Proceedings of the 4th International Conference on Statistics: Theory and Applications (ICSTA'22)
Prague, Czech Republic – July 28- 30, 2022
Paper No. 167
DOI: 10.11159/icsta22.167

Comparison Of Two Mean Vectors Under Differential Privacy For High-Dimensional Data

Caizhu Huang[1], Di Wang[2], Yan Hu[2], Nicola Sartori[1]
[1]Department of Statistical sciences, University of Padova, Padova, Italy
caizhu.huang@phd.unipd.it; nicola.sartori@unipd.it;
[2] Division of Computer, Electrical and Mathematical Sciences and Engineering, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
di.wang@kaust.edu.sa; yan.hu@kaust.edu.sa.

Abstract - The multivariate hypothesis testing problem is a more interesting task of the statistical inference for high-dimensional data nowadays, in which the dimension p of the observation vectors is diverging and could even be larger than the sample size. However, in many applications of multivariate hypotheses problems, the data are highly sensitive and require privacy protection. Here we consider a private non-parametric projection test for the comparison of the high-dimensional multivariate mean vectors that guarantees strong differential privacy. The empirical evidence shows that the non-parametric projection test under differential privacy gives accurate inference under the null hypothesis and a higher power under the local alternative hypothesis.

***Keywords*:** Differential privacy; High-dimensional data; Non-parametric test; Two-sample means.

1. Introduction

With the rapidly development of science and technology, the collection data with a complex structure and sharp increase in data dimensions is widespread in many fields, such as biomedical science, finance and so on [1]. For example, in modern biomedicine, DNA microarrays expression tracks usually thousands of genes simultaneously of each subject. Moreover, there are many applications where the data analyzed includes sensitive information. The data owner releases some synthetical results based on data, without revealing the sensitive information of subjects but reflecting the real distribution of the data as much as possible [2]. Recently, many researches in computer sciences field focus on the statistical hypothesis problem based on differential privacy (DP). Kifer and Rogers (2017) [3] developed a private test statistics based on mahalanobis distance to the hypothesis problem of goodness of fit and independence testing. Martin et al. (2021) [4] showed that the comparison of multivariate population means under DP are of interest in the statistical inference. The bootstrap algorithm for Hotelling T^2 test under DP yields a reliable test decision when the dimension p of the observation vectors is fixed. See also [5]–[7]. However, in high-dimensional data analysis, in which the dimension p is diverging, even larger than the sample sizes, the classical statistical inference under DP, such as Hotelling T^2, chi-square tests, can be not defined since the sample covariance matrix is singular.

To address the comparison of mean vectors under DP in such high dimensional scheme, we here consider several non-private parametric methods. Bai and Saranadasa (1996) [8] proposed a modification of the Hotelling T^2 statistic by using the L_2-norm but not involving the inverse of the covariance matrix which is specified later. See also [1], [9]–[11]. However, the underlying distribution of many real data tends to be non-normal even after log transformation or can be not known. In this sense, it is interested to apply the non-parametric test for high-dimensional mean vectors test without the assumption of underlying distribution of data. Wang and Xu (2021) [12] proposed an approximate randomization test procedure based on the statistics proposed by [11]. Their proposal does not need the condition on the structure of covariance matrix and the balance of sample sizes. However, the randomization test based on the statistic proposed by [11] tends to have unsatisfactory power performance.

In this paper, we develop a non-parametric permutation test based on the projection test under DP for high dimensional mean vectors hypothesis problem. The parametric projection test is proposed by our preview work. Several simulation studies are conducted for comparing the proposed permutation test based on the projection test with the

permutation test based on the tests proposed by [8]. The numerical results show that the non-parametric tests under DP still maintain a high accuracy in terms of the Type I error under the null hypothesis, even in the setting with small sample size. On the other hand, under the local alternative hypothesis, the non-parametric projection test enjoys the powerful performance, while the non-parametric tests based on the statistics proposed by [8] has unsatisfactory power. Indeed, the power behaviour of the non-parametric projection test depends on the choice of the linear coefficient a in the projection statistic.

2. Main Results

Let $X_{i1}, X_{i2}, \dots, X_{in_i} \in \mathbb{R}^p, i = 1,2,$ be independent and identically distributed (iid) p-dimensional random variable with mean vector μ_i and identical covariance matrix $\Sigma = \Gamma\Gamma^T$ with a $p \times q$ matrix Γ. The data $X_i = \{X_{ij}, j = 1, \dots, n_i\}$ generates from the independent component model with structure as follows

$$X_{ij} = \Gamma Z_{ij} + \mu_i, \tag{1}$$

where Z_{ij} is iid q-dimensional random variable with zero mean and unit variance (see [11] for more details). In this paper, we consider a non-parametric projection test under DP for testing the equality of two mean vectors hypothesis

$$H_0: \mu_1 = \mu_2 \quad \text{vs} \quad H_1: \mu_1 = \mu_2, \tag{2}$$

where the dimension p is larger than the sample size $n_i, i = 1,2$. In this sense, the classical Hotelling T^2 test is not available since the sample covariance matrices are not well defined. Here we introduce two suitable statistics for testing problem (2).

First, we introduce the non-private test statistics. Let $\bar{X}_i = \sum_{j=1}^{n_i} X_{ij}/n_i$ denote the sample mean vector of data X_i and $S_n = \{(n_1 - 1)S_1 + (n_2 - 1)S_2\}/(n_1 + n_2 - 2)$ denote the pooled sample covariance matric with the sample covariance matrix $S_i, i = 1,2$. The statistic proposed by [8] takes a form as

$$T_{BS} = (\bar{X}_1 - \bar{X}_2)^T(\bar{X}_1 - \bar{X}_2) - \tau tr(S_n), \tag{3}$$

where $\tau = (n_1 + n_2)/(n_1 n_2)$ and $tr(S_n)$ is a trace operator of matrix S_n. Under some mild conditions, the statistic T_{BS} (3) after suitable location and scale transformation approximates to the null standard normal distribution [8]. In order to improve the power, we use a projection statistic which is based on the L_2 - norm by adding a projection term and is defined by

$$T_{PT} = (\bar{X}_1 - \bar{X}_2)^T(\bar{X}_1 - \bar{X}_2) + k_n(\bar{X}_1 - \bar{X}_2)^T aa^T(\bar{X}_1 - \bar{X}_2), \tag{4}$$

where k_n is a positive constants sequence with $k_n \to +\infty$ and $k_n/\sqrt{p} \to 0$. The linear coefficient vector a is a p-dimensional constant unit vector, i.e. $\|a\| = 1$. Due to the contribution of the second term of (4), the projection statistic (4) achieves to improve the power of the test.

Secondly, we construe the private test statistic for the statistics T_{BS} and T_{PT}. In order to guarantee the DP property, data $X_i, i = 1,2$ both restrict from the p-dimensional cube $[-m, m]^p$ with $m > 0$. Following to Section 3 of [4], we can privatize the statistics by privatizing each entries in T_{BS} and T_{PT} [4] by using the composition theorems of DP. For the privatization of the sample means, we use the popular Laplace Mechanism which results in $\bar{X}_i^{DP} = \bar{X}_i + Y_i$ fulfilling $\varepsilon/2$-DP, if $Y_i = (Y_{i1}, \dots, Y_{ip})^T$ consists of independent random variables $Y_{ik} \sim Lap\left\{0, \frac{2mp}{n_i(\varepsilon/2)}\right\}$. For the privatization of the trace of sample pooled covariance matrix S_n, we also use the Laplace Mechanism to define differentially private estimate $tr^{DP}(S_n) = tr(S_n) + Y$ with the scalar random noise $Y \sim Lap\left\{0, \frac{2mp}{n_i(\varepsilon/2)}\right\}$, satisfying $\varepsilon/2$-DP. Then, the private statistics with respect to T_{BS} and T_{PT} are, respectively,

$$T_{BS}^{DP} = (\bar{X}_1^{DP} - \bar{X}_2^{DP})^T(\bar{X}_1^{DP} - \bar{X}_2^{DP}) - \tau tr^{DP}(S_n),$$

and

$$T_{PT}^{DP} = (\bar{X}_1^{DP} - \bar{X}_2^{DP})^T(\bar{X}_1^{DP} - \bar{X}_2^{DP}) + k_n(\bar{X}_1^{DP} - \bar{X}_2^{DP})^T aa^T(\bar{X}_1^{DP} - \bar{X}_2^{DP}).$$

In order to investigate the performance of the statistics of T_{BS}^{DP} and T_{PT}^{DP} for the high-dimensional hypothesis (2), we now define the corresponding non-parametric version of private statistics mentioned above. According to the theory of the permutation, the probability distribution is invariable between the original data and permuted data under the null hypothesis [13]. Let $X = (X_{11}, \dots, X_{1n_1}, X_{21}, \dots, X_{2n_i})$ and the corresponding permutations denote by X^* which is obtained as $X^* = \{X(u_l^*), l = 1, \dots, n_1 + n_2\}$, where $u_l^* = (u_1^*, \dots, u_{n_1+n_2}^*)$ is a random permutation of unit labels $1, \dots, n_1 + n_2$. In this sense, the corresponding permutation statistics of T_{BS}^{DP} and T_{PT}^{DP} can be defined by using the permuted data X^*, denoted by T_{BS}^{DP*} and T_{PT}^{DP*}, respectively. Therefore, we are able to define the p-value function of the permutation statistics. In order to simplify the abuse notation, we denote the permutation statistics by T^*. In practice, suppose that we have B random permutations. Under the null hypothesis, an unbiased and strongly consistent estimate of p-value can be defined as $\hat{\lambda} = \sum_{k=1}^{B} I(T_k^* \geq T^0)/B$, where T^0 represents the observed value of statistic on the observed data X and $I_A(x)$ is the indicator function of the set A that takes value 1 if $x \in A$ and 0 otherwise. Hence, denote the nominal significant level by α, the null hypothesis is rejected if $\hat{\lambda} < \alpha$.

3. Simulation studies

The performance of the non-parametric projection test for the hypothesis (2) in the high-dimensional framework, in which the dimension p is larger than the sample size $n_i, i = 1,2$, is here assessed via Monte Carlo simulations based on $R = 10{,}000$ replications. The empirical distribution of p-value of non-parametric projection test is compared with a non-parametric competitor proposed by [8]. We also investigate the performance in terms of type I error and the local power.

Samples Z_{ij} in (1) of size $n_i, i = 1,2$ are generate from the p-variate $U(-\sqrt{3}, \sqrt{3})$ under the null hypothesis H_0, where $U(-\sqrt{3}, \sqrt{3})$ is a Uniform distribution with a range between $-\sqrt{3}$ and $\sqrt{3}$ with mean 0 and variance 1. Without loss of generality, suppose that the expectation of data X_i are zeroes under H_0 and the covariance matrix $\Sigma = I_p$ where I_p is $p \times p$ identity matrix. For the non-parametric projection test, let $k_n = \sqrt{p/\log(p)}$ and a has three choices: $a_1 = (1,0, \dots, 0)$, $a_2 = (\sqrt{p/2}1_{p/2}, 0_{1-p/2})$ and $a_3 = \sqrt{p}/p1_p$, where 1_p is a p-variate vector of ones. If p is an odd integer number, we take the ceiling integer of the value of $p/2$. In order to guarantee the stronger privacy, we set up small $\varepsilon = \frac{1}{20}, \frac{1}{4}$. The various simulation setups of the sample sizes and dimension are detailed below:

(1) Under H_0, we set up fixed and small sample size $n = n_1 = n_2 = 5$, and various dimension $p = 20,\ 50,\ 100,\ 150,\ 200$. In this setting, the permutations $B = 252$.
(2) Under H_0, we set up fixed dimension $p = 200$, and various sample size $n = 20, 50, 100, 150$. In this setting, let $B = 10{,}000$.
(3) Under H_1, we consider the two mean vectors as $\mu_1 = 0_p$ and $\mu_2 = \delta 1_p/\sqrt{np}$ with various $\delta = 3,5,7,9$. Here we set up the sample size $n = 5$, $p = 50$.

Table 1 Empirical probability of Type I error for the non-parametric projection test with different a_1, a_2 and a_3 (npPT1, npPT2 and npPT3, respectively) and non-parametric test proposed by [8] (npBS) with the fixed sample size $n = 5$ at the nominal level α = 0.05

	$\varepsilon = 1/20$				$\varepsilon = 1/4$			
p	npPT1	npPT2	npPT3	npBS	npPT1	npPT2	npPT3	npBS
20	0.045	0.047	0.045	0.046	0.045	0.047	0.045	0.046
50	0.042	0.043	0.042	0.042	0.042	0.043	0.042	0.042
100	0.048	0.044	0.047	0.047	0.048	0.044	0.047	0.047
150	0.045	0.044	0.048	0.046	0.045	0.044	0.048	0.046
200	0.042	0.045	0.045	0.043	0.042	0.045	0.045	0.043

Table 1 shows the empirical Type I error for two non-parametric tests with the small sample size. The evidences show that the non-parametric approaches maintain a high accuracy in terms of the Type I error under the null hypothesis. As the sample size increases, the more accurate Type I error is obtained (see Table 2). We also investigate the empirical powers of different statistics under the local alternative hypothesis H_1. Table 3 gives the empirical local power. As δ increases, the power of all tests increase. In particular, the empirical local powers of the non-parametric projection test with the quantity a_2 and a_3 are more powerful than the test proposed by [8] (BS). In addition, the empirical local power of npPT1 is comparable to that of npBS, which confirms that the projection test is powerful when the direction of a is the same as $\mu_1 - \mu_2$.

Table 2 Empirical probability of Type I error for the non-parametric projection test with different a_1, a_2 and a_3 (npPT1, npPT2 and npPT3, respectively) and non-parametric test proposed by [8] (npBS) with the fixed dimension $p = 200$ at the nominal level $\alpha = 0.05$

	$\varepsilon = 1/20$				$\varepsilon = 1/4$			
n	npPT1	npPT2	npPT3	npBS	npPT1	npPT2	npPT3	npBS
20	0.050	0.050	0.051	0.051	0.050	0.050	0.051	0.051
50	0.048	0.050	0.049	0.048	0.048	0.050	0.049	0.048
100	0.052	0.050	0.051	0.050	0.052	0.050	0.052	0.050
150	0.046	0.049	0.048	0.045	0.046	0.049	0.048	0.045

Table 3 Empirical local power for the non-parametric projection test with different a_1, a_2 and a_3 (npPT1, npPT2 and npPT3, respectively) and non-parametric test proposed by [8] (npBS) with the fixed sample size $n = 5$ and dimension $p = 50$.

	$\varepsilon = 1/20$				$\varepsilon = 1/4$			
δ	npPT1	npPT2	npPT3	npBS	npPT1	npPT2	npPT3	npBS
3	0.055	0.085	0.116	0.051	0.055	0.085	0.115	0.051
5	0.065	0.135	0.198	0.062	0.065	0.135	0.197	0.062
7	0.080	0.198	0.271	0.077	0.080	0.197	0.270	0.076
9	0.099	0.251	0.325	0.095	0.099	0.250	0.324	0.094

4. Conclusion

This work extended a non-parametric projection statistic to test the equality of two mean vectors under DP for high dimensional data. The non-parametric projection test is defined by adding a projection term based on L_2-norm of two sample mean vectors, which achieves improved power under some mild conditions. The simulation results show that the non-parametric projection test under DP enjoys a comparable Type I error and a higher power than its competitor even for very small sample sizes. In particular, the non-parametric tests are more accurate in terms of Type I errors when increasing the sample sizes. However, the non-parametric approach costs much time with large sample sizes. In this sense, further research should focus on the asymptotic theory of a parametric test for such high dimensional hypothesis under DP.

References

[1] J.-T. Zhang, J. Guo, B. Zhou, and M.-Y. Cheng, "A Simple Two-Sample Test in High Dimensions Based on L^2-Norm," *J. Am. Stat. Assoc.*, vol. 115, no. 530, pp. 1011–1027, Apr. 2020, doi: 10.1080/01621459.2019.1604366.

[2] R. Busa-Fekete, D. Fotakis, and M. Zampetakis, "Private and Non-private Uniformity Testing for Ranking Data," p. 13.

[3] D. Kifer and R. Rogers, "A New Class of Private Chi-Square Tests," p. 10.

[4] M. Dunsche, T. Kutta, and H. Dette, "Multivariate Mean Comparison under Differential Privacy," *ArXiv211007996 Cs Math Stat*, Oct. 2021, Accessed: Apr. 27, 2022. [Online]. Available: http://arxiv.org/abs/2110.07996

[5] M. Gaboardi, H. W. Lim, R. Rogers, and S. P. Vadhan, "Differentially Private Chi-Squared Hypothesis Testing: Goodness of Fit and Independence Testing," *Proceeding 33rd Int. Conf. Mach. Learn.*, vol. 48, p. 10, 2016.

[6] J. Awan and A. Slavkovic, "Differentially Private Inference for Binomial Data," *ArXiv190400459 Cs Math Stat*, Mar. 2019, Accessed: Mar. 03, 2022. [Online]. Available: http://arxiv.org/abs/1904.00459

[7] S. Couch, Z. Kazan, K. Shi, A. Bray, and A. Groce, "Differentially Private Nonparametric Hypothesis Testing," *ArXiv190309364 Cs Stat*, Mar. 2019, Accessed: Mar. 17, 2022. [Online]. Available: http://arxiv.org/abs/1903.09364
[8] Z. Bai and H. Saranadasa, "EFFECT OF HIGH DIMENSION: BY AN EXAMPLE OF A TWO SAMPLE PROBLEM," p. 19.
[9] L. Zhang, T. Zhu, and J.-T. Zhang, "Two-sample Behrens–Fisher problems for high-dimensional data: a normal reference scale-invariant test," *J. Appl. Stat.*, pp. 1–21, Oct. 2020, doi: 10.1080/02664763.2020.1834516.
[10] X. Cui, R. Li, G. Yang, and W. Zhou, "Empirical likelihood test for a large-dimensional mean vector," *Biometrika*, vol. 107, no. 3, pp. 591–607, Sep. 2020, doi: 10.1093/biomet/asaa005.
[11] S. X. Chen and Y.-L. Qin, "A two-sample test for high-dimensional data with applications to gene-set testing," *Ann. Stat.*, vol. 38, no. 2, Apr. 2010, doi: 10.1214/09-AOS716.
[12] R. Wang and W. Xu, "An approximate randomization test for high-dimensional two-sample Behrens-Fisher problem under arbitrary covariances," *ArXiv210801860 Math Stat*, Dec. 2021, Accessed: Apr. 22, 2022. [Online]. Available: http://arxiv.org/abs/2108.01860
[13] H. Huang and F. Pesarin, "Nonparametric tests for repeated observations with ordered categorical data," p. 12.