Comparison of High-dimensional Covariance Matrix Testing between RMT and Graph-based Methods

Yan-Yu Chen, Mingshuo Liu

December 28, 2024

Contributions: Equal contributions. Everyone participates each step.
Keywords: High-dimensional data, Marčenko–Pastur distributions, Covariance hypothesis test, Similarity graph.

1 Motivation

High-dimensional data arise in many modern scientific fields, in which cases the classical asymptotic theory fails to apply to traditional statistics, e.g., two-sample tests. In classical asymptotic theory, the number of variables p is a fixed constant. Let $y_n := \frac{p}{n}$ be the ratio that measures the difference between the sample size n and p. It is clear that $y_n \to 0$ as n goes to infinity in the classical settings. However, in high-dimensional scenario, p is no longer a constant and grows as n increases. When n goes to infinity, y_n will not vanishes but converges to a non-zero y since p also grows to infinity. This fundamental difference of high-dimensional scenario completely break the limiting behavior of traditional statistics. They even cause very large deviation from the truth in high dimensional cases, as pointed out by Dempster (1958), T^2-test has much failure in high dimensional data. Therefore, necessary correction or novel methods should be developed.

Several novel statistics have been proposed to overcome the difficulty of high-dimensional data. Modern random matrix theory (RMT) has emerged as a particularly useful framework for analyzing high-dimensional data. In the meanwhile, Friedman and Rafsky (1979) developed nonparametric testing for two sample that can be applied to data with arbitrary dimension using the graph-based approach. In this project, we aim to compare the power of two methods in the scenario that n, $p \to \infty$ with $y_n \to y$ for the two-sample test. Specifically, we have a two-sample covariance testing problem in high-dimensional scenario

- Data: $\mathbf{X}_1 = \{X_{1i}\}_{i=1}^{n_1}$ with $X_{1i} \sim N_p(\mu_1, , \Sigma_1)$; $\mathbf{X}_2 = \{X_{2i}\}_{i=1}^{n_2}$ with $X_{2j} \sim N_p(\mu_2, , \Sigma_2)$
- Scenario: n, $p \to \infty$ with $y_n := \frac{p}{n} \to y \in (0, 1)$
- Hypothesis: $H_0 : \Sigma_1 = \Sigma_2$ v.s. $H_a : \Sigma_1 \neq \Sigma_2$
- Goal: Compare the methods to find the powerful level-α test

In the simulation study, we compare the power of the following methods. One method proposed by Bai et al. (2009) copes with high-dimensional effects using RMT and the other presented by Chen and Friedman (2017) utilized similarity graphs to construct a powerful test statistic We expect Chen's method will outperform Bai's eventually as $y \to 1$. Because Bai's method is valid for all $y \in (0, 1)$, we believe that Bai's power should be greater than Chen's method when y is in the middle of $(0,1)$. However, as y increases to 1, Bai's method should be more unreliable. On the other hand, Chen's method does not have such restrictions and takes advantage of the non-parametric method.

The rest of the paper was organized as following. Preliminary and useful RMT results were recalled in Section 2. We devoted the methodology of this two methods to Sections 3 with the minimum pre-requested background. Due to the page limit, some detail expressions in Section 2 and 3 are omitted, but can be found in their original paper. In Section 4, we compared their numerical performance in our simulation study. Finally, in Section 5, we summarized our finding.

2 Background

In this section, we first review traditional likelihood ratio method (LRT), and then empathized some useful results selected from Paul and Aue (2014). They will be applied to derive Bai's corrected likelihood ratio test (CLRT). Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be two samples. For traditional LRT, we calculate their sample covariance matrices, S_1 and S_2. We define the likelihood ratio as

$$L_1 = \frac{\sup_{\theta\in\Theta} L(\theta|\mathbf{X}_1, \mathbf{X}_2)}{\sup_{\theta\in\Theta_0} L(\theta|\mathbf{X}_1, \mathbf{X}_2)} = \frac{|S_1|^{n_1/2}|S_2|^{n_2/2}}{\left|(n_1/N)S_1 + (n_2/N)S_2\right|^{N/2}},$$

where $\theta = \{\mu_1, \mu_2, \Sigma_1, \Sigma_2\}$ and $N = n_1 + n_2$. The LRT statistic can now calculated as $T_N = -2\log L_1$. In classical scenario, Wilks' theorem implies $T_N \xrightarrow{d} \chi^2_{df-df_0}$, $df - df_0 = \frac{p(p+1)}{2}$. However, in high-dimensional scenario, analysis using RMT shows that $T_N \xrightarrow{a.s.} \infty$, which leads to many false rejections of H_0. Therefore, the LRT must be corrected to extend to the high-dimensional regime.

Next, we review the result from RMT. Suppose that X is an $N \times N$ matrix with eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$. The empirical distribution of the eigenvalues of X, referred to as the empirical spectral distribution (ESD) of X, is the function $N^{-1}\sum_{i=1}^N \delta_{\lambda_i}$, where δ_y denotes the Dirac mass at y. If X is Hermitian, so that the eigenvalues of X are real, we can define the empirical distribution function of X as $F^X(x) = N^{-1}\sum_{j=1}^N \mathbf{1}_{\lambda_j \le x}$ for $x \in \mathbb{R}$. For two-sample problem, we particularly analyze the ESD $F_n^{V_n}$, where $V_n := S_1 S_2^{-1}$ is the F-matrix.

Let $\widetilde{\mathcal{A}}$ be the set of analytic complex functions, and $\widetilde{G}_n(f)$ be the empirical process such that

$$\widetilde{G}_n(f) := p\int_{-\infty}^{\infty} f(x)[F_n^{V_n} - F_{y_{n_1}, y_{n_2}}]dx, \qquad f \in \widetilde{\mathcal{A}},$$

where $F_n^{V_n}$ is the empirical spectral distribution (ESD) of $V_n := S_1 S_2^{-1}$ for S_1 and S_2 being the associated sample covariance matrices with $n = (n_1, n_2)$, and $F_{y_{n_1}, y_{n_2}}$ is the limiting distribution of $F_n^{V_n}$. Bai presented the following theorem from Zheng (2012), which mainly establish the theoretical guarantee of Bai's CLRT.

Theorem 1. *Let $f_1, \ldots, f_k \in \widetilde{\mathcal{A}}$. For each p, (ξ_{ij_1}) and (η_{ij_2}) are i.i.d. real variables, $1 \le i \le p, 1 \le j_1 \le n_1, 1 \le j_2 \le n_2$. $E\xi_{11} = E\eta_{11} = 0$, $E|\xi_{11}|^2 = E|\eta_{11}|^2 = 1$, and $E|\xi_{11}|^4 = E|\eta_{11}|^4 < \infty$. Furthermore, $y_{n_1} \to y_1 \in (0,1)$, $y_{n_2} \to y_2 \in (0,1)$. Then, the random vector $\left(\widetilde{G}_n(f_1), \ldots, \widetilde{G}_n(f_k)\right)$ weakly converges to a k-dimensional Gaussian vector with the mean vector $m(f_j)$ and the covariance function $v(f_j, f_\ell)$.*

Both $m(f_j)$ and $v(f_j, f_\ell)$ can be expressed by using contour integrals, and were shown in Bai's paper.

3 Methodology

We divided the discussion into two subsections according to the methods

3.1 CLRT

Bai proposed a remedy scaling of the LR statistic T_N such that the CLRT statistics weakly converges to the Gaussian vector with the mean $m(f)$ and covariance function $v(f)$ specified in Bai's theorem. Let L_1 be the traditional likelihood ratio test (LRT) statistic, $f(x) = \log\big((n_1/N)x + n_2/N\big) - (n_1/N)\log x$, $\tilde{G}(f) = -\frac{2\log L_1}{N} - p \cdot F_{y_{n_1},y_{n_2}}(f)$. The we apply the previous theorem and obtain Bai's CLRT.

Theorem 2. *Assuming that the conditions of the previous theorem hold under H_0, and*

$$f(x) = \log\big(y_{n_1} + y_{n_2}x\big) - \frac{y_{n_2}}{y_{n_1} + y_{n_2}}\log x - \log\big(y_{n_1} + y_{n_2}\big).$$

Then, under H_0 and $n_1 \wedge n_2 \to \infty$,

$$\widetilde{T}_N = v(f)^{-1/2}\left[-\frac{2\log L_1}{N} - p \cdot F_{y_{n_1},y_{n_2}}(f) - m(f)\right] \Rightarrow N(0,1),$$

where $N = n_1 + n_2$.

Both $m(f)$ and $v(f)$ are dependent only on (y_1, y_2) and again can be found explicitly in Bai's paper. This result, unlike tradictional LRT, is distribution-fee. For non-Gaussian data, CLRT is a generalized pseudo-likelihood ratio test (or Gaussian LRT). However, there is no easy way to generalize CLRT from $y \in (0,1)$ to $y > 1$ since L_1 will become undefined.

3.2 Graph-based two-sample test

Chen presented a novel test statistic based on a similarity graph constructed on the pooled observations from two samples. This non-parametric method has good property of asymptotic distribution free, which is shown to be powerful under location and scale alternatives. It transforms the original data into graphs, and do inference based on the constructed graphs. It also allows the case $p \geq n$, which is not possible in CLRT. So Chen's method can be a competitor with CLRT.

To be specific, Chen uses the kind of graphs, under which two observations are easier to be connected, if their distances are closer. One test statistics is defined as :

$$S = (R_1 - \mu_1, R_2 - \mu_2)\,\Sigma^{-1}\begin{pmatrix} R_1 - \mu_1 \\ R_2 - \mu_2 \end{pmatrix} \tag{1}$$

where $\mu_1 = \mathbf{E}(R_1), \mu_2 = \mathbf{E}(R_2)$, and Σ is the covariance matrix of the vector $(R_1, R_2)'$ under permutation null distribution, which is shown asymptotically to be χ_2 distribution.

Another test statistics we consider is

$$Z_m = \max(Z_w, |Z_{\text{diff}}|) \tag{2}$$

is shown to be asymptotically bi-variate Gaussian distributed, where

$$\begin{aligned} Z_w &= \frac{qR_1 + pR_2 - \big(q\mathrm{E}(R_1) + p\mathrm{E}(R_2)\big)}{pq\sqrt{|G|}} \\ Z_{\text{diff}} &= \frac{R_1 - R_2 - \big(\mathrm{E}(R_1) - \mathrm{E}(R_2)\big)}{\sqrt{rpq|G|}} \end{aligned} \tag{3}$$

Equation 1 and 2 are related test statistics extracted from the graph edges connection, and details for explaining them are omitted here.

4 Simulation

In this project, we compared two methods, CLRT, as a remedy scaling of the LR statistic, and graph-based methods for high dimensional two sample test problem. For simplicity, we take $n = n_1 = n_2$, to compare these two methods. Choosing $\sigma_1 = I_p$, we simulate the alternatives for

- $H_1 : \Sigma_1 \Sigma_2^{-1} = \text{diag}(3, 1, \ldots, 1)$
- $H_2 : \Sigma_1 = a\Sigma_2, a \in \mathbb{R}$, a is a pre-specified scalar standing for their difference extent that changes with dimension to provide comparable power, $a = c(1.4, 1.15, 1.1, 1.1, 1.1, 1.05)$

Under several different choices of y_n and a, we expect to observe the power of Chen's eventually dominates Bai's as $y_n \to 1$. We summarized our results as follows.

Table 1: Size and power comparison under H_1 when $\alpha = 0.05$

(n, p)	CLR		LR		max		gen	
	size	power	size	power	size	power	size	power
100 5	0.065	**0.957**	0.069	0.963	0.055	0.457	0.047	0.471
400 20	0.068	**0.997**	0.086	1.000	0.055	0.343	0.051	0.323
800 40	0.066	**1.000**	0.143	1.000	0.053	0.308	0.044	0.292
1600 80	0.045	**1.000**	0.269	1.000	0.067	0.271	0.062	0.247
800 400	0.062	**0.119**	1.000	1.000	0.050	0.066	0.055	0.065
1000 500	0.072	**0.122**	1.000	1.000	0.056	0.071	0.046	0.067
2000 1000	0.066	**0.114**	1.000	1.000	0.046	0.071	0.048	0.053
3200 1600	0.077	**0.132**	1.000	1.000	0.049	0.069	0.059	0.070

Table 2: Size and power comparison under H_2 when $\alpha = 0.05$

(n, p)	CLR		LR		max		gen	
	size	power	size	power	size	power	size	power
100 5	0.065	0.642	0.069	0.668	0.055	**0.786**	0.047	0.783
400 20	0.068	0.473	0.086	0.642	0.055	**0.996**	0.051	0.995
800 40	0.066	0.438	0.143	0.710	0.053	**1.000**	0.044	**1.000**
1600 80	0.045	1.000	0.269	1.000	0.067	**1.000**	0.062	**1.000**
800 400	0.062	0.283	1.000	1.000	0.050	**1.000**	0.055	**1.000**
1000 500	0.072	0.366	1.000	1.000	0.056	**1.000**	0.046	**1.000**
2000 1000	0.066	0.797	1.000	1.000	0.046	**1.000**	0.048	**1.000**
3200 1600	0.077	0.312	1.000	1.000	0.049	**1.000**	0.059	**1.000**

- When y_n is relatively small, CLRT is shown to control the size well, which indeed corrects the traditional LRT. However, if y_n is relatively larger, CLRT's size did not controlled wll for n ranges from hundreds to even thousands. Thus, their convergence rate is fairly slow. For the graph-based method, the size, oppositely, can be controlled well across all scenarios.

- When there is sparse difference in the covariance structure, CLRT performs better than the graph-based method. However, when the difference is not sparse enough, the graph-based method is better. It follows from that the graph methods are more robust by combining information from all dimensions together.
- Concerning the running time, CLRT is more time consuming than graph-based methods, even thousands times slower when dimension is in thousand scale.
- CLRT is limited to $y \in (0, 1)$, while the graph-based methods can allow for $y > 1$.
- CLRT is only designed to detect the covariance difference, while graph-based methods are designed for both mean and covariance differences.

5 Conclusion

Although in practice we may still apply both methods to test the covariance difference between two samples, in principle, when we already know that the structure difference can only be sparse covariance difference, and $y < \frac{1}{2}$, we recommend using CLRT method, otherwise, we had better implement graph-based methods.

References

Bai, Z., Jiang, D., Yao, J.-F., and Zheng, S. (2009). Corrections to lrt on large-dimensional covariance matrix by rmt. *The Annals of Statistics*, 37(6B):3822–3840.

Chen, H. and Friedman, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American statistical association*, 112(517):397–409.

Dempster, A. P. (1958). A high dimensional two sample significance test. *The Annals of Mathematical Statistics*, pages 995–1010.

Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the wald- wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717.

Paul, D. and Aue, A. (2014). Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150(4):1–29.

Zheng, S. (2012). Central limit theorems for linear spectral statistics of large dimensional f-matrices. *Journal of the American statistical association*, 48(2):444 – 476.