A Permutation Test for High-Dimensional Covariance Matrix

Shanshan Li*

University of Electronic Science and Technology of China

*Corresponding author: 201821110328@std.uestc.edu.cn

Abstract. In the case of "big p and small n", classical statistical methods and theories are difficult to apply to high-dimensional data problems. This article considers testing the equality of two sample covariance matrices. When both the dimension p and the sample size n tend to infinity, a permutation method is proposed. The new test method eliminates the limitation of sample distribution and dimension. Numerical research shows that the proposed test method has good results in both normal and non-Gaussian distributions under high-dimensional data.

Keywords: permutation test, covariance matrix, high-dimensional data, Eigenvalue.

1. Introduction

In practical applications, we often encounter situations where the dimension P of the data is getting higher and higher, and far exceeds the number of samples n, that is, p>>n. We call this type of "big p, small n" data High-dimensional data. In classic statistics, it is often assumed that the sample size n is large and the data dimension p is fixed, or p<n, so classic statistics has certain limitations or no longer applicable when dealing with large-dimensional data problems.

In this article, we consider the equality test of the covariance matrices of two populations in the case of high-dimensional data, assuming that $X_1$ and $X_2$ are unknown distributions with dimension p, $X_{11}, X_{12}, \ldots, X_{1n}$ and $X_{21}, X_{22}, \ldots, X_{2n}$ are independent of the populations $X_1$ and $X_2$, respectively. The p-dimensional random vector samples are subject to P and Q respectively. The sample sizes are m and n, where the dimension p may be much larger than $(n_1, n_2)$. The covariance matrices of the population X and Y are $\Sigma_1=\text{Var}(X_1)$ and $\Sigma_2=\text{Var}(X_2)$, respectively. Consider the test for the equality of the two covariance matrices:

$$H_0: \Sigma_1 = \Sigma_2 \quad H_1: \Sigma_1 \neq \Sigma_2$$

(1)

In this article, the hypothesis test is mainly performed on two high-dimensional covariance matrices (1). This article proposes to use permutation test to test the equality of two samples of high-dimensional covariance matrix. In the test problem, the distribution of the samples is not limited, and the dimensionality. The number p is much larger than the sample size n.

This type of test problem usually uses traditional likelihood ratio statistics when low-dimensional $p \leq \min \{n_1, n_2\}$, see Anderson (2003).
\[ T_{LR} = n_1 \log |S_1| + n_2 \log |S_2| - (n_1 + n_2) \log |S| \]

It only converges to the \( \chi^2_{p(p+1)/2} \) distribution when \( p \) is fixed, \( n_1, n_2 \to \infty \) and the population is a normal distribution. Were

\[ S = \frac{n_1}{n_1 + n_2} S_1 + \frac{n_2}{n_1 + n_2} S_2 \quad S_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)' \]

Dempster (1958) [1] first proposed a new statistic for testing two multidimensional populations. Bai and Silverstein (2009) [2] proved that when the dimensionality of the sample does not satisfy \( p<<n \), the eigenvalues of the sample covariance matrix are very different from the eigenvalues of the overall covariance matrix, so high-dimensional data is encountered. At that time, classical statistical methods are no longer applicable.

Schott (2007) [4] is based on the likelihood ratio statistical method. Under the condition that the two populations are normal and the dimension \( p \) and the sample size meet a certain order relationship, the F-norm is used to express the difference between the two covariance matrices. The distance, put forward statistics. Srivastava (2005) [5] assumes that under the condition of a normal population, the covariance matrix is the test statistic of the diagonal matrix.

For the rest of this paper, the basic principles of permutation test are mainly introduced in section 2, and permutation test is applied to the test problem of the equality of two-sample high-dimensional covariance matrix. In Section 3, we conducted a simulation study to prove that this method has high applicability for high-dimensional two-sample covariance test. Section 4 summarizes the paper and discusses the potential of applying the method in this paper to high-dimensional data sets.

2. Test

Li and Chen (2012) [7] transform the test problem (1) into an objective function: \( \text{tr}(\Sigma_1 - \Sigma_2)^2 = 0 \). Therefore, the statistic of hypothesis test (1) can be obtained as \( \text{tr}(\Sigma_1 - \Sigma_2)^2 \), which is the F-norm of \( \Sigma_1 - \Sigma_2 \). In other words, we only need to test:

\[ \text{tr}(\Sigma_1 - \Sigma_2)^2 = \text{tr}(\Sigma_1)^2 + \text{tr}(\Sigma_2)^2 - 2\text{tr}(\Sigma_1 \Sigma_2) \tag{2} \]

Each term of this formula will be estimated separately, and the unbiased estimator of \( \text{tr}(\Sigma_1^2) \) will be obtained as:

\[ A_n = \frac{1}{n_b(n_b-1)} \sum_{i,j} \left( X'_{bi} X_{bj} \right)^2 - \frac{2}{n_b(n_b-1)(n_b-2)} \sum_{i,j,k} \left( X'_{bi} X_{bj} \right) \left( X'_{bi} X_{bk} \right) \]

\[ + \frac{1}{n_b(n_b-1)(n_b-2)(n_b-3)} \sum_{i,j,k,l} X_{bi} X_{bj} X_{bk} X_{bl} \]

Where \( \Sigma^* \) represents the sum of unequal indexes. For example, \( \Sigma^*_{n,j,k} \) represents the sum of items \( \{(i, j, k) : i \neq j, j \neq k, i \neq k \} \). Similarly, the unbiased estimator of \( \text{tr}(\Sigma_1 \Sigma_2) \) is:
\begin{equation}
C_{n_1n_2} = \frac{1}{n_1n_2} \sum_{i,j} (X_{1i}^j X_{2j}) - \frac{1}{n_1(n_1-1)} \sum_{i,k} \sum_{j,k} X_{1i1j} X_{1i2k} X_{1k1j} \\
- \frac{1}{n_2(n_2-1)} \sum_{i,k} \sum_{j,k} X_{2i1j} X_{2i2k} X_{2k1j} + \frac{1}{n_1n_2(n_1-1)(n_2-1)} \sum_{i,k} \sum_{j,k} X_{1i1j} X_{1i2k} X_{2k1j} X_{2k2j}
\end{equation}

Therefore, the test statistics

\begin{equation}
T_{n_1n_2} = A_n + A_{n_2} - 2C_{n_1n_2}
\end{equation}

Obviously, \( T_{n_1n_2} \) is an unbiased estimator of (2). This method does not need to assume a normal distribution, and the conditions of sample dimension \( p \) and sample size \( n_1, n_2 \) are wider, and are suitable for high-dimensional data.

The permutation test can control the probability of making Type I errors within a limited sample. The basic idea is: under the premise that the original hypothesis \( H_0 \) holds, construct a test statistic \( T \), which is different from hypothesis testing. Then replace the sample information according to a certain principle to obtain the replacement sample. Its distribution is the replacement distribution of the test statistic. According to the probability \( P \) of the null hypothesis distribution of the observation sample appearing in the distribution, compare it with a given threshold (usually 0.05).

In this article, we propose algorithm 1, the sample data are \((X_{11}, X_{12}, \ldots, X_{n_1})\) and \((X_{21}, X_{22}, \ldots, X_{n_2})\), where \( N = n_1 + n_2 \). Assuming that \( \sigma^N \) is all permutation possibilities \( (\sigma \in \sigma^N) \), bring the original sample \( Z^N \) into the test statistic \( T_{n_1n_2} \) to get the initial value of the permutation test \( T_0 = T_{n_1n_2}(Z^N) \), bring the permuted sample into the statistic, and recalculate the value of the test statistic \( T_{n_1n_2}(Z_{\sigma}^N) \).

However, we can obtain an approximate p-value by using random permutation to approximate the permutation distribution, so assuming that the number of permutations is \( B \), calculate the value of the statistic \( T_{\text{perm}}^{i} = 1, \ldots, B \). The algorithm is as follows:

**Algorithm 1** The two-sample covariance matrix permutation \( \text{Perm}(T) \)

**Input:** sample \( \{X_{1i}\}_{i=1}^{n_1}, \{X_{2j}\}_{j=1}^{n_2} \), permutations number \( B \), significance level \( \alpha \)

1. Combine two sample matrices \( Z^N = (Z_1, Z_2, \ldots, Z_N) = (X_{11}, X_{12}, \ldots, X_{n_1}, X_{21}, X_{22}, \ldots, X_{n_2}) \)

2. Calculate the initial value of the test statistic \( T_0 = T_{n_1n_2}(Z^N) \)

3. For \( i = 1 \) to \( i = B \), do

   Random replacement, get a replacement sample \( Z_{\sigma}^N = (Z_{\sigma(1)}, Z_{\sigma(2)}, \ldots, Z_{\sigma(N)}) \)

   Calculate the value of the test statistic \( T_i^{\text{perm}} = T_{n_1n_2}(Z_{\sigma}^N) \)

End For

4. The estimated value of the permutation test is approximately:

   \[ p = \frac{1}{B+1} \left( 1 + \sum_{i=1}^{B} I(T_i^{\text{perm}} \geq T_0) \right) \]

6. If \( p < \alpha \), reject the null hypothesis. Otherwise, accept the null hypothesis.

**3. Simulation studies**

In order to study the effect of the test method \( \text{Perm}(T) \) proposed in the previous two chapters and compare it with the methods in other articles, including Li and Chen (2012) [7] using the unbiased
estimator (LC) of Frobenius distance, Cai, Liu and Xia (2013) [8] use the maximum distance (Cai) of the elements of the sample covariance matrix. In this chapter, we use Monte Carlo simulation to estimate the size and power of the traditional test and the test proposed in this article. We set the following simulation:

For \(i = 1, \ldots, n_1\), generate samples by \(X_i = \sum \sqrt{2} Z_i\); for \(j = 1, \ldots, n_2\), generate samples by \(X_j = \sum \sqrt{2} Z_{m+j}\), where \(\{Z_{m+j}\}_{i=1, m+n}\) is a p-dimensional i.i.d random variable composed of \(Z_k\), \(k = 1, \ldots, p\). We consider the following distribution of \(Z_k\): 

(i) Standard normal distribution, so the multidimensional Gaussian samples \(X_1\) and \(X_2\) are obtained.

(ii) For the Gamma distribution, the sample \(Z_k\) is independently and identically distributed in \(\text{Gamma}(2.5,0.5)\).

For the simulation in this article, we calculate the size and power under the volume test program, the sample size is selected as \(n = n_2 = 20, 40, 80\), and the dimension is selected as \(p = 80, 160, 320, 640\). Since it is not feasible to calculate all possible permutations, we use 1000 permutations to calculate the p-value. The size and power of the test are based on 1,000 repetitions. The significance level \(\alpha = 0.05\).

| \(p\) | \(n = n_2\) | Empirical size \(\text{Perm(T)}\) | Empirical size \(\text{LC}\) | Empirical size \(\text{Cai}\) | Empirical power \(\text{Perm(T)}\) | Empirical power \(\text{LC}\) | Empirical power \(\text{Cai}\) |
|------|-------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| 80   | 20          | 0.062                         | 0.059                         | 0.079                         | 0.989                         | 0.950                         | 0.984                         |
|      | 40          | 0.057                         | 0.061                         | 0.058                         | 1                             | 0.987                         | 1                             |
|      | 80          | 0.054                         | 0.062                         | 0.053                         | 1                             | 1                             | 1                             |
| 160  | 20          | 0.054                         | 0.053                         | 0.069                         | 0.972                         | 0.9831                        | 0.960                         |
|      | 40          | 0.048                         | 0.036                         | 0.058                         | 0.997                         | 0.991                         | 0.995                         |
|      | 80          | 0.047                         | 0.047                         | 0.043                         | 1                             | 1                             | 1                             |
| 320  | 20          | 0.053                         | 0.059                         | 0.103                         | 1                             | 0.990                         | 0.964                         |
|      | 40          | 0.051                         | 0.053                         | 0.065                         | 1                             | 1                             | 0.987                         |
|      | 80          | 0.052                         | 0.048                         | 0.052                         | 1                             | 1                             | 0.992                         |
| 640  | 20          | 0.047                         | 0.046                         | 0.114                         | 0.982                         | 0.901                         | 0.863                         |
|      | 40          | 0.053                         | 0.057                         | 0.057                         | 1                             | 0.994                         | 0.975                         |
|      | 80          | 0.051                         | 0.059                         | 0.036                         | 1                             | 1                             | 0.981                         |

Regarding Empirical sizes, whether it is the normal distribution in Table 1 or the Gamma distribution in Table 2, for large-dimensional p or large sample size n, the empirical sizes of \(\text{Perm(T)}\) and LC test are close to 0.05. When n/p is small, the empirical size of the Cai test is \(\geq 0.1\). For high-dimensional normal data, \(\text{Perm(T)}\) and LC tests are tested well in Empirical size, while under non-Gaussian distribution, \(\text{Perm(T)}\) test is more effective.
Table 2. Empirical sizes and powers under Gamma distribution

| p   | $n_1=n_2$ | Empirical size | Power          |
|-----|-----------|----------------|----------------|
|     |           | Perm(T)       | LC             | Cai            |
| 80  | 20        | 0.042 0.039 0.048 | 0.814 0.865 0.377 |
|     | 40        | 0.061 0.050 0.013 | 0.887 0.866 0.498 |
|     | 80        | 0.047 0.039 0.010 | 0.927 0.983 0.615 |
| 160 | 20        | 0.051 0.051 0.139 | 0.802 0.807 0.624 |
|     | 40        | 0.053 0.043 0.075 | 0.992 0.937 0.797 |
|     | 80        | 0.042 0.034 0.040 | 0.995 0.976 0.843 |
| 320 | 20        | 0.035 0.048 0.163 | 0.900 0.889 0.760 |
|     | 40        | 0.046 0.043 0.065 | 0.999 0.976 0.955 |
|     | 80        | 0.051 0.044 0.039 | 1.0 1.0 0.965  |
| 640 | 20        | 0.043 0.052 0.204 | 0.968 0.962 0.805 |
|     | 40        | 0.049 0.054 0.058 | 1.0 0.943 0.927  |
|     | 80        | 0.047 0.052 0.032 | 1.0 0.997 0.994  |

In empirical power, the test based on Perm(T) has greater empirical power than other tests, especially at p=320, this advantage is more obvious, so Perm(T) is the most powerful.

Figure 1. Empirical sizes under normal distribution

Figure 2. Empirical sizes under Gamma distribution

Figure 1 and Figure 2 depict Empirical sizes under the normal distribution and the Gamma distribution, respectively. We can see that the Empirical size of Perm(T) is close to 0.05 regardless of the normal distribution or the non-Gaussian distribution. And as p increases, the Empirical size of Perm(T) fluctuates less, so the performance of Perm(T) is more robust than the other two methods.

4. Conclusions

In this paper, the two-sample covariance matrix equality test in the case of high-dimensional data refers to the test statistics proposed by Li and Chen (2012), which will be applied to the permutation
test. Therefore, there is no need to prove the distribution of statistics. Under the condition that the dimensionality approaches infinity, the test method Perm(T) is proposed. The simulation experiment results show that Perm(T) is more robust when testing the equality of two high-dimensional overall covariance matrices, and it has good results in the case of normal distribution and non-normal distribution.

References
[1] Dempster, A. P. A High Dimensional Two Sample Significance Test [J]. The Annals of Mathematical Statistics, 1958, 29(4):995-1010.
[2] Bai Z D, Saranadasa H. Effect of high dimension: By an example of a two sample problem [J]. Statistica Sinica, 1996, 6(2).
[3] Silverstein B J W. Exact Separation of Eigenvalues of Large Dimensional Sample Covariance Matrices [J]. The Annals of Probability, 1999, 27(3): 1536-1555.
[4] James, R, Schott. A test for the equality of covariance matrices when the dimension is large relative to the sample sizes [J]. Computational Statistics & Data Analysis, 2007.
[5] Srivastava, M. S. Some tests concerning the covariance matrix in high dimensional data [J]. Journal of the Japan Statistical Society, 2005, 35: 251–272.
[6] Srivastava M S, Yanagihara H. Testing the equality of several covariance matrices with fewer observations than the dimension [J]. Journal of Multivariate Analysis, 2010, 101(6): 1319-1329.
[7] Chen S. Two Sample Tests for High Dimensional Covariance Matrices[J]. MPRA Paper, 2012, 105(490): 810-819.
[8] Cai T T, Ma Z. Optimal hypothesis testing for high dimensional covariance matrices [J]. Bernoulli, 2013, 19(5B): 2359-2388.
[9] Cai T, Liu W, Xia Y. Two-sample test of high dimensional means under dependence [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2014, 76(2): 349-372.