文章编号:2095-6134(2022)05-0586-07

Signed-rank-based test for high dimensional mean vector*

LIU Yan[1,2], LI Shiming[3], ZHANG Sanguo[1,2†]

(1 *School of Mathematical Sciences*, *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*;
2 *Key Laboratory of Big Data Mining and Knowledge Management*, *Chinese Academy of Sciences*, *Beijing* 100049, *China*;
3 *Beijing Tongren Eye Center*, *Beijing Tongren Hospital*, *Capital Medical University*, *Beijing* 100730, *China*)

(Received 8 April 2020; Revised 5 November 2020)

Liu Y, Li S M, Zhang S G. Signed-rank-based test for high dimensional mean vector[J]. Journal of University of Chinese Academy of Sciences, 2022,39(5):586-592. DOI:10.7523/j.ucas.2020.0059.

Abstract This work is concerned with tests for one-sample mean vectors under high dimensional cases. Existing high dimensional tests for mean vectors base on the assumption of elliptical distribution have been proposed recently. To extend to more distributions, we propose a signed-rank-based test. The proposed test statistic is robust and scalar-invariant. Asymptotic properties of the test statistic are established. Numerical studies show that the proposed test has a good control of the type-I error and is more efficiency. We also employ the proposed method to analyze an ophthalmic data.

Keywords high dimensional analysis; signed-rank; one-sample test; scalar-invariance

CLC number: O212.1 **Document code**: A **DOI**: 10.7523/j.ucas.2020.0059

基于符号秩的高维均值检验

刘琰[1,2], 李仕明[3], 张三国[1,2]

(1 中国科学院大学数学科学学院, 北京 100049; 2 中国科学院大数据挖掘与知识管理重点实验室, 北京 100049;
3 首都医科大学附属北京同仁医院眼科中心, 北京 100730)

摘 要 研究高维情形下一样本均值检验的问题。已有的一些高维均值检验方法假设样本具有椭球等高分布。为应用到更多的分布,提出基于符号秩的均值检验统计量。所提方法是稳健的且具有刻度变换不变性。建立了所提出检验统计量的渐近性质,数值模拟表明该方法可以很好地控制第一类错误,且功效更高。还将该方法应用到眼科数据中。

关键词 高维数据分析;符号秩;一样本检验;标度不变性

Suppose that $\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n \in \mathbb{R}^p$ are independent and identically distribution random samples with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. And consider the following test

$$H_0:\boldsymbol{\mu}=\boldsymbol{\mu}_0 \quad \text{vs.} \quad H_1:\boldsymbol{\mu}\neq\boldsymbol{\mu}_0. \tag{1}$$

under $n<p$. This is the so-called "large p, small

* Supported by Beijing Natural Science Foundation (Z190004, JQ20029), Key Program of Joint Funds of the National Natural Science Foundation of China (U19B2040), and Capital Health Research and Development of Special (2020-2-1081)

† Corresponding author, E-mail: sgzhang@ucas.ac.cn

n" paradigm. When p is fixed and under the assumption of normal distribution, a traditional method to test (1) is Hotelling's test statistic. However, Hotelling's test is not defined in the case of $p > n$ because of the singularity of the sample covariance matrix. It is a challenge to the traditional method in high dimensional situation.

The challenge of testing (1) in high dimensional situation has attracted many researchers. Ref. [1] constructed the test statistics which avoid the inverse of the sample covariance matrix. but the test statistics can only be applied to the case of $p/n \to c \in (0,1)$, which means that the increasing rate of the sample dimension should be same as the sample size. Ref. [2] proposed a new test statistic without any direct relationship between p and n. In practice, different components may have different scales. Therefore, scalar-invariant is an important property to a test statistic. Ref. [3], Ref. [4] and Ref. [5] constructed a test statistic with the property of scalar-invariant and under the assumption that $p = o(n^2)$. Ref. [6] proposed a scalar-invariant test that allows the dimension to be arbitrarily large. But their test is not location shift invariant. However, under heavy-tailed distributions, which frequently arise in genomics and quantitative finance, the asymptotic properties of the above test statistics are not established, a natural result is that these tests tend to have unsatisfactory power. Under the assumption of elliptical distributions, Ref. [7] proposed a novel non-parametric test based on spatial-signs, which is more powerful than the test in Ref. [2] for heavy-tailed multivariate distributions and has similar power to the test in Ref. [2] for multivariate normal distribution. But their test is not scalar-invariant. Ref. [8] proposed a novel scalar-invariant test based on multivariate-sign, which is more powerful than the test in Ref. [5] for heavy-tailed multivariate distributions. And their method is under the assumption that $\log(p) = o(n)$.

We propose a novel test for hypothesis (1) based on signed-rank method and our study have two main contributions. Firstly, the proposed test statistic works for more distributions because signed-rank method only requires that the distribution of the samples is symmetric. And the test statistic is available when p is arbitrarily large. Secondly, we show that, under null hypothesis, the proposed test statistic is asymptotically normal. Moreover, the simulation study shows that our method is scalar-invariant and robust, and is more efficient without the assumption of elliptical distributions.

1 A signed-rank-based high dimensional test

1.1 The proposed test statistic

Suppose that $\boldsymbol{X}_i, i = 1, \cdots, n$ are independent and identically distribution random samples with dimension p. We denote that $\boldsymbol{X}^{(k)} = (X_{1k}, \cdots, X_{nk})$, $k = 1, \cdots, p$ as the sample of the k-th dimension. And, let $(r_{1k}, \cdots, r_{nk})$ be the rank of $(|X_{1k}|, \cdots, |X_{nk}|)$. To test hypothesis (1), we proposed a test statistic based on signed-rank functions, which are defined as:

$\boldsymbol{U}_i = \mathrm{diag}\{\mathrm{sign}(X_{i1}), \cdots, \mathrm{sign}(X_{ip})\}(r_{i1}, \cdots, r_{ip})^{\mathrm{T}}$, where $i = 1, 2, \cdots, n$. Then, we consider the following U-statistic:

$$T_n = \frac{\sum_{i \neq j}^{n} U_i^{\mathrm{T}} U_j}{2n(n-1)}. \tag{2}$$

Set $\boldsymbol{s}_i = (s_{i1}, \cdots, s_{ip})^{\mathrm{T}}$ with covariance matrix $\boldsymbol{\Sigma}_s > 0$, where $s_{ij} = \mathrm{sign}(X_{ij})$. To establish the asymptotic properties of the U statistic under the null hypothesis, we need following conditions:

A1. $\mathrm{P}(s_{ij} = 1) = \mathrm{P}(s_{ij} = -1) = \frac{1}{2}$ for each i and j; $|X_{ij}| \neq |X_{kj}|$ for any $i \neq k$ and each j, and $X_{ij} \neq 0$ for each i and j; s_{ij} and r_{ij} are uncorrelated for each i and j.

A2. $\mathrm{tr}(\boldsymbol{\Sigma}_s^4) = o(\mathrm{tr}^2(\boldsymbol{\Sigma}_s^2))$.

Remark 1.1 Condition A1 is necessary condition of the signed-rank test under null hypothesis and it indicates that the random samples have symmetric distributions. Under the first term in condition A1, we have $E(s_{ij}) = 0$. Under the second term in condition A1, $r_{ij} \neq r_{kj}$ for any $i \neq k$ and each j so that $(r_{1j}, \cdots, r_{nj})$ is a permutation of all the

elements in $\{1,\cdots,n\}$. Condition A2 is similar to that applied in Ref. [2], and it is a quite mild condition on the eigenvalues of $\boldsymbol{\Sigma}_s$.

Under H_0, and then suppose condition A1 hold, it is easy to show that

$$E(T_n)=0,$$

and

$$Var(T_n)=\frac{1}{2n(n-1)}\mathrm{tr}(\boldsymbol{\Sigma}_u^2),$$

where $\boldsymbol{\Sigma}_u=E(U_1U_1^{\mathrm{T}})$.

Theorem 1.1 in the following establishes the asymptotic normality of T_n.

Theorem 1.1 Under H_0, and then suppose conditions A1 and A2 hold, as $n\to\infty$ and $p\to\infty$,

$$\frac{T_n}{\sqrt{2n(n-1)\mathrm{tr}(\boldsymbol{\Sigma}_u^2)}}\xrightarrow{d}N(0,1). \qquad (3)$$

Theorem 1.1 implies that we can reject H_0 if $T_n>z_\alpha(2n(n-1)\mathrm{tr}(\boldsymbol{\Sigma}_u^2))^{1/2}$, where z_α is the upper α-quantile of $N(0,1)$. The proof of Theorem 1.1 is conventional, so we omit the details. If someone needs detailed proof, please contact us.

1.2 Computational issue

In practice, in oder to estimate $Var(T_n)$, the estimator for $\mathrm{tr}(\boldsymbol{\Sigma}_u^2)$ is needed. Similar to the estimator used by Ref. [2], we propose the following estimator:

$$\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_u^2)}=\frac{1}{n(n-1)}\mathrm{tr}\Big\{\sum_{j\neq k}^{n}(U_j-\bar{U}_{(j,k)})U_j^{\mathrm{T}}(U_k-\bar{U}_{(j,k)})U_k^{\mathrm{T}}\Big\}$$

where $\bar{U}_{(j,k)}$ is the sample mean of U after excluding U_j and U_k, and $\bar{U}_{(j)}$ is the sample mean of U after excluding U_j. Then, we have the estimator of $Var(T_n)$: $\widehat{Var(T_n)}=\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_u^2)}/2n(n-1)$. And, under H_0, we could show $\widehat{Var(T_n)}/Var(T_n)\xrightarrow{p}1$.

When n and p are large, the computation of $\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_u^2)}$ is too complex. In order to improve the speed of calculation, inspired by Ref. [7], we could transform $\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_u^2)}$ to:

$$\begin{aligned}\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_u^2)}=&-\frac{n}{(n-1)(n-2)^2}\sum_{j=1}^{n}(U_j^{\mathrm{T}}U_j)^2+\\&\frac{1}{n(n-1)(n-2)^2}\Big(\sum_{j=1}^{n}U_j^{\mathrm{T}}U_j\Big)^2+\\&\frac{n-1}{n(n-2)^2}\mathrm{tr}\Big\{\Big(\sum_{j=1}^{n}U_jU_j^{\mathrm{T}}\Big)^2\Big\}+\\&\frac{1-2n}{n(n-1)}\bar{U}^{*\mathrm{T}}\Big(\sum_{j=1}^{n}U_jU_j^{\mathrm{T}}\Big)\bar{U}^*+\\&\frac{4n-2}{n(n-1)(n-2)}\sum_{j=1}^{n}U_j^{\mathrm{T}}U_jU_j^{\mathrm{T}}\bar{U}^*-\\&\frac{2}{n(n-1)}\sum_{j=1}^{n}\bar{U}^{*\mathrm{T}}\bar{U}^*U_j^{\mathrm{T}}U_j+\\&\frac{(n-2)^2}{n(n-1)}\|\bar{U}^*\|^4.\end{aligned}$$

where $\bar{U}^*=\frac{1}{n-2}\sum_{k=1}^{n}U_k$.

2 Simulation study

We compare the performance of the proposed test (SR) with five alternatives: Ref. [1] (BS), Ref. [2] (CQ), Ref. [5] (SKK), Ref. [7] (WPL), Ref. [8] (FZW). All the following simulations are replicated 1 000 times. And, we set $n=20$, 50 and $p=200$, 1 000.

Example 1 We generate $\boldsymbol{X}_i$ from p-variate normal distribution $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Two different choices of Σ are considered as follows: 1) $\boldsymbol{\Sigma}_1=R$; 2) $\boldsymbol{\Sigma}_2=D^{1/2}RD^{1/2}$. Where $R=(\sigma_{jk})$ with $\sigma_{jk}=0.5^{|j-k|}$ for $1\leqslant j,k\leqslant p$, and $D=\mathrm{diag}\{d_1,\cdots,d_p\}$ with $d_j=1I\{1\leqslant j\leqslant p/4\}+2I\{p/4+1\leqslant j\leqslant p/2\}+3I\{p/2+1\leqslant j\leqslant 3p/4\}+4I\{3p/4+1\leqslant j\leqslant p\}$ for $1\leqslant j\leqslant p$. Without loss of generality, we set $\mu_j=\eta$ for $j=1,\cdots,p$, and $\eta=:\|\mu\|^2/\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}^2)}=c$. To calculate the power, we set $c=0.1$ and 0.15 when $n=20$, and $c=0.075$ and 0.1 when $n=50$. And we could calculate the size when we select $c=0$.

Table 1 stands for the performance of the six tests in Example 1. We can see that the power of SR is similar to those of BS, CQ and WPL when $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_1$, and is more than those of BS, CQ and WPL when $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_2$. It indicates that SR has better performance when the scales of different components are different. For example, when $(n,p)=(20,200)$, $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_2$ and $c=0.1$, the power of SR, BS,

Table 1 The empirical size and power at the significance level of 5% in Example 1

(n,p)	$\boldsymbol{\Sigma}$	c	BS	CQ	SKK	WPL	FZW	SR
(20,200)	$\boldsymbol{\Sigma}_1$	0(size)	0.055	0.066	0.044	0.065	0.055	0.061
		0.100	0.395	0.395	0.337	0.427	0.375	0.379
		0.150	0.612	0.607	0.551	0.571	0.579	0.536
	$\boldsymbol{\Sigma}_2$	0(size)	0.066	0.059	0.041	0.058	0.053	0.062
		0.100	0.407	0.420	0.492	0.394	0.550	0.547
		0.150	0.616	0.603	0.759	0.601	0.807	0.779
(20,1 000)	$\boldsymbol{\Sigma}_1$	0(size)	0.067	0.065	0.025	0.066	0.014	0.062
		0.100	0.405	0.398	0.237	0.401	0.164	0.400
		0.150	0.621	0.626	0.413	0.621	0.347	0.589
	$\boldsymbol{\Sigma}_2$	0(size)	0.052	0.050	0.016	0.051	0.010	0.050
		0.100	0.396	0.402	0.390	0.400	0.304	0.571
		0.150	0.626	0.630	0.678	0.626	0.590	0.814
(50,200)	$\boldsymbol{\Sigma}_1$	0(size)	0.051	0.057	0.046	0.044	0.058	0.058
		0.075	0.735	0.737	0.689	0.724	0.789	0.745
		0.100	0.880	0.878	0.837	0.870	0.900	0.871
	$\boldsymbol{\Sigma}_2$	0(size)	0.050	0.065	0.042	0.064	0.059	0.052
		0.075	0.744	0.751	0.871	0.745	0.909	0.889
		0.100	0.891	0.893	0.967	0.885	0.986	0.968
(50,1 000)	$\boldsymbol{\Sigma}_1$	0(size)	0.046	0.063	0.030	0.063	0.051	0.051
		0.075	0.796	0.785	0.692	0.789	0.752	0.738
		0.100	0.927	0.927	0.879	0.929	0.902	0.909
	$\boldsymbol{\Sigma}_2$	0(size)	0.054	0.051	0.031	0.059	0.049	0.058
		0.075	0.786	0.785	0.913	0.784	0.929	0.946
		0.100	0.931	0.910	0.976	0.917	0.992	0.991

CQ, and WPL are 0.547, 0.407, 0.420, and 0.394, respectively. And we observe that SR has better performance in power than SKK and FZW when $p \gg n$. The reason is that SKK and FZW are under the assumptions that p cannot be much larger than n. For example, when $(n,p) = (20,1\ 000)$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1$ and $c = 0.15$, the power of SR, SKK and FZW are 0.589, 0.413 and 0.347 respectively.

Example 2 In this example, $\boldsymbol{X}_i$ is generated from p-variate t-distribution with 3 degrees of freedom. The setting of mean vector $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ are the same as those in Example 1. And we select $c = 0.1$ and 0.15 for $\boldsymbol{\mu}$ to calculate the power.

Table 2 shows the simulation results in Example 2. We can see that SR have better performance in power than that of other five tests in all settings. For example, when $(n,p) = (50,200)$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1$ and $c = 0.15$, the power of SR is 0.773 and the power of the other tests in this setting are 0.419, 0.538, 0.549, 0.577, and 0.610 respectively. For t-distribution is a common heavy-tailed distribution, the results in this table indicate that SR is robust. Table 3 shows the performance of the six tests in Example 3. It shows that SR are more powerful than other five tests in all settings. For example, when $(n,p) = (20,1\ 000)$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_2$, and $c = 0.15$, the power of BS, CQ, SKK, WPL, FZW, and SR are 0.626, 0.615, 0.695, 0.653, 0.650, and 0.949, respectively. Laplace distribution is not a elliptical distribution, and Table 3 shows that SR is more effective in this situation.

Example 3 In this example, $\boldsymbol{X}_i$ is generated from p-variate Laplace distribution. And we consider the same setting of mean vector $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ as those in Example 1. To calculate the power, we select $c = 0.1$ and $c = 0.15$ when $n = 20$, and $c = 0.05$ and $c = 0.075$ when $n = 50$.

Table 3 shows the performance of the six tests in Example 3. It shows that SR are more powerful than other five tests in all settings. For example, when $(n,p) = (20,1\ 000)$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_2$, and $c = 0.15$, the power of BS, CQ, SKK, WPL, FZW, and SR are 0.626, 0.615, 0.695, 0.653, 0.650, and 0.949, respectively. Laplace distribution is not a elliptical distribution, and Table 3 shows that SR is more effective in this situation.

Table 2 The empirical size and power at the significance level of 5% in Example 2

(n,p)	$\boldsymbol{\Sigma}$	c	BS	CQ	SKK	WPL	FZW	SR
(20,200)	$\boldsymbol{\Sigma}_1$	0(size)	0.029	0.047	0.028	0.065	0.046	0.061
		0.10	0.082	0.141	0.112	0.164	0.134	0.200
		0.15	0.109	0.231	0.197	0.213	0.227	0.340
	$\boldsymbol{\Sigma}_2$	0(size)	0.034	0.055	0.032	0.065	0.046	0.058
		0.10	0.070	0.146	0.181	0.163	0.214	0.331
		0.15	0.132	0.217	0.286	0.207	0.311	0.453
(20,1 000)	$\boldsymbol{\Sigma}_1$	0(size)	0.013	0.055	0.007	0.058	0.004	0.049
		0.10	0.032	0.131	0.046	0.140	0.024	0.207
		0.15	0.069	0.160	0.068	0.199	0.042	0.283
	$\boldsymbol{\Sigma}_2$	0(size)	0.018	0.064	0.010	0.061	0.006	0.061
		0.10	0.040	0.138	0.064	0.137	0.036	0.283
		0.15	0.075	0.190	0.148	0.175	0.108	0.473
(50,200)	$\boldsymbol{\Sigma}_1$	0(size)	0.029	0.059	0.030	0.051	0.055	0.064
		0.10	0.253	0.327	0.374	0.390	0.399	0.539
		0.15	0.419	0.538	0.549	0.577	0.610	0.773
	$\boldsymbol{\Sigma}_2$	0(size)	0.035	0.063	0.043	0.061	0.043	0.064
		0.10	0.221	0.320	0.514	0.370	0.594	0.740
		0.15	0.416	0.512	0.783	0.587	0.808	0.906
(50,1 000)	$\boldsymbol{\Sigma}_1$	0(size)	0.021	0.049	0.022	0.052	0.046	0.066
		0.10	0.163	0.343	0.291	0.361	0.332	0.530
		0.15	0.331	0.509	0.512	0.537	0.555	0.789
	$\boldsymbol{\Sigma}_2$	0(size)	0.016	0.056	0.022	0.043	0.057	0.045
		0.10	0.195	0.324	0.474	0.380	0.522	0.768
		0.15	0.328	0.571	0.786	0.555	0.789	0.964

Table 3 The empirical size and power at the significance level of 5% in Example 3

(n,p)	$\boldsymbol{\Sigma}$	c	BS	CQ	SKK	WPL	FZW	SR
(20,200)	$\boldsymbol{\Sigma}_1$	0(size)	0.051	0.068	0.034	0.063	0.045	0.062
		0.100	0.384	0.383	0.334	0.408	0.388	0.492
		0.150	0.586	0.591	0.563	0.617	0.597	0.719
	$\boldsymbol{\Sigma}_2$	0(size)	0.059	0.065	0.039	0.066	0.042	0.057
		0.100	0.364	0.393	0.537	0.403	0.549	0.715
		0.150	0.597	0.588	0.799	0.623	0.789	0.909
(20,1 000)	$\boldsymbol{\Sigma}_1$	0(size)	0.053	0.053	0.012	0.059	0.011	0.060
		0.100	0.380	0.391	0.201	0.407	0.166	0.519
		0.150	0.610	0.603	0.425	0.637	0.356	0.771
	$\boldsymbol{\Sigma}_2$	0(size)	0.053	0.044	0.008	0.056	0.007	0.056
		0.100	0.378	0.391	0.428	0.390	0.317	0.723
		0.150	0.626	0.615	0.695	0.653	0.650	0.949
(50,200)	$\boldsymbol{\Sigma}_1$	0(size)	0.059	0.063	0.042	0.052	0.063	0.058
		0.075	0.508	0.508	0.495	0.515	0.507	0.618
		0.100	0.750	0.748	0.699	0.737	0.766	0.843
	$\boldsymbol{\Sigma}_2$	0(size)	0.053	0.057	0.030	0.055	0.055	0.062
		0.075	0.522	0.519	0.670	0.507	0.701	0.811
		0.100	0.724	0.732	0.892	0.734	0.901	0.977
(50,1 000)	$\boldsymbol{\Sigma}_1$	0(size)	0.052	0.049	0.023	0.053	0.038	0.059
		0.075	0.485	0.523	0.413	0.504	0.472	0.619
		0.100	0.774	0.801	0.690	0.775	0.743	0.895
	$\boldsymbol{\Sigma}_2$	0(size)	0.053	0.060	0.020	0.050	0.046	0.042
		0.075	0.516	0.516	0.663	0.513	0.725	0.864
		0.100	0.798	0.783	0.906	0.789	0.960	0.986

Example 4 In this example, we generate X_i from a mixed distribution. Firstly, we generate Z_{ij} from normal distribution for $1 \leqslant j \leqslant 2p/5$, generate Z_{ij} from t distribution with 3 degrees of freedom for

$2p/5+1 \leqslant j \leqslant 7p/10$, and generate Z_{ij} from Laplace distribution for $7p/10+1 \leqslant j \leqslant p$, and all Z_{ij} have mean 0 and variance 1. Then we let $\boldsymbol{X}_i = \boldsymbol{\Gamma} Z_i + \boldsymbol{\mu}$, where $\boldsymbol{\Gamma}$ is a $p \times p$ matrix with $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}} = \boldsymbol{\Sigma}$, and $Z_i = \{Z_{i1}, \cdots, Z_{ip}\}^{\mathrm{T}}$. And we consider the same setting of mean vector $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ as those in Example 1. To calculate the power, we select $c = 0.1$ and $c = 0.15$ when $n = 20$, and $c = 0.05$ and $c = 0.075$ when $n = 50$.

Table 4 stands for the simulation results in Example 4. We can see that the power of SR is more than those of the other five tests in all settings. For example, when $(n,p) = (50,1\ 000)$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_2$ and $c = 0.075$, the power of SR is 0.757 and the power of the other tests in this setting are 0.214, 0.271, 0.548, 0.299 and 0.613 respectively. In practice, the variates usually have different distributions. Hence, the results in Table 4 indicate that SR is supposed to have better performancein application.

Moreover, we plot the empirical distributions of SR with the settings of four examples and compare them with the standard normal distribution. And, Fig. 1 confirms the asymptotic normal distributions of SR given in Theorem 1.1.

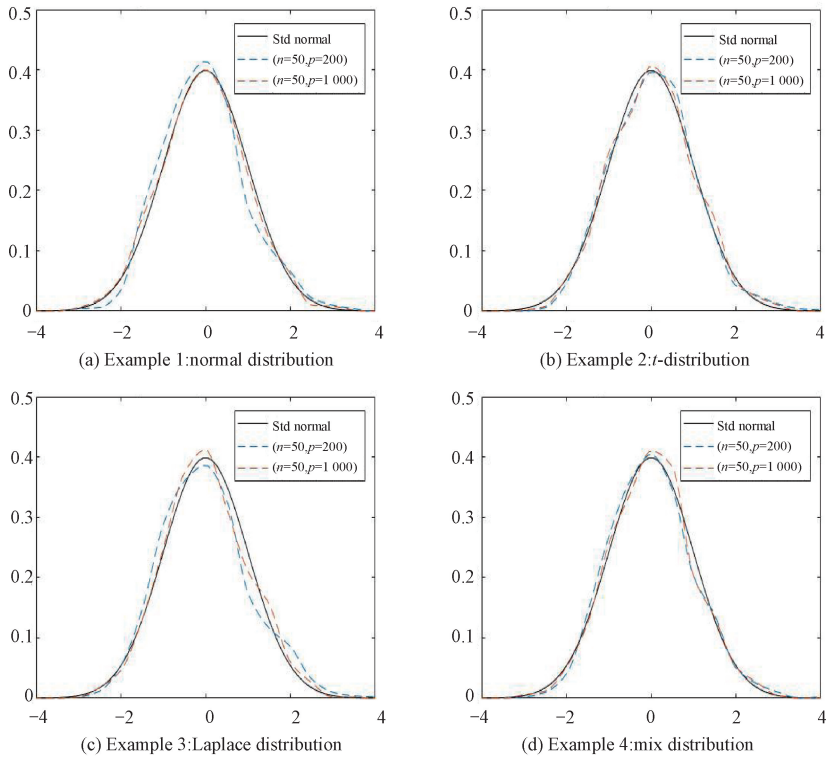
Table 4 The empirical size and power at the significance level of 5% in Example 4

(n,p)	$\boldsymbol{\Sigma}$	c	BS	CQ	SKK	WPL	FZW	SR
(20,200)	$\boldsymbol{\Sigma}_1$	0(size)	0.034	0.064	0.047	0.060	0.065	0.062
		0.100	0.166	0.213	0.270	0.235	0.286	0.369
		0.150	0.259	0.334	0.427	0.362	0.449	0.553
	$\boldsymbol{\Sigma}_2$	0(size)	0.034	0.062	0.028	0.052	0.038	0.046
		0.100	0.167	0.221	0.424	0.239	0.446	0.561
		0.150	0.296	0.374	0.659	0.385	0.683	0.785
(20,1 000)	$\boldsymbol{\Sigma}_1$	0(size)	0.025	0.061	0.015	0.056	0.008	0.053
		0.100	0.112	0.190	0.136	0.207	0.094	0.371
		0.150	0.190	0.338	0.307	0.331	0.223	0.560
	$\boldsymbol{\Sigma}_2$	0(size)	0.031	0.060	0.013	0.063	0.002	0.045
		0.100	0.132	0.218	0.283	0.215	0.200	0.581
		0.150	0.226	0.344	0.579	0.368	0.469	0.841
(50,200)	$\boldsymbol{\Sigma}_1$	0(size)	0.037	0.059	0.040	0.065	0.064	0.059
		0.075	0.225	0.276	0.374	0.302	0.417	0.477
		0.100	0.352	0.417	0.557	0.471	0.624	0.694
	$\boldsymbol{\Sigma}_2$	0(size)	0.050	0.063	0.039	0.053	0.067	0.063
		0.075	0.261	0.335	0.579	0.302	0.627	0.713
		0.100	0.407	0.454	0.820	0.508	0.851	0.893
(50,1 000)	$\boldsymbol{\Sigma}_1$	0(size)	0.038	0.053	0.037	0.055	0.045	0.057
		0.075	0.155	0.254	0.308	0.271	0.363	0.454
		0.100	0.323	0.428	0.525	0.409	0.583	0.719
	$\boldsymbol{\Sigma}_2$	0(size)	0.039	0.056	0.025	0.060	0.042	0.046
		0.075	0.214	0.271	0.548	0.299	0.613	0.757
		0.100	0.368	0.456	0.825	0.517	0.856	0.944

3 Real data application

In this section, we employ the proposed signed-rank-based method to study an ophthalmic data. This data is collected by the Beijing Tongren Eye Center and Anyang Eye Hospital. We take the data of the fifth and sixth grades of a class in the data, Apply the proposed method to study whether the visual factors and their interaction with eye habits are different in different grades.

Firstly, we remove the visual factors and their interaction with eye habits with missing values greater than 15%, and impute the sample mean into the missing values for the remaining 945 factors. Then, we let $\boldsymbol{X}_i$ be the difference between the visual factors and their interaction with eye habits of the i-th student in the sixth grade and those in the fifth grade. And, we calculate standard deviations of each dimension in $\boldsymbol{X}$, and show the distribution of the standard deviations in Fig. 2. It shows that these standard deviations are different, so the scalar-invariance method are supposed to have better performance in the analysis of this data. Applying the proposed SR method, we obtain a p-value $<10^{-9}$,



(a) Example 1:normal distribution
(b) Example 2:t-distribution
(c) Example 3:Laplace distribution
(d) Example 4:mix distribution

Fig. 1 T_n under the null hypothesis with four different distributions of X

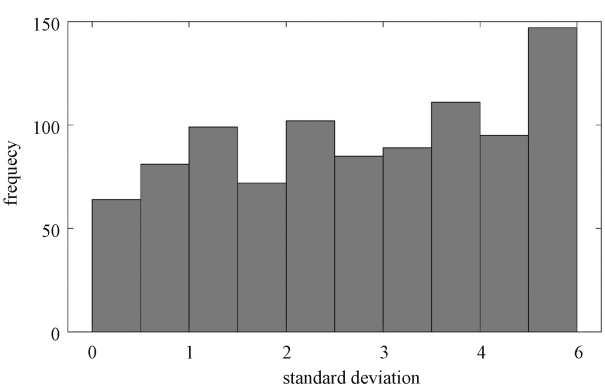


Fig. 2 The distribution of the standard deviations

which illustrates that the visual factors and their interaction factors of eye habits are different in different grades. Through CQ, WPL and FZW methods, the p-values obtained are 0.491 0, 0.491 3 and $<10^{-9}$ respectively. For the standard deviations of each dimension in the sample are different, the CQ and WPL methods are relatively ineffective, while the p-values obtained through FZW and SR methods are small.

References

[1] Bai Z D, Hewa S. Effect of high dimension: by an example of a two sample problem[J]. Statistica Sinica, 1996, 6(2): 311-329.

[2] Chen S X, Qin Y L. A two sample test for high dimensional data with applications to gene-set testing[J]. Annals of Statistics, 2010, 38(2): 808-835. DOI: 10.1214/09-aos716.

[3] Srivastava M S, Meng D. A test for the mean vector with fewer observations than the dimension[J]. Journal of Multivariate Analysis, 2008, 99(3): 386-402. DOI: 10.1016/j.jmva.2006.11.002.

[4] Srivastava M S. A test for the mean vector with fewer observations than the dimension under non-normality[J]. Journal of Multivariate Analysis, 2009, 100(3): 518-532. DOI: 10.1016/j.jmva.2008.06.006.

[5] Srivastava M S, Shota K, Yutaka K. A two sample test in high dimensional data[J]. Journal of Multivariate Analysis, 2013, 114: 349-358. DOI: 10.1016/j.jmva.2012.08.014.

[6] Park J Y, Deepak N A. A test for the mean vector in large dimension and small samples[J]. Journal of Statistical Planning and Inference, 2013, 143(5): 929-943. DOI: 10.1016/j.jspi.2012.11.001.

[7] Wang L, Peng B, Li R Z. A high-dimensional nonparametric multivariate test for mean vector[J]. Journal of the American Statistical Association, 2015, 110(512): 1658-1669. DOI: 10.1080/01621459.2014.988215.

[8] Feng L, Zou C L, Wang Z J. Multivariate-sign-based high-dimensional tests for the two-sample location problem[J]. Journal of the American Statistical Association, 2016, 111(514): 721-735. DOI: 10.1080/01621459.2015.1035380.