \mathbb{E}X_1 = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}, \Sigma := \mathrm{Cov}(X_1) = \mathbb{E}(X_1 - \mu)(X_1 - \mu)'$$

(In previous classes, $p = d$ and $n = T$.)

- Interested in two-sample p-dimensional mean test, namely,

$$X_1, \ldots, X_{n_1} \sim F_1 \text{ with mean } \mu_1, \Sigma_1,$$

$$Y_1, \ldots, Y_{n_2} \sim F_2 \text{ with mean } \mu_2, \Sigma_2,$$

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2.$$

Hotelling's T^2 for fixed $p \ll n$

- If the dimension p is smaller than the sample sizes n_1 and n_2, the state-of-the-art method is Hotelling's T^2 test.

$$T^2 = (\overline{X} - \overline{Y})' \left\{ S_n \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right\}^{-1} (\overline{X} - \overline{Y})$$

$$S_n = \frac{1}{n} \left\{ (n_1 - 1)\hat{\Sigma}_1 + (n_2 - 1)\hat{\Sigma}_2 \right\}, \quad n = n_1 + n_2 - 2.$$

- Under $H_0 : \mu_1 = \mu_2$ and Gaussianity, we have

$$\frac{n_1 + n_2 - p - 1}{pn} T^2 \sim F(p, n_1 + n_2 - p - 1)$$

- Under $H_1 : \mu_1 = \mu_2$, it is non-central F-distribution.

Hotelling's T^2 in high dimension

- If $p > n_1 + n_2 - 2$, then S_n is not invertible.
- Poor power when $p \approx n$. For example, Yin, Bai and Krishnaiah (1988) show that when $p/n \to c$, the smallest and the largest eigenvalues of the sample covariance $\hat{\Sigma}$ do not converge to the respective eigenvalues of Σ.
- Therefore, Hotelling's T^2 cannot be used for HD mean test.
- Here, we overview several high-dimensional mean tests based on
  - Sum-of-squares type statistics
  - Max-type statistics

  and their refinements and extensions to TS setting.

Sum of squares type tests

It starts with Bai and Saranadasa (1996) assuming $\Sigma_1 = \Sigma_2 = \Sigma$.

- Just get rid of S_n^{-1} in Hotelling's T^2, and work with

$$(\overline{X} - \overline{Y})'(\overline{X} - \overline{Y})$$

- Subtract mean and divided by standard deviation gives the test statistic as

$$T_{BS} = \frac{(\overline{X} - \overline{Y})'(\overline{X} - \overline{Y}) - \frac{n_1+n_2}{n_1 n_2} tr(S_n)}{\frac{n_1+n_2}{n_1 n_2}\sqrt{\frac{2(n+1)n}{(n+2)(n-1)}(tr(S_n^2) - n^{-1}(tr(S_n))^2)}},$$

where $n = n_1 + n_2 - 2$.

- For example ($\mu_1 = 0$)

$$\mathbb{E}\overline{X}'\overline{X} = \mathbb{E}\frac{1}{n_1^2}\sum_{s,t} X_t' X_s$$

$$= \mathbb{E}\frac{1}{n_1^2}\sum_{t,s} tr(X_s X_t') = \frac{1}{n_1} tr(\Sigma_1) \approx \frac{1}{n_1} tr(\hat{\Sigma}_1) \qquad (1)$$

B&S test

- CLT: Assume factor-like model

$$X_i = \Gamma z_i + \mu_1, \quad Y_i = \Gamma z_i + \mu_2,$$

Γ is a $p \times m$ matrix ($m \leq \infty$) with $\Gamma\Gamma' = \Sigma$ (hence common covariance), z_i are i.i.d. random vectors with some moments conditions.

$$p/n \to c \in [0, \infty), \quad \lambda_{max}(\Sigma) = o(\sqrt{p}).$$

Then, under $H_0 : \mu_1 = \mu_2$

$$T_{BS} \to \mathcal{N}(0, 1)$$

- Note that the dimension could be larger than the sample size ($n = n_1 + n_2 - 2$).

B&S test

About the assumption on $\lambda_{\max}(\Sigma)$:

- Small exercise in (1) gives

$$\|\overline{X} - \mu\|^2 = O\left(\frac{tr(\Sigma)}{n}\right)$$

- Similarly, we can show that

$$\mathrm{Var}((\overline{X} - \overline{Y})'(\overline{X} - \overline{Y})) = O(n^{-2} tr(\Sigma^2)).$$

- Hence, eigenvalue condition says that the variance term vanishes as sample size increases:

$$\frac{1}{n^2} tr(\Sigma^2) \leq \frac{p(\lambda_{max}^2(\Sigma))}{n^2} = o(p^2/n^2) \to 0.$$

Extensions of B&S

- Many extensions are suggested, for example, Srivastava and Du (2008) suggested weighted version

$$(\overline{X} - \overline{Y})' D_s^{-1} (\overline{X} - \overline{Y}), \quad D_s = \mathrm{diag}(s_{11}, \ldots, s_{pp}),$$

  where s_{ii} are the diagonal elements of pooled sample covariance S.
- However, B&S assumes $p/n \to c$, so it is not working for ultra high dimension when $p/n \to \infty$.
- Chen and Qin (2010) modified B&S by removing cross-term $\sum_t X_t' X_t$. Essentially, p and n are related in the proof by $\lambda_{\max}(\Sigma)$ condition which involves the square term calculation $X_t' X_t$, $Y_t' Y_t$.

Extension of B&S

- The CQ test statistic is given by

$$T_n = \frac{\sum_{s\neq t} X_s' X_t}{n_1(n_1-1)} + \frac{\sum_{s\neq t} Y_s' Y_t}{n_2(n_2-1)} - 2\frac{\sum_{s,t} X_s' Y_t}{n_1 n_2}$$

and satisfies

$$\frac{T_n - \|\mu_1 - \mu_2\|^2}{\sqrt{Var(T_n)}} \to \mathcal{N}(0,1)$$

as $p, n \to \infty$ but with only

$$\frac{tr(\Sigma^4)}{tr^2(\Sigma^2)} \to 0 \quad as \quad p \to \infty.$$

Max-type tests

- Cai et al. (2014) suggested max-type test statistic:

$$T_{CLX} = \frac{n_1 n_2}{n_1 + n_2} \max_{1 \le i \le p} \frac{|\overline{X}^{(i)} - \overline{Y}^{(i)}|^2}{s_{ii}}$$

- Then, under suitable conditions, it converges to Type I extreme value Gumbel distribution.

$$P(T_{CLX} - 2\log p + \log\log p \le x) \to \exp\left(-\frac{1}{\sqrt{\pi}}\exp(-x/2)\right).$$

- Finite sample improvement using bootstrap. For example, Chernozhukov et al. (2013) proposed a (Gaussian) multiplier bootstrap and Chang et al. (2017) proposed a Gaussian parametric bootstrap.

Multiplier/Wild Bootstrap

- Multiplier boostrap or wild bootstrap for IID observations was originally proposed to replicate residuals in regression with nonconstant variance. For example, assume that

$$\mathbb{E}u_t = 0, \quad \mathbb{E}u_t^2 = \sigma_t^2.$$

  Then, multiplier bootstrap gives

$$\mathbb{E}\epsilon_t u_t = \mathbb{E}\epsilon_t \mathbb{E}u_t = 0$$

$$\mathbb{E}\epsilon_t^2 u_t^2 = \mathbb{E}\epsilon_t^2 \mathbb{E}u_t^2 = \sigma_t^2.$$

- Hence the key condition for WB is zero mean, unit variance. Further assumption $\mathbb{E}\epsilon_t^3 = 1$ gives more efficiency.
- For high dimensional WB, normal distribution is widely used for the reasons given in later slides.

Gaussian Approximation for IID HD

- The key result is due to Chernozhukov et al. (2013). Gaussian approximation for HD IID observations. For IID HD obsrevations $X_1, \ldots, X_n$ with mean μ and $\Sigma = \mathbb{E}X_t X_t'$,

$$\sup_{u \geq 0} \left| P(\sqrt{n}|\overline{X} - \mu|_\infty \geq u) - P(\sqrt{n}|\overline{Z} - \mu|_\infty \geq u) \right| \to 0,$$

  where $Z_1, \ldots, Z_n$ are i.i.d $\mathcal{N}(\mu, \Sigma)$ and $|\nu|_\infty = \max_{j \leq p} \nu_j$.
- Main idea of proof is first to approximate $\max$ by smooth differentiable function

$$F_\beta(z) = \beta^{-1} \log \left(\sum_{j=1}^{p} \exp(\beta z_j) \right), \quad \beta > 0$$

  and use the bounds

$$0 \leq F_\beta(z) - \max_{1 \leq j \leq p} z_j \leq \frac{\log p}{\beta}$$

Gaussian Approximation for IID HD

- Then, the maximum of non-Gaussian random variables can be approximated by that of Gaussian with the following error bound:

$$|\mathbb{E}\{g(F_\beta(\sqrt{n}|\overline{X}-\mu|)) - g(F_\beta(\sqrt{n}|\overline{Z}-\mu|))\}| \leq D_n$$

  for $g \in C_b^3(\mathbb{R})$.
- By using Taylor expansion of F_β and anti-concentration inequality for Gaussian random variable due to Nazarov (2003)

$$P(Z \leq z + a) - P(Z \leq z) \leq Ca\sqrt{\log p},$$

  we bound the KS distance.

Multiplier Bootstrap

- Gaussian multiplier boostrap is obtained by considering

$$\max_{1\leq j\leq p}\sum_{t=1}^{n} X_{jt}\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0,1)$$

- This works because comparing the distribution functions of maxima of two-Gaussian vectors V and W gives

$$\sup_{u\in\mathbb{R}}\left|P(\max_{1\leq j\leq p} V_j \leq u) - P(\max_{1\leq j\leq p} W_j \leq u)\right| = O(\Delta_0^{1/3}),$$

where

$$\Delta_0 = \max_{1\leq j,k\leq p}|\sigma_{jk}^V - \sigma_{jk}^W|$$

- Very roughly speaking:

$$\max\sum_{t=1}^{n} X_t \approx \max\sum_{t=1}^{n} Z_t \approx \max\sum_{t=1}^{n} X_t\epsilon_t$$

Finite Sample Performance

- Many simulations/empirical analyses suggest that max-type tests perform well in sparse signals in the sense that means possibly differ in only a small number of coordinates. In contrast, SS-type works better for "dense signals" as the opposite of sparsity.
- In principal, however, all tests are related to estimation of Σ in some way. If the dimension is too high, this is a non-trivial task and it is hard to expect good performance.
- This leads to the development of thresholding/screening before applying mean tests.

Thresholding/Screening

- Basic idea is to reduce dimension before applying tests.
- Chen et al. (2018) suggests thresholding for their SS-type test as

$$L_1(s) = \sum_{k=1}^{p} nT_{nk} I\left\{nT_{nk} + 1 > 2s\log p\right\}, n = \frac{n_1 n_2}{n_1 + n_2}, s \in (0,1),$$

where

$$T_{nk} = \frac{\sum_{s \neq t} X_s^{(k)} X_t^{(k)}}{n_1(n_1-1)} + \frac{\sum_{s \neq t} Y_s^{(k)} Y_t^{(k)}}{n_2(n_2-1)} - 2\frac{\sum_{s,t} X_s^{(k)} Y_t^{(k)}}{n_1 n_2}.$$

Then, for $s \in (0,1)$,

$$\frac{L_1(s) - \mathbb{E}L_1(s)}{\sqrt{\mathrm{Var}(L_1(s))}} \to \mathcal{N}(0,1)$$

as $n \to \infty$, $p/n \to \infty$.

Thresholding/Screening

- Chang et al. (2017) suggest the screening for max-type test with significance level α. Select components satisfying

$$\left| \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\overline{X}^{(i)} - \overline{Y}^{(i)}}{\sqrt{s_{ii}}} \right| > \sqrt{2 \log p} + \frac{1}{\sqrt{2 \log p}} + \sqrt{-2 \log \alpha}$$

  and perform max-type tests such as CLX with Gaussian parametric bootstrap for p-value calculation.

Illustration with gene sets

- Technically gene sets are defined in gene ontology (GO) system that provides structured and controlled vocabularies producing names of gene sets.
- Two treatments for cancer are given, and interested whether the population mean of each treatment group is the same.

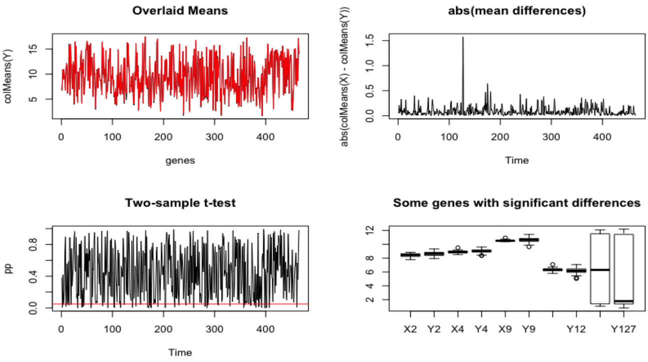

Illustration with gene sets

- Readily implemented in `highmean`, `HDtest` packages in R.
- Result gives the following:

```
$pval
Bai1996
0.04818344

$pval
Chen2010
0.04570195

$pval
Cai2014
0.1331871
```

Extensions to TS setting

- Only few papers address the case for temporally dependent observations.
- Extension of B&S to time series context is done by Ayyala et al. (2017).
- Max-type test is considered in Zhang and Chen (2014) and Zhang and Wu (2018).
- For the shortness sake, let us consider one-sample test, that is, H_0: $\mu = 0$ versus H_1: $\mu \neq 0$.

SS-type test for TS

- Observe for TS that

$$\mathbb{E}(\overline{X}'\overline{X}) = \frac{1}{n^2}\mathbb{E}\sum_{s,t} X_s' X_t = \frac{1}{n^2}\sum_{s,t} tr\mathbb{E}X_t X_s'$$

$$= \frac{1}{n^2}\sum_{s,t} tr(\mathbb{E}X_t X_s') = \frac{1}{n}\sum_{h=-(n-1)}^{n-1}\left(1-\frac{|h|}{n}\right)\gamma_X(h) = \frac{1}{n}tr(\Omega_n),$$

  where Ω_n is the long-run variance!
- Hence, B&S test in TS context is based on

$$T_A = \frac{\overline{X}'\overline{X} - n^{-1}tr(\Omega_n)}{\sqrt{\mathrm{Var}(\overline{X}'\overline{X})}} \to N(0,1)$$

- See Ayyala et al. (2017) for the test statistic with estimation of $tr(\Omega_n)$ and variance term.

Max-type test for TS

- Test statistic is given by

$$\sqrt{n} \max_{1 \leq j \leq p} |\overline{X}_j|$$

- p-value is calculated from block multiplier (wild) bootstrap (BWB).
- Block Wild Bootstrap (BWB) sample is obtained by

$$X^*_{jt} = X_{jt}\epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1)$$

where t-th observations falling into ith block share the same multiplier ϵ_i.

Gaussian Approximation for dependent HD TS

- Extension depends heavily on the measure of dependence in HDTS.
- Zhang and Cheng (2018) extend this to weakly dependent series $\{X_t\}$ under functional dependence with some restrictions.
- Furthermore, Zhang and Wu (2017) extended further scaled maximum under general functional dependence measure,

$$\sup_{u\geq 0}\left|P(\sqrt{n}|D_0^{-1/2}(\overline{X}-\mu)|_\infty\geq u)-P(\sqrt{n}|D_0^{-1/2}(\overline{Z}-\mu)|_\infty\geq u)\right|\to 0,$$

where $D_0=\mathrm{diag}(\Omega)$ is the diagonal matrix of long-run variance Ω.

Test for Covariance

- For IID samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, wish to test the hypotheses

$$H_0 : \Sigma_1 = \Sigma_2 \quad \text{versus} \quad H_1 : \Sigma_1 \neq \Sigma_2,$$

- Cai et al. (2013) suggest the max-type statistic

$$M_n = \max_{1 \le i \le j \le p} \frac{(\hat{\sigma}_{ij1} - \hat{\sigma}_{ij2})^2}{\theta_{ij1}/n_1 + \theta_{ij2}/n_2},$$

where

$$\hat{\sigma}_{ij1} = \frac{1}{n_1} \sum_{t=1}^{n_1} (X_{ti} - \overline{X}_i)(X_{tj} - \overline{X}_j), \hat{\sigma}_{ij2} = \frac{1}{n_2} \sum_{s=1}^{n_2} (Y_{si} - \overline{Y}_i)(Y_{sj} - \overline{Y}_j)$$

$$\hat{\theta}_{ij1} = \frac{1}{n_1} \sum_{t=1}^{n_1} [(X_{ti} - \overline{X}_i)(X_{tj} - \overline{X}_j) - \hat{\sigma}_{ij1}]^2,$$

$$\hat{\theta}_{ij2} = \frac{1}{n_2} \sum_{s=1}^{n_2} [(Y_{si} - \overline{Y}_i)(Y_{sj} - \overline{Y}_j) - \hat{\sigma}_{ij2}]^2.$$

Test for Covariance

- Then, under the null, it converges to Gumbel distribution

$$P(M_n - 4\log p + \log\log p \leq t) \to \exp\left(-\frac{1}{\sqrt{8\pi}}\exp(-t/2)\right)$$

- Zhang and Wu (2015) also showed the Gaussian approximation result:

$$\sup_{u\geq 0}\left|P(\sqrt{n}\max_{ij}\frac{|\hat{\sigma}_{ij}-\sigma_{ij}|}{\tau_{ij}} \geq u) - P(\max_{ij}\frac{|Z_{ij}|}{\tau_{ij}} \geq u)\right| \to 0,$$

where $Z \sim N(0, T)$, and τ_{ij} is the ij element of T which is the covariance matrix of $\mathrm{vec}(X_t X_t' - \mathbb{E}X_t X_t')$.

Test for Autocovariance

- Baek et al. (2019+) further interested in

$$H_0 : \gamma_X(h) = \gamma_Y(h), \quad h = 0, \ldots, \pm K, \tag{2}$$

for fixed K, where $\gamma_X(h) = \mathbb{E}X_{t+h}X_t'$ and $\gamma_Y(h) = \mathbb{E}Y_{t+h}Y_t'$, $h \in \mathbb{Z}$, (assume $\mathbb{E}X_t = 0$, $\mathbb{E}Y_t = 0$) are the (matrix) autocovariance functions (ACVFs) of the two series.

- It is equivalent to

$$H_0 : \mathbb{E}Z_t = 0, \tag{3}$$

where

$$Z_t = \begin{pmatrix} \mathrm{vech}(X_tX_t' - Y_tY_t') \\ \mathrm{vec}(X_{t+1}X_t' - Y_{t+1}Y_t') \\ \vdots \\ \mathrm{vec}(X_{t+K}X_t' - Y_{t+K}Y_t') \end{pmatrix}. \tag{4}$$

Test for Autocovariances and discussion

- Previously introduced SS-type, Max-type testing procedures can be adapted to transformed data observations $\{Z_1, \dots, Z_{n-K-1}\}$.
- Determining optimal block size is also interesting and important for finite sample performance.
- We can embed factor models for the tests of autocovariances.

References

- Ayyala, D. N., Park, J. and Roy, A. (2017). Mean vector testing for high-dimensional dependent observations, Journal of Multivariate Analysis 153, 136–155.
- Bai, Z. and Sarandasa, H. (1996). Effect of high dimension: By an example of a two sample problem. Statist. Sinica 6, 311–329.
- Cai, T., Liu, W. D. and Xia, Y.(2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings.J.Am.Stat.Assoc.108, 265–277.
- Cai, T. T., Liu, W. and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. Journal of the Royal Statistical Society, Series B, 76, 349–372.
- Chang, J. , Zheng, C. , Zhou, W. and Zhou, W. (2017). Simulation-based hypothesis testing of high dimensional means under covariance heterogeneity. Biom, 73: 1300–1310. doi:10.1111/biom.12695
- Chen, S. X. and Qin, Y. (2010). A two sample test for high dimensional data with applications to gene-set testing. The Annals of Statistics, 38, 808–835.
- Chen, S. X., Li, J. and Zhong, P. S. (2018) Two-sample and ANOVA tests for high dimensional means. To appear in The Annals of Statistics.
- Chernozhukov, V., Chetverikov, D. and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. Ann. Statist. 41 2786–2819.
- Nazarov, F. (2003). On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure. In Geometric Aspects of Functional Analysis. Lecture Notes in Math. 1807, 169–187. Springer, Berlin.
- Srivastava M. S. and Du M. (2008). A test for the mean vector with fewer observations than the dimension, J. Multivariate Anal. 99, 386–402.
- Yin, Y., Bai, Z. and Krishnatah, P. R. (1988). On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. Probab. Theory Related Fields 78 509–521.
- Zhang, D. and Wu, W. B. (2017). Gaussian approximation for high dimensional time series, The Annals of Statistics 45(5), 1895–1919.
- Zhang, X. and Cheng, G. (2018). Gaussian approximation for high dimensional vector under physical dependence. Bernoulli 24, no. 4A, 2640–2675.