

Supplemental material to some permutation tests for high dimensional mean vectors

Abstract

We present additional simulation results analysing the empirical distribution of p -values across various approaches for data X generated from a normal distribution. Additionally, we conduct extended simulation studies on data generated from a gamma distribution. We also examine the empirical type I error of various tests based on larger sample sizes for generating data. The results show that nonparametric tests remain highly accurate even in scenarios involving small sample sizes and strong correlation assumptions. However, parametric tests are unstable in scenarios with stronger correlations.

S1 Simulation studies with small sample sizes

The performance of the nonparametric (i.e. permutation) projection test for the hypotheses (1) in the paper of the high-dimensional framework, where the sample sizes $n_i, i = 1, 2$, are smaller than the dimension p , i.e. $p/n_i \rightarrow [1, \infty)$, is here assessed via Monte Carlo simulations based on $R = 10,000$ replications. Since we only consider the scenario of equal sample sizes, $n_1 = n_2 = 5$, we used all $\binom{10}{5} = 252$ distinct permutations in place of B random permutation. We compare the empirical distribution of the p -value statistic of nonparametric projection tests with two versions of nonparametric competitors proposed by Bai & Saranadasa (1996) and Chen & Qin (2010), as well as with parametric projection tests. We also investigate the performance of all the tests in terms of type I error and local power. For nonparametric and parametric projection tests, we consider $k_n = \sqrt{p/\log(p)}$ and three choices for the quantity α : $\alpha_1 = (1, 0, \dots, 0)$, $\alpha_2 = (1/\sqrt{[p/2]}1_{[p/2]}, 0_{[p/2]})$ and $\alpha_3 = 1/\sqrt{p}1_p$, where 1_p is a p -dimensional vector of 1. In order to facilitate the presentation of results, let us denote the three versions of the nonparametric projection tests as npPT1, npPT2 and npPT3, and the proposed test of Bai & Saranadasa (1996) as npBS and the test of Chen & Qin (2010) as npCQ. Finally, we denote the corresponding parametric projection tests PT1, PT2 and PT3 using α_1 , α_2 and α_3 , respectively.

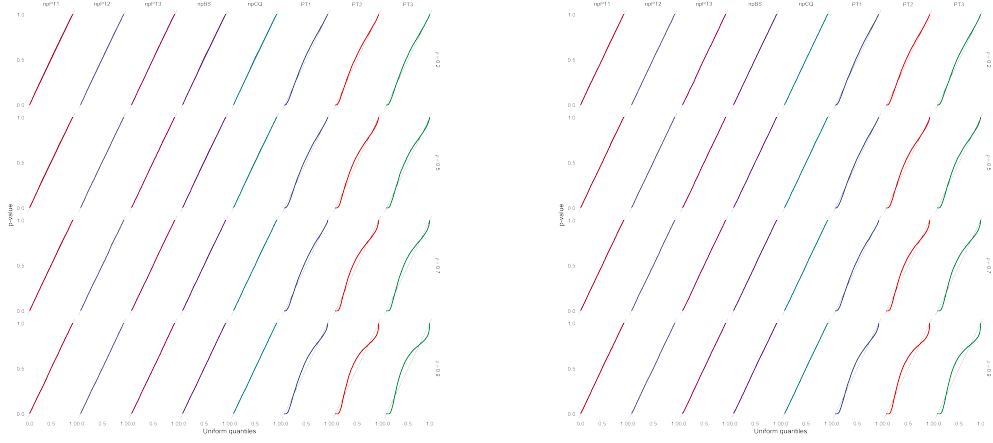


Figure S2: Empirical distribution of p -value for various correlation ρ with small sample sizes $n_1 = n_2 = 5$ of the nonparametric and parametric tests, compared with the $U(0, 1)$ given by the gray diagonal lines. The data X generates from normal (left panel) and gamma (right panel) distributions with $\Sigma = (\sigma_{j,k})_{p \times p} = \rho^{|j-k|}$. The simulated setups are detailed in the setting A3.

S1.1 Simulation results for normal scenario

The left-hand panels of Figures S1 - S3 display the empirical distribution of the p -value for various nonparametric and parametric tests. The results confirm the findings presented in the paper, which show that nonparametric projection tests are more accurate.

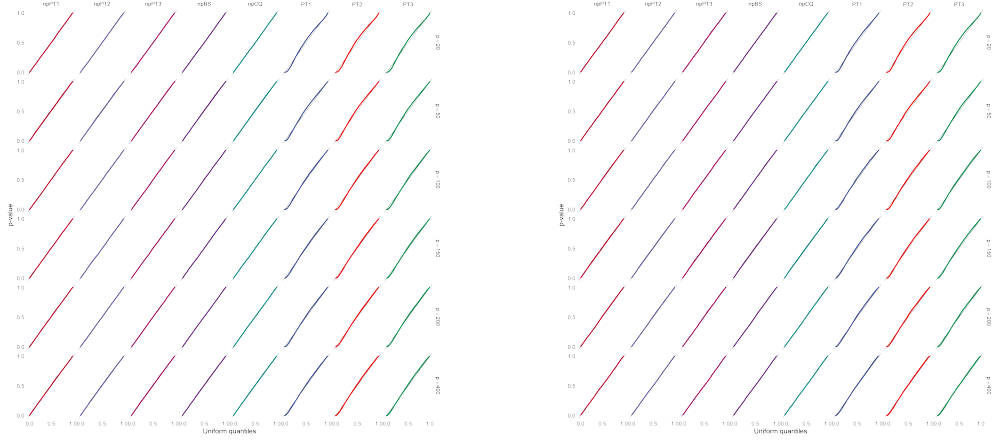


Figure S1: Empirical distribution of p -value for various p of the nonparametric and parametric tests, compared with the $U(0, 1)$ given by the gray diagonal lines. The data X generates from normal (left panel) and gamma (right panel) distributions with $\Sigma = I_p$, respectively. The simulated setups are detailed in the setting A1.

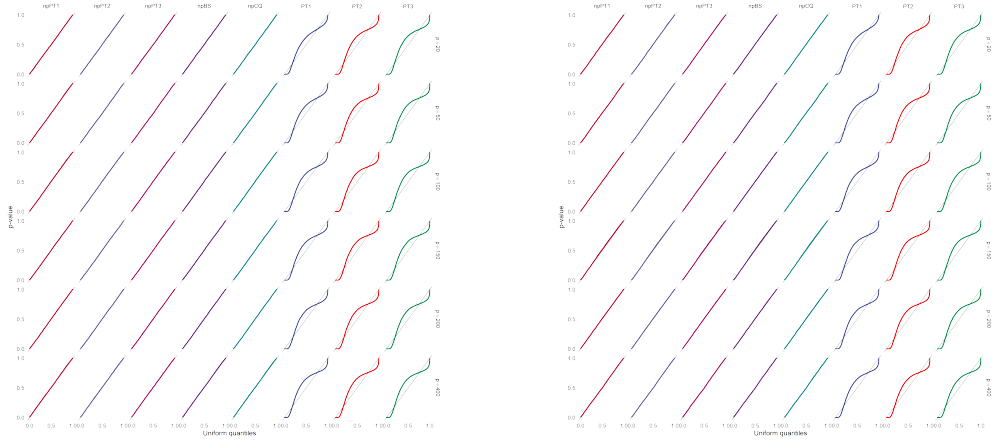


Figure S3: Empirical distribution of p -value for various p of the nonparametric and parametric tests, compared with the $U(0,1)$ given by the gray diagonal lines. The data X generates from normal (left panel) and gamma (right panel) distributions with $\Sigma = \rho I_p + (1 - \rho)1_p 1_p^T$. The simulated setups are detailed in the setting A2.

S1.2 Simulation results for gamma scenario

Extended simulation studies are also considered in a setting where simulated samples of size $n_i, i = 1, 2$, are generated from the gamma distribution, $z_{ijk} \sim Ga(4, 2) - 2$, where $Ga(4, 2)$ denotes a gamma distribution with a shape parameter 4 and a scale parameter 2, under the null hypothesis H_0 . Particular interest is focused on situations where $n_i, i = 1, 2$ is small, as the parametric tests do not perform well in these cases. The details of several simulation scenarios are follows:

- A1 under H_0 , we consider the covariance matrix as $\Sigma = I_p$. We set up fixed and small sample size $n_1 = n_2 = 5$, and dimensions $p \in \{20, 50, 100, 150, 200, 400\}$;
- A2 under H_0 , we consider a general autoregressive correlative structure of the covariance matrix Σ , such as: $\Sigma = (\sigma_{j,k})_{p \times p} = \rho^{|j-k|}$ with correlations $\rho \in \{0.2, 0.5, 0.7, 0.9\}$, and fixed $n_1 = n_2 = 5$ and $p = 50$;
- A3 under H_0 , here we consider a strong correlation covariance matrix as the compound symmetry matrix $\Sigma = \rho I_p + (1 - \rho)1_p 1_p^T$ with $\rho = 0.5$, by fixing the sample size $n_1 = n_2 = 5$ and the dimensions at $p \in \{20, 50, 100, 150, 200, 400\}$;
- A4 under H_1 , we consider the two mean vectors as $\mu_1 = 0_p$ and $\mu_2 = \delta 1_p / \sqrt{n_1 p}$ with size-effect $\delta \in \{3, 5, 7, 9\}$, by setting up the sample size at $n_1 = n_2 = 5$, the dimension at $p = 50$ and the correlation at $\rho = 0$.

Table S1 reports the empirical type I error in scenarios A1, A2 and A3, with data generated from a gamma distribution. The nonparametric approaches perform reliably under H_0 , while the parametric versions of the projection test perform worse in this small sample

Table S1: Empirical type I errors of nonparametric and parametric tests at the nominal level 0.05, with $n_i = 5$. The data X generates from the gamma distribution. Results for scenarios A1, A2 and A3.

A1	p	npPT1	npPT2	npPT3	npBS	npCQ	PT1	PT2	PT3
	20	0.045	0.052	0.045	0.047	0.047	0.093	0.097	0.095
	50	0.046	0.047	0.049	0.046	0.046	0.086	0.092	0.091
	100	0.048	0.048	0.050	0.050	0.050	0.083	0.089	0.087
	150	0.044	0.047	0.045	0.044	0.044	0.078	0.081	0.083
	200	0.049	0.046	0.048	0.047	0.047	0.083	0.082	0.083
	400	0.048	0.046	0.048	0.047	0.047	0.082	0.082	0.080
A2	p	npPT1	npPT2	npPT3	npBS	npCQ	PT1	PT2	PT3
	20	0.048	0.050	0.049	0.048	0.048	0.119	0.124	0.121
	50	0.052	0.053	0.053	0.052	0.052	0.124	0.127	0.126
	100	0.046	0.046	0.047	0.046	0.046	0.116	0.119	0.118
	150	0.050	0.051	0.050	0.050	0.050	0.119	0.121	0.123
	200	0.048	0.047	0.047	0.048	0.048	0.121	0.121	0.122
	400	0.050	0.049	0.049	0.049	0.049	0.122	0.122	0.122
A3	ρ	npPT1	npPT2	npPT3	npBS	npCQ	PT1	PT2	PT3
	0.2	0.047	0.047	0.049	0.050	0.050	0.090	0.094	0.099
	0.5	0.049	0.048	0.050	0.050	0.050	0.095	0.104	0.109
	0.7	0.048	0.052	0.052	0.047	0.047	0.102	0.110	0.116
	0.9	0.051	0.052	0.052	0.051	0.051	0.112	0.119	0.124

size setting. The right-hand panels of Figures S1–S3 display the empirical distribution of p -value statistics for all the considered tests.

The results reported in Table S1 and the right-hand panels of Figures S1 - S3 are similar to those obtained from the normal data in the paper. The nonparametric projection test performs well in all settings.

In addition, we investigate the empirical powers of different statistics under some local alternatives H_1 , as detailed in A4. Figure S4 displays the empirical local power of all tests. The nonparametric and parametric projection tests demonstrated good power. Furthermore, the empirical local power of npPT1 and parametric PT1 tests are comparable to those of the BS and CQ tests, confirming the empirical result by Huang et al. (2024) (see Figure 1). We conjecture that this result could be connected to the fact that, for small sample sizes n and large dimension p , the minimal sufficient statistics for any underlying distribution are almost always the entire $n \times p$ data sets.

S2 Simulation results with larger sample sizes

Extended simulation studies are considered, involving larger sample sizes, in which simulated samples of size $n_i, i = 1, 2$ are generated from the normal distribution $N(0_p, \Sigma)$ and

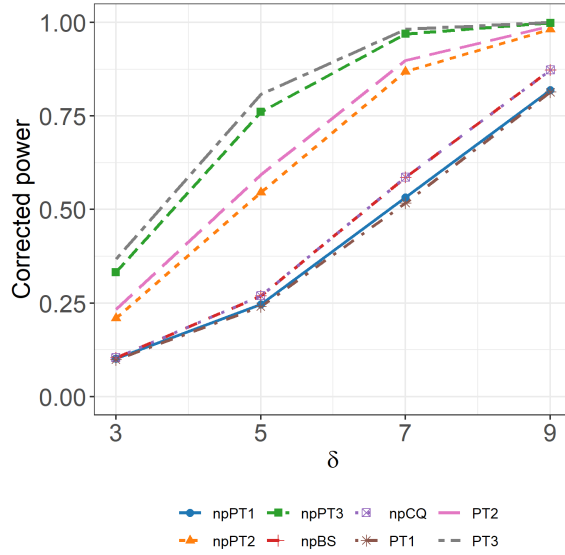


Figure S4: Local power for various δ of the nonparametric and parametric tests. The data X generating from gamma distribution with $\Sigma = I_p$. The simulated setups are detailed in the setting A4.

the gamma distribution with $z_{ijk} \sim Ga(4, 2) - 2$ under H_0 . For the quantities α and k_n of nonparametric and parametric projection tests, we consider the same settings as in Section S1. With a larger sample size, we consider two scenarios.

- A5 under H_0 , we consider the covariance matrix $\Sigma = I_p$. We set up fixed dimension $p = 200$ and various sample sizes $n_i \in \{15, 20, 50, 100, 150, 200\}$, $i = 1, 2$;
- A6 under H_0 , we consider a general autoregressive correlative structure of the covariance matrix $\Sigma = (\sigma_{j,k})_{p \times p} = \rho^{|j-k|}$ with correlations $\rho \in \{0.2, 0.5, 0.7, 0.9\}$ and fixed $n_1 = n_2 = 100$ and $p = 200$.

Figure S5 shows the empirical distribution of the p -value of eight nonparametric and parametric tests with different and larger sample sizes and zero correlation. Nonparametric tests maintain high accuracy across all sample sizes. Parametric tests become more accurate as sample sizes increase. However, Figure S6 shows that parametric tests are not reliable even with larger sample sizes, $n_1 = n_2 = 100$, in a strong correlation setting. As before, nonparametric tests also work in the presence of strong correlations. See Tables S2-S3 for more details about the empirical type I errors of the nonparametric and parametric tests at the nominal level of 0.05.

Table S2: Empirical type I errors of nonparametric and parametric tests at the nominal level 0.05, with $p = 200$. The data X generates from the normal distribution. Results for scenarios A5 and A6.

A5	p	npPT1	npPT2	npPT3	npBS	npCQ	PT1	PT2	PT3
	15	0.050	0.051	0.051	0.049	0.049	0.062	0.063	0.062
	20	0.054	0.049	0.050	0.050	0.050	0.064	0.058	0.062
	50	0.051	0.050	0.050	0.053	0.053	0.060	0.060	0.058
	100	0.052	0.052	0.051	0.051	0.051	0.059	0.059	0.059
	150	0.051	0.049	0.048	0.050	0.050	0.059	0.056	0.056
	200	0.050	0.048	0.048	0.048	0.048	0.057	0.055	0.057
A6	ρ	npPT1	npPT2	npPT3	npBS	npCQ	PT1	PT2	PT3
	0.2	0.052	0.052	0.051	0.052	0.052	0.059	0.062	0.059
	0.5	0.053	0.052	0.051	0.055	0.055	0.061	0.064	0.065
	0.7	0.052	0.050	0.050	0.054	0.054	0.065	0.064	0.065
	0.9	0.050	0.051	0.052	0.050	0.050	0.068	0.066	0.071

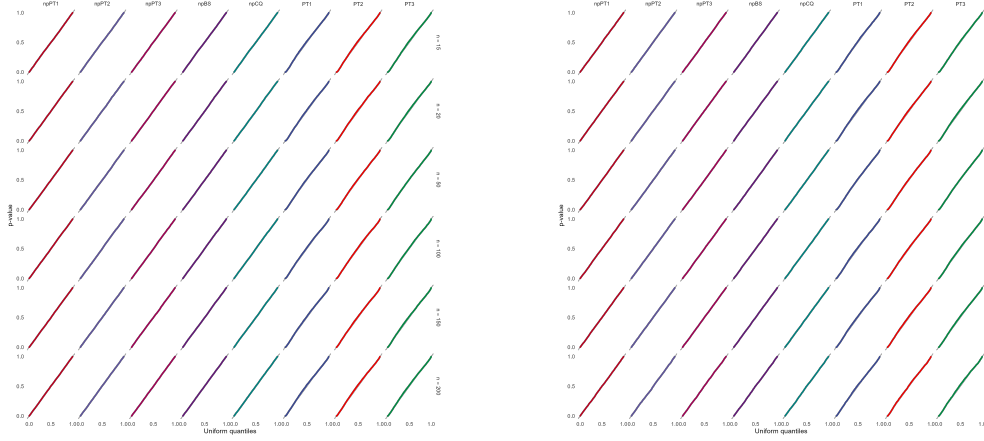


Figure S5: Empirical distribution of p -value for various sample sizes n_i of the nonparametric and parametric tests, compared with the $U(0, 1)$ given by the gray diagonal lines. The data X generates from normal (left panel) and gamma (right panel) distributions with $\Sigma = I_p$, respectively. The simulated setups are detailed in the setting A5.

References

- BAI, Z. & SARANADASA, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statistica Sinica* **6**, 311.
- CHEN, S. & QIN, Y. (2010). A two-sample test for high-dimensional data with application to gene-set testing. *Annals of Statistics* **38**, 808.

Table S3: Empirical type I errors of nonparametric and parametric tests at the nominal level 0.05, with $p = 200$. The data X generates from the gamma distribution. Results for scenarios A5 and A6.

A5	p	npPT1	npPT2	npPT3	npBS	npCQ	PT1	PT2	PT3
	15	0.051	0.052	0.050	0.051	0.051	0.061	0.063	0.062
	20	0.048	0.048	0.050	0.046	0.046	0.058	0.056	0.061
	50	0.055	0.052	0.052	0.054	0.054	0.064	0.060	0.060
	100	0.049	0.051	0.052	0.053	0.053	0.058	0.059	0.059
	150	0.050	0.050	0.048	0.052	0.052	0.058	0.058	0.057
	200	0.048	0.049	0.049	0.050	0.050	0.053	0.056	0.057
A6	ρ	npPT1	npPT2	npPT3	npBS	npCQ	PT1	PT2	PT3
	0.2	0.050	0.047	0.050	0.051	0.051	0.058	0.058	0.061
	0.5	0.049	0.049	0.050	0.050	0.050	0.057	0.064	0.063
	0.7	0.051	0.050	0.051	0.050	0.050	0.061	0.065	0.067
	0.9	0.051	0.050	0.051	0.051	0.051	0.068	0.069	0.069

HUANG, C., CUI, X. & KENNE PAGUI, E. C. (2024). Two-sample mean vector projection test in high-dimensional data. *Computational Statistics* **39**, 1061–1091.

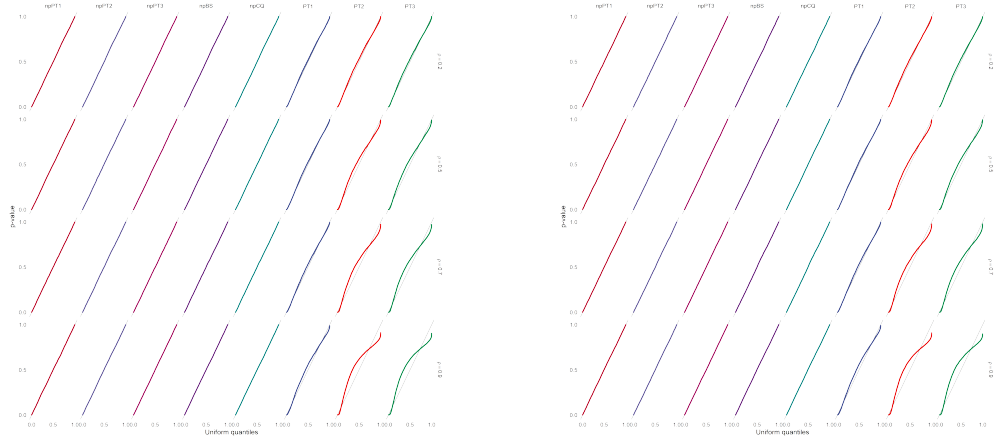


Figure S6: Empirical distribution of p -value for various correlation ρ with larger sample sizes $n_1 = n_2 = 100$ of the nonparametric and parametric tests, compared with the $U(0, 1)$ given by the gray diagonal lines. The data X generates from normal (left panel) and gamma (right panel) distributions with $\Sigma = (\sigma_{j,k})_{p \times p} = \rho^{|j-k|}$. The simulated setups are detailed in the setting A6.