A rank-based adaptive independence test for high-dimensional data

Xiangyu Shi[a], Ruiyuan Cao[a], Jiang Du[a,b,*], Zhuqing Miao[a]

[a]*Faculty of Science, Beijing University of Technology, Beijing 100124, China*
[b]*Beijing Institute of Scientific and Engineering Computing, Beijing, 100124, China*

Abstract

There are lots of methods for independence test for high-dimensional random vector. However, it is difficult for practitioner to choose a powerful test because the true alternative hypothesis is unknown. Comnbining the L_2-type with the L_∞-type test statistic, we propose a rank-based test method. From a technical point of view, the proposed test is distribution-free and consequently the corresponding critical values can be obtained by Monte Carlo methods. Compared with permutation or bootstrap test methods, the proposed statistic saves calculation cost. Simulation results show that the resulting method has excellent performance with finite sample size. We also provide a real data application to demonstrate the practicality and effectiveness of the proposed test method.

Keywords: High-dimensional data, rank statistics, L_2-type test statistic, L_∞-type test statistic, independence test.

1. Introduction

The nonparametric association measure is an important and popular tool for the statistician to handle independence test. The correlation coefficients based on rank are the powerful and useful methods for discovering dependent relationship in data. Two classical rank-based tests for nonparametric independence are Spearman's rank correlation coefficient (Hotelling and Pabst [13]) and Kendall's rank correlation coefficient (Kendall [15]). The test methods based on rank have the following several advantages. First, rank-based nonparametric measure does not require strong assumptions on the joint probability dis-

*Corresponding author
Email address: dujiang84@163.com (Jiang Du)

tribution of random vector, for example the second order moment condition. Second, the rank-based nonparametric test statistics are more robust than Pearson's correlation when the data are contaminated with outliers, heavy tailed and skewed distribution. In addition, the resulting test methods are distribution free, consequently the corresponding critical values can always be obtained by Monte Carlo methods.

Let $\boldsymbol{X} = (X_1, \ldots, X_d)^\top \in \mathbb{R}^d$ be a d-dimensional continuous random vector. $\boldsymbol{x}_i (i \in \{1, \ldots, n\})$ is the independent and identically distributed (i.i.d) copy of $\boldsymbol{X}$. Set $\mathcal{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$, n is the sample size. We aim to test the following independence problem,

$$H_0 : X_1, \ldots, X_d \text{ are mutually independent.} \tag{1}$$

In the traditional framework of testing independence (i.e., $n > d$), many classical test statistics exist including the likelihood ratio test (Anderson et al. [1]), Roy's largest root test (Roy [19]) and Nagao's test (Nagao [18]), among others. However, the study of high-dimensional data (i.e., $n < d$) has become particularly important with the increase in computer storage capacity. As traditional statistical methods are no longer applicable, much work on high-dimensional data has been investigated. See, for instance, Han et al. [10], Han et al. [11], Bao [2] and Cai et al. [4].

There are many methods for applied researchers to implement independence test. Schott [20] developed a simple test procedure based on sample correlation matrix for high-dimensional data and proved that the resulting statistic converges to a normal distribution under the null hypothesis. Based on pairwise distance covariance, Yao et al. [25] constructed an L_2-type test to inspect the mutual independence and banded dependence structure of high-dimensional data. It explains the non-linear and non-monotonic dependence of the data on each other. Using the idea of Cramér-type moderate deviation theorem, Drton et al. [6] proposed a maximum test statistic for pairwise rank correlation based on Hoeffding's D(Hoeffding [12]), Blum-Kiefer-Rosenblatt's R(Blum et al. [3]), and Bergsma-Dassios-Yanagimoto's $\tau^\star$(Yanagimoto [24]), and the proposed statistics are rate-optimal under the Gaussian copula model for sparse alternatives. Nevertheless, existing quadratic-based tests often suffer from low power under sparse alternative hypotheses. On the other hand, for dense alternatives, extreme value tests can have low power.

To tackle these challenges, there existed a number of contributions for solving test problem in the high-dimensional data. Among them, based on a screening technique, Fan et al. [8] proposed a new technique called "power-enhanced components" to enhance the power of quadratic statistics under sparse alternatives. Feng et al. [9] presented a max-sum test statistic based on sample correlation to test cross-sectional dependence of high-dimensional panel data. Moreover, they proved the asymptotic independence of the maximum and quadratic test statistic. By combining a class of power-sum tests, Xu et al. [23] proposed a new method for computing adaptive tests of asymptotic p-values, yielding high testing power for various alternative hypotheses in a high-dimensional setting. Chen and Feng [5] combined the L_2-type test and L_∞-type test to propose a Fisher's combination test statistic for considering one-sample means and two-sample means problems.

In real data analysis, it is a difficult task for the applied researchers to choose a powerful and implemental test methods mainly due to the fact that the asymptotic null distribution is unfeasible or the true alternative hypothesis is unknown. Combining L_2-type and L_∞-type test statistics with rank method, we propose a rank-based adaptive independence test method in this paper. Moreover, the algorithm corresponding to the proposed method is presented. The main contributions of this paper are listed as follows.

1. We introduce a combination of L_2-type and L_∞-type test statistics to solve the high dimensional independence problem, and propose a new test method that is adaptive to the underlying data. We call it RAT for short. At the same time, the corresponding algorithm is also provided;

2. Since RAT is based on rank method, the corresponding critical value table with sample size n and dimension d can be obtained via Monte Carlo simulation. This improves the efficiency of the calculation compared with the permutation test or bootstrap test methods;

3. Simulation experiments and example analyses illustrate the well finite sample performance of the proposed method. In particular, we have numerically compared RAT with L_2-type and L_∞-type statistics in our simulation studies.

The rest of this paper is organized as follows. In Section 2, we first review the algorithms for L_2-type and L_∞-type test statistics, and then introduce the algorithmic details of the proposed adaptive test statistic RAT. In Section 3, to demonstrate the finite sample performance of the proposed method, we conduct simulation studies and real data analysis. Section 4 summarizes our conclusions.

2. Methodology

2.1. Review of rank-based tests

The symbol F denotes a joint cumulative distribution function for the continuous random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^\top \in \mathbb{R}^d$ under consideration, and $F_i(i \in \{1, \ldots, d\})$ is the respective marginal cumulative distribution function. Throughout this paper, $\mathcal{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ is a sample consisting of n independent observations of $\boldsymbol{X}$. Let R_{ik} and R_{il} be the rank of x_{ik} and x_{il} for $1 \leq k < l \leq d$ and $i \in \{1, \ldots, n\}$, respectively. We now describe in detail the three types of rank correlation explored in this paper.

Definition 2.1 (Kendall's tau)**.** *Kendall's tau is defined, for $k \neq l \in \{1, \ldots, d\}$, by*

$$\tau_{kl} \equiv \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \mathrm{sign}(x_{ik} - x_{jk}) \mathrm{sign}(x_{il} - x_{jl}), \tag{2}$$

where the sign function $\mathrm{sign}(\cdot)$ is defined as $\mathrm{sign}(x) = x/|x|$ with the convention $0/0 = 0$. Moreover, Mao [17] derived the forms of the first eight-order moments of τ_{kl} under H_0, and then obtained the following results

$$\begin{aligned} &\mathrm{E}_{H_0}(\tau_{kl}) = 0, \\ &\mathrm{Var}_{H_0}(\tau_{kl}) = \mathrm{E}_{H_0}(\tau_{kl}^2) = \frac{2(2n+5)}{9n(n-1)}, \\ &\mathrm{Var}_{H_0}(\tau_{kl}^2) = \frac{8(n-2)(100n^3 + 492n^2 + 731n + 279)}{2025n^3(n-1)^3}. \end{aligned} \tag{3}$$

Definition 2.2 (Spearman's rho)**.** *Spearman's rho is defined, for $k \neq l \in \{1, \ldots, d\}$, by*

$$\rho_{kl} \equiv 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_{ik} - R_{il})^2. \tag{4}$$

According to Mao [16], if X_k is independent of X_l, then

$$
\begin{aligned}
&\mathrm{E}_{H_0}(\rho_{kl}) = 0, \\
&\mathrm{Var}_{H_0}(\rho_{kl}) = \mathrm{E}_{H_0}(\rho_{kl}^2) = \frac{1}{n-1}, \\
&\mathrm{Var}_{H_0}(\rho_{kl}^2) = \frac{2(25n^3 - 57n^2 - 40n + 108)}{25(n+1)n(n-1)^3}.
\end{aligned}
\tag{5}
$$

Definition 2.3 (Spearman's footrule)**.** *Spearman's footrule is defined, for $k \neq l \in \{1, \ldots, d\}$, by*

$$
\varphi_{kl} \equiv 1 - \frac{3}{n^2 - 1} \sum_{i=1}^{n} |R_{ik} - R_{il}|. \tag{6}
$$

In terms of Lemma 3.1 and Lemma 3.2 in Shi et al. [21], under H_0,

$$
\begin{aligned}
&\mathrm{E}_{H_0}(\varphi_{kl}) = 0, \\
&\mathrm{Var}_{H_0}(\varphi_{kl}) = \mathrm{E}_{H_0}(\varphi_{kl}^2) = \frac{2n^2 + 7}{5(n+1)(n-1)^2}, \\
&\mathrm{Var}_{H_0}(\varphi_{kl}^2) = \frac{2(28n^5 - 14n^4 + 172n^2 + 1159n + 1726)}{175(n+1)^3(n-1)^4}.
\end{aligned}
\tag{7}
$$

Based on the above nonparametric association measures, we construct L_2-type and L_∞-type statistics for the independence test. For $\xi_{kl} \in \{\tau_{kl}, \rho_{kl}, \varphi_{kl}\}$, define $\omega_1 = \frac{d(d-1)}{2}\mathrm{E}_{H_0}(\xi_{kl}^2)$, $\omega_2 = \frac{d(d-1)}{2}\mathrm{Var}_{H_0}(\xi_{kl}^2)$ and $\omega_3 = \mathrm{Var}_{H_0}(\xi_{kl})$. Consider the L_2-type test statistics

$$
S_\xi := \omega_2^{-1/2} \left(\sum_{k<l} \xi_{kl}^2 - \omega_1 \right), \tag{8}
$$

and L_∞-type test statistics

$$
M_\xi := \omega_3^{-1} \max_{k<l} \xi_{kl}^2 - 4\log d + \log\log d. \tag{9}
$$

Since the test statistics S_ξ and M_ξ are free of the data-generating process, the corresponding distributions can be approximated via Monte Carlo techniques. Consequently, the critical values or p-value can be obtained via Monte Carlo simulation. Let $\hat{F}^{\tau}_{S,n,d:B}(\cdot)$, $\hat{F}^{\rho}_{S,n,d:B}(\cdot)$ and $\hat{F}^{\varphi}_{S,n,d:B}(\cdot)$ be the empirical distributions, and let $F^{\tau}_{S,n,d:B}(\cdot)$, $F^{\rho}_{S,n,d:B}(\cdot)$ and $F^{\varphi}_{S,n,d:B}(\cdot)$

be their population counterparts. For a given significance level $\alpha \in (0,1)$, the $\alpha/2$ and $1-\alpha/2$ quantiles of S_ξ are provided as follows:

Step 1. For $b \in \{1,\ldots,B\}$, we generate $\mathcal{X}_n^{(b)} \in \mathbb{R}^{n\times d}$ as an $n \times d$ matrix with all entries independently drawn from the standard normal distribution, which yield rank statistics $\{\tau_{kl}^{(b)}, k<l\}$, $\{\rho_{kl}^{(b)}, k<l\}$ and $\{\varphi_{kl}^{(b)}, k<l\}$.

Step 2. With the above rank statistics, we calculate the values of $\omega_2^{-1/2}\left(\sum_{k<l}(\xi_{kl}^{(b)})^2-\omega_1\right)$ in Eq. (8), where $\xi_{kl}^{(b)} \in \{\tau_{kl}^{(b)}, \rho_{kl}^{(b)}, \varphi_{kl}^{(b)}\}$.

Step 3. After collecting each statistic $S_\xi^{(b)}$, the empirical distribution functions $\hat{F}^{\xi}_{S,n,d:B}(\cdot)$ are obtained.

Step 4. According to the definition of quantile, we may obtain the $\alpha/2$ quantiles and the $1-\alpha/2$ quantiles of $\hat{F}^{\xi}_{S,n,d:B}(\cdot)$, i.e.,

$$\begin{aligned}\hat{q}^{\xi}_{\alpha/2;S,n,d} &\equiv \inf\{x : \hat{F}^{\xi}_{S,n,d:B}(x) \geq \alpha/2\},\\ \hat{q}^{\xi}_{1-\alpha/2;S,n,d} &\equiv \inf\{x : \hat{F}^{\xi}_{S,n,d:B}(x) \geq 1-\alpha/2\}.\end{aligned} \tag{10}$$

The L_∞-type test statistics are usually powerful against sparse alternatives. Let $\hat{F}^{\tau}_{M,n,d:B}(\cdot)$, $\hat{F}^{\rho}_{M,n,d:B}(\cdot)$ and $\hat{F}^{\varphi}_{M,n,d:B}(\cdot)$ be the empirical distributions, and let $F^{\tau}_{M,n,d:B}(\cdot)$, $F^{\rho}_{M,n,d:B}(\cdot)$ and $F^{\varphi}_{M,n,d:B}(\cdot)$ be their population counterparts. For a given significance level $\alpha \in (0,1)$, the $(1-\alpha)$ quantiles of M_ξ are provided as follows:

Step 1. For $b \in \{1,\ldots,B\}$, we generate $\mathcal{X}_n^{(b)} \in \mathbb{R}^{n\times d}$ as an $n \times d$ matrix with all entries independently drawn from a standard normal distribution, which yield rank statistics $\{\tau_{kl}^{(b)}, k<l\}$, $\{\rho_{kl}^{(b)}, k<l\}$ and $\{\varphi_{kl}^{(b)}, k<l\}$.

Step 2. We calculate the values of $\omega_3^{-1}\max_{k<l}(\xi_{kl}^{(b)})^2 - 4\log d + \log\log d$ in Eq. (9), where $\xi_{kl}^{(b)} \in \{\tau_{kl}^{(b)}, \rho_{kl}^{(b)}, \varphi_{kl}^{(b)}\}$.

Step 3. After collecting each statistic $M_\xi^{(b)}$, the empirical distribution functions $\hat{F}^{\xi}_{M,n,d:B}(\cdot)$ are obtained.

Step 4. According to the definition of quantile, we may obtain the $1-\alpha$ quantiles of $\hat{F}^{\xi}_{M,n,d:B}(\cdot)$,

$$\hat{q}^{\xi}_{1-\alpha;M,n,d} \equiv \inf\{x : \hat{F}^{\xi}_{M,n,d:B}(x) \geq 1-\alpha\}. \tag{11}$$

Hence, we propose the following level α tests $S_{\xi,\alpha}$ and $M_{\xi,\alpha}$ under H_0:

$$\begin{aligned} S_{\xi,\alpha} &\equiv I\left(S_\xi \leq \hat{q}^{\xi}_{\alpha/2;S,n,d} \text{ or } S_\xi \geq \hat{q}^{\xi}_{1-\alpha/2;S,n,d}\right), \\ M_{\xi,\alpha} &\equiv I\left(M_\xi \geq \hat{q}^{\xi}_{1-\alpha;M,n,d}\right), \end{aligned} \tag{12}$$

where $I(\cdot)$ represents the indicator function. The procedure of simulation-based threshold size-α tests is outlined in Algorithm 1.

To our best knowledge, L_2-type and L_∞-type statistics are sensitive to the alternative hypothesis. Therefore, in the next subsection, we combine L_2-type and L_∞-type statistics to construct a rank-based adaptive test method.

2.2. *The rank-based adaptive test (RAT) method*

We now introduce the rank-based adaptive test method to solve the independence test of unknown alternative hypothesis. We first define some notations as follows. For $\xi \in \{\tau, \rho, \varphi\}$, let

$$P_{S_\xi} = 1 - \Phi(S_\xi) \text{ and } P_{M_\xi} = 1 - F(M_\xi),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal and $F(\cdot)$ is the Gumbel distribution $\exp\left(-e^{-y/2}/\sqrt{8\pi}\right)$. To enable automatic adaptation to the underlying data, we propose the rank-based adaptive test method,

$$\mathrm{RAT} = \min\{P_{S_\tau}, P_{S_\rho}, P_{S_\varphi}, P_{M_\tau}, P_{M_\rho}, P_{M_\varphi}\}. \tag{13}$$

P_{S_ξ} and P_{M_ξ} are the p-values of S_ξ and M_ξ tests, respectively, but RAT is no longer a genuine p-value. Let $\hat{F}_{n,d:B}(\cdot)$ be the empirical distribution, and let $F_{n,d:B}(\cdot)$ be its population counterpart. Since the rank-based test statistic is free-distribution, the simulation-based critical value with sample size n and dimension d can be obtained by Monte Carlo methods. Hence, we provide the quantile of RAT in Eq. (13).

Algorithm 1 Simulation-based threshold size-α tests

Input: The observed n-sample $\mathcal{X}_n$, a significance level α, the independent sample generation times B.

Output: $S_{\xi,\alpha}$, $M_{\xi,\alpha}$.

1: Randomly generate a set of B i.i.d. samples $\mathcal{X}_n^{(b)}$ from a standard normal distribution(to estimate the quantiles), all independent of $\mathcal{X}_n$.
2: **for** $b = 1, \dots, B$ **do**
3: Compute $S_\xi^{(b)}$ and $M_\xi^{(b)}$ in Eq. (8) and (9).
4: **end for**
5: Collect each statistics $S_\xi^{(b)}$ and $M_\xi^{(b)}$, and compute the empirical distribution functions $\hat{F}^\xi_{S,n,d:B}(\cdot)$ and $\hat{F}^\xi_{M,n,d:B}(\cdot)$.
6: Compute the Monte Carlo estimators $\hat{q}^\xi_{\alpha/2;S,n,d}$, $\hat{q}^\xi_{1-\alpha/2;S,n,d}$ and $\hat{q}^\xi_{1-\alpha;M,n,d}$ of the quantiles $q^\xi_{\alpha/2;S,n,d}$, $q^\xi_{1-\alpha/2;S,n,d}$ and $q^\xi_{1-\alpha;M,n,d}$ based on Eq. (10) and (11), respectively.
7: Compute S_τ, S_ρ and S_φ based on Eq. (8) and reject the null hypothesis if

$$\begin{aligned} S_\tau \le \hat{q}^\tau_{\alpha/2;S,n,d} \text{ or } S_\tau \ge \hat{q}^\tau_{1-\alpha/2;S,n,d};\\ S_\rho \le \hat{q}^\rho_{\alpha/2;S,n,d} \text{ or } S_\rho \ge \hat{q}^\rho_{1-\alpha/2;S,n,d};\\ S_\varphi \le \hat{q}^\varphi_{\alpha/2;S,n,d} \text{ or } S_\varphi \ge \hat{q}^\varphi_{1-\alpha/2;S,n,d}. \end{aligned}$$

8: Compute M_τ, M_ρ and M_φ based on Eq. (9) and reject the null hypothesis if

$$M_\tau \ge \hat{q}^\tau_{1-\alpha;M,n,d}, \quad M_\rho \ge \hat{q}^\rho_{1-\alpha;M,n,d} \text{ and } M_\varphi \ge \hat{q}^\varphi_{1-\alpha;M,n,d}.$$

9: **return** $S_{\tau,\alpha}$, $S_{\rho,\alpha}$, $S_{\varphi,\alpha}$, $M_{\tau,\alpha}$, $M_{\rho,\alpha}$, $M_{\varphi,\alpha}$.

Step 1. For $b \in \{1, \ldots, B\}$, we generate $\mathcal{X}_n^{(b)} \in \mathbb{R}^{n \times d}$ as an $n \times d$ matrix with all entries independently drawn from a standard normal distribution, which yield rank statistics $\{\tau_{kl}^{(b)}, k < l\}$, $\{\rho_{kl}^{(b)}, k < l\}$ and $\{\varphi_{kl}^{(b)}, k < l\}$.

Step 2. Calculate the values of $\min\{P_{S_\tau}, P_{S_\rho}, P_{S_\varphi}, P_{M_\tau}, P_{M_\rho}, P_{M_\varphi}\}$ in Eq. (13).

Step 3. After collecting each statistic $\mathrm{RAT}^{(b)}$, the empirical distribution functions $\hat{F}_{n,d:B}(\cdot)$ are obtained.

Step 4. According to the definition of quantile, we may obtain the α quantile of $\hat{F}_{n,d:B}(\cdot)$,

$$\hat{q}_{\alpha;n,d} \equiv \inf\{x : \hat{F}_{n,d:B}(x) \geq \alpha\}. \tag{14}$$

Moreover, we propose the following size-α tests RAT_α under H_0,

$$\mathrm{RAT}_\alpha \equiv I\left(\mathrm{RAT} \leq \hat{q}_{\alpha;n,d}\right). \tag{15}$$

RAT is summarized in Algorithm 2.

Algorithm 2 Simulation-based threshold RAT

Input: The observed n-sample $\mathcal{X}_n$, a significance level α, the independent sample generation times B.

Output: RAT_α.

1: Randomly generate a set of B i.i.d. samples from a standard normal distribution(to estimate the quantiles), all independent of $\mathcal{X}_n$.
2: **for** $b = 1, \ldots, B$ **do**
3: Compute $\mathrm{RAT}^{(b)}$ in Eq. (13)
4: **end for**
5: Collect each statistics $\mathrm{RAT}^{(b)}$, and compute the empirical distribution function $\hat{F}^{\xi}_{n,d:B}(\cdot)$.
6: Compute the Monte Carlo estimator $\hat{q}_{\alpha;n,d}$ of the quantile $q_{\alpha;n,d}$ based on Eq. (14).
7: Compute P_{S_ξ} and P_{M_ξ}.
8: Compute test statistic $\mathrm{RAT} = \min\{P_{S_\tau}, P_{S_\rho}, P_{S_\varphi}, P_{M_\tau}, P_{M_\rho}, P_{M_\varphi}\}$ and reject the null hypothesis if
$$\mathrm{RAT} \leq \hat{q}_{\alpha;n,d}.$$
9: **return** RAT_α.

3. Numerical analysis

3.1. Monte Carlo simulation

In this subsection, we use generated data to numerically invesitage the finite-sample performance of the proposed RAT algorithm comparing with the rank-based method in Eq. (8) and (9).

Set the dimension $d \in \{50, 100, 200, 400\}$ and the sample size $n \in \{50, 100, 200\}$. For each test, the significance level is $\alpha = 0.05$. For each case, the replication number is 1000. The following three examples are considered to generate the data $\mathcal{X}_n = (x_{ij})_{n\times d}$:

Example 1. *The components of the data $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ are independently generated in the manner shown below.*
(a) $\boldsymbol{X} \sim N_d(0, I_d)$ (standard multivariate normal distribution).
(b) $\{X_i\}_{i=1}^d$ are i.i.d. with a $Cauchy(0, 1)$(Cauchy distribution).
(c) $x_{ij} = y_{ij} + z_{ij}$, where $y_{ij} \overset{i.i.d}{\sim} N(0, 1)$. $Z = (z_{ij})_{n\times d}$ is a sample matrix with 5% of entries created independently from $N(0, 100)$ and the remainder 0. Moreover, y_{ij} and z_{ij} are independent.

This example is designed to invesitage the empirical sizes of the proposed RAT test method under different data generation processes.

Example 2. *The data $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ are generated in the manner shown below and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independently generated.*
(a) $\{\boldsymbol{x}_i\}_{i=1}^d \sim N_d(0, \Sigma_\rho)$ with $\Sigma_\rho = \rho I_d + (1-\rho)e_d e_d^\top$ and $\rho = 0.03$, where I_d is the d-dimension identity matrix and e_d is the d-dimensional column vector of ones.
(b) $X_j = Z_j + \frac{1}{10d}\sum_{i\neq j} Z_i$ for $1 \leq j \leq d$ with $Z_1, \ldots, Z_d$ from $Cauchy(0, 1)$.
(c) $x_{ij} = y_{ij} + z_{ij}$. $\{\boldsymbol{y}_i\}_{i=1}^d$ is independently generated from normal distribution $N_d(0, \Sigma_\rho)$ with $\Sigma_\rho = \rho I_d + (1-\rho)e_d e_d^\top$ and $\rho = 0.03$. $Z = (z_{ij})_{n\times d}$ is the sample matrix with 5% of entries created independently from the normal distribution $N(0, 100)$ and the remaining elements being 0.

This example is designed to invesitage the empirical powerful of the proposed RAT test method under dense alternatives.

Example 3. *The data* $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ *are generated in the manner shown below and* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ *are independently generated.*

(a) $\{\boldsymbol{x}_i\}_{i=1}^d \sim N_d(0, \Sigma)$ *with* $\Sigma = (\sigma_{ij})_{d\times d}$, $\sigma_{11} = \ldots = \sigma_{dd} = 1$, $\sigma_{12} = \sigma_{21} = \frac{5}{2}\sqrt{\frac{\log d}{n}}$ *and* $\sigma_{ij} = 0$ *otherwise.*

(b) $X_1 = Z_1 + \sqrt{\frac{\log d}{n}} Z_2$, $X_2 = Z_2 + \sqrt{\frac{\log d}{n}} Z_1$ *and* $X_j = Z_j$ *for* $3 \le j \le d$ *with* $Z_1, \ldots, Z_d$ *from* $Cauchy(0, 1)$.

(c) $x_{ij} = y_{ij} + z_{ij}$. $\{\boldsymbol{y}_i\}_{i=1}^d$ *being generated independently from* $N_d(0, \Sigma)$ *with* $\Sigma = (\sigma_{ij})_{d\times d}$, $\sigma_{ii} = 1$, $\sigma_{12} = \sigma_{21} = \frac{5}{2}\sqrt{\frac{\log d}{n}}$ *and* $\sigma_{ij} = 0$ *otherwise.* $Z = (z_{ij})_{n\times d}$ *is the sample matrix with* 5% *of entries is independently generated from* $N(0, 100)$ *and the rest being* 0.

This example is designed to invesitage the empirical power of the proposed RAT test method under sparse alternatives.

Tables 1-3 show the empirical sizes and powers of the proposed test with a simulation-based critical value ($B = 2000$). The permutation-based strategy is an alternative to the simulation-based approach, but we find that simulations based on the critical null distribution are easier to analyze and have the advantage of allowing the approximation error to be arbitrarily tiny with a larger Monte Carlo sample. The simulation-based technique can save computing expenses because the proposed statistics are constructed based on rank and do not rely on the underlying data generation process.

The empirical size approximates the nominal size, as seen in Table 1. All dimensions are well controlled, thus avoiding the size distortion caused by the slow convergence of the L_∞-type statistics to the extreme distribution. With the dense alternative, Table 2 shows that the L_2-type test statistics have a higher empirical powers, while the L_∞-type has a lower empirical powers, regardless of whether the underlying data are constant-tailed, infinite variance, or outliers. In Table 3 under the sparse alternative, the L_∞-type statistics have higher empirical powers. On the contrary, the L_2-type has a lower empirical powers. Indeed, it is well known that there is no the uniformly most powerful test. Fortunately, in most cases, the RAT statistic performs well under both sparse and dense alternatives, which illustrates the benefits of aggregation.

Table 1: Empirical sizes of tests.

n	d	S_ρ	S_τ	S_φ	M_ρ	M_τ	M_φ	RAT
Example 1(a)								
50	50	0.056	0.060	0.051	0.041	0.044	0.042	0.049
	100	0.049	0.050	0.062	0.048	0.050	0.045	0.061
	200	0.051	0.052	0.052	0.069	0.066	0.061	0.058
	400	0.057	0.059	0.045	0.041	0.036	0.059	0.047
100	50	0.048	0.053	0.053	0.067	0.071	0.062	0.070
	100	0.057	0.051	0.048	0.051	0.051	0.040	0.052
	200	0.044	0.038	0.042	0.051	0.055	0.057	0.047
	400	0.070	0.069	0.079	0.050	0.050	0.055	0.054
200	50	0.048	0.047	0.049	0.039	0.035	0.045	0.063
	100	0.058	0.056	0.056	0.063	0.057	0.047	0.056
	200	0.049	0.049	0.048	0.042	0.049	0.039	0.038
	400	0.056	0.060	0.053	0.046	0.049	0.047	0.035
Example 1(b)								
50	50	0.053	0.055	0.042	0.052	0.058	0.038	0.050
	100	0.049	0.050	0.053	0.039	0.044	0.055	0.050
	200	0.057	0.059	0.056	0.053	0.047	0.062	0.060
	400	0.061	0.061	0.050	0.054	0.037	0.039	0.051
100	50	0.059	0.066	0.066	0.052	0.055	0.044	0.065
	100	0.066	0.070	0.061	0.044	0.041	0.057	0.055
	200	0.059	0.050	0.057	0.052	0.069	0.070	0.074
	400	0.037	0.032	0.047	0.051	0.057	0.052	0.037
200	50	0.047	0.047	0.049	0.050	0.052	0.066	0.065
	100	0.045	0.043	0.050	0.059	0.058	0.049	0.055
	200	0.045	0.043	0.050	0.059	0.058	0.049	0.055
	400	0.054	0.056	0.047	0.052	0.048	0.049	0.048
Example 1(c)								
50	50	0.053	0.053	0.046	0.067	0.070	0.043	0.051
	100	0.040	0.042	0.048	0.037	0.029	0.048	0.046
	200	0.055	0.056	0.056	0.042	0.042	0.049	0.057
	400	0.067	0.069	0.054	0.052	0.043	0.063	0.058
100	50	0.067	0.069	0.054	0.052	0.043	0.063	0.058
	100	0.055	0.060	0.053	0.048	0.050	0.059	0.058
	200	0.052	0.047	0.051	0.061	0.061	0.068	0.055
	400	0.054	0.051	0.059	0.045	0.058	0.051	0.046
200	50	0.046	0.046	0.046	0.047	0.043	0.050	0.057
	100	0.051	0.048	0.043	0.069	0.069	0.043	0.057
	200	0.048	0.048	0.049	0.043	0.040	0.046	0.029
	400	0.057	0.059	0.054	0.048	0.054	0.046	0.055

Table 2: Empirical powers of tests in dense cases.

n	d	S_ρ	S_τ	S_φ	M_ρ	M_τ	M_φ	RAT
Example 2(a)								
50	50	0.181	0.182	0.167	0.058	0.055	0.075	0.262
	100	0.371	0.380	0.391	0.053	0.064	0.105	0.541
	200	0.787	0.792	0.787	0.081	0.083	0.131	0.915
	400	0.980	0.980	0.979	0.076	0.058	0.122	0.996
100	50	0.390	0.410	0.419	0.082	0.083	0.114	0.549
	100	0.851	0.866	0.862	0.079	0.077	0.135	0.905
	200	0.995	0.994	0.995	0.099	0.107	0.183	0.999
	400	1.000	1.000	1.000	0.104	0.110	0.157	1.000
200	50	0.860	0.861	0.859	0.121	0.108	0.145	0.908
	100	0.999	0.998	0.999	0.171	0.170	0.181	1.000
	200	1.000	1.000	1.000	0.117	0.119	0.170	1.000
	400	1.000	1.000	1.000	0.153	0.167	0.217	1.000
Example 2(b)								
50	50	0.873	0.874	0.871	0.287	0.320	0.367	0.910
	100	0.910	0.911	0.913	0.202	0.228	0.393	0.946
	200	0.959	0.961	0.959	0.296	0.314	0.460	0.980
	400	0.977	0.977	0.975	0.270	0.257	0.470	0.988
100	50	0.989	0.990	0.990	0.524	0.560	0.662	0.995
	100	0.999	0.999	0.999	0.490	0.544	0.724	0.999
	200	1.000	1.000	1.000	0.546	0.589	0.721	1.000
	400	1.000	1.000	1.000	0.517	0.557	0.733	1.000
200	50	1.000	1.000	1.000	0.806	0.813	0.910	1.000
	100	1.000	1.000	1.000	0.843	0.853	0.921	1.000
	200	1.000	1.000	1.000	0.818	0.848	0.928	1.000
	400	1.000	1.000	1.000	0.833	0.860	0.945	1.000
Example 2(c)								
50	50	0.124	0.127	0.115	0.066	0.067	0.074	0.184
	100	0.232	0.237	0.241	0.057	0.062	0.081	0.377
	200	0.588	0.604	0.584	0.068	0.067	0.100	0.790
	400	0.935	0.935	0.928	0.055	0.041	0.118	0.987
100	50	0.225	0.238	0.245	0.085	0.086	0.116	0.388
	100	0.643	0.658	0.652	0.079	0.080	0.139	0.759
	200	0.956	0.954	0.960	0.088	0.103	0.155	0.984
	400	1.000	1.000	1.000	0.068	0.078	0.131	1.000
200	50	0.655	0.660	0.654	0.091	0.089	0.126	0.762
	100	0.974	0.974	0.975	0.102	0.096	0.125	0.989
	200	1.000	1.000	1.000	0.088	0.099	0.139	1.000
	400	1.000	1.000	1.000	0.118	0.118	0.188	1.000

Table 3: Empirical powers of tests in sparse cases.

n	d	S_ρ	S_τ	S_φ	M_ρ	M_τ	M_φ	RAT
Example 3(a)								
50	50	0.065	0.068	0.062	0.896	0.907	0.891	0.858
	100	0.063	0.069	0.077	0.958	0.961	0.965	0.939
	200	0.042	0.042	0.044	0.991	0.991	0.990	0.986
	400	0.059	0.060	0.052	1.000	1.000	1.000	1.000
100	50	0.050	0.058	0.060	0.846	0.845	0.807	0.785
	100	0.070	0.066	0.061	0.882	0.892	0.879	0.845
	200	0.037	0.031	0.040	0.916	0.928	0.896	0.873
	400	0.053	0.052	0.061	0.928	0.939	0.920	0.896
200	50	0.068	0.067	0.070	0.795	0.789	0.753	0.744
	100	0.060	0.060	0.056	0.829	0.830	0.753	0.759
	200	0.046	0.045	0.045	0.832	0.837	0.786	0.778
	400	0.055	0.055	0.051	0.889	0.898	0.847	0.835
Example 3(b)								
50	50	0.075	0.077	0.067	0.907	0.942	0.948	0.914
	100	0.061	0.062	0.070	0.911	0.947	0.967	0.929
	200	0.054	0.056	0.057	0.925	0.952	0.966	0.934
	400	0.061	0.063	0.052	0.939	0.964	0.979	0.953
100	50	0.088	0.091	0.097	0.990	0.995	0.997	0.994
	100	0.077	0.080	0.076	0.984	0.991	0.995	0.988
	200	0.059	0.051	0.062	0.984	0.994	0.996	0.990
	400	0.048	0.044	0.054	0.981	0.993	0.994	0.990
200	50	0.105	0.107	0.110	0.997	0.999	0.999	0.999
	100	0.063	0.062	0.065	0.998	0.998	0.999	0.998
	200	0.066	0.063	0.061	0.999	1.000	1.000	1.000
	400	0.051	0.052	0.049	1.000	1.000	1.000	1.000
Example 3(c)								
50	50	0.079	0.079	0.072	0.604	0.648	0.691	0.604
	100	0.057	0.059	0.062	0.624	0.708	0.794	0.684
	200	0.062	0.064	0.060	0.711	0.796	0.871	0.801
	400	0.064	0.067	0.046	0.755	0.858	0.941	0.891
100	50	0.051	0.058	0.061	0.551	0.573	0.613	0.523
	100	0.063	0.068	0.060	0.545	0.566	0.623	0.544
	200	0.054	0.044	0.048	0.595	0.641	0.692	0.613
	400	0.046	0.043	0.053	0.619	0.660	0.731	0.625
200	50	0.064	0.064	0.066	0.475	0.491	0.512	0.474
	100	0.056	0.054	0.051	0.548	0.555	0.539	0.510
	200	0.056	0.055	0.056	0.504	0.524	0.538	0.454
	400	0.067	0.071	0.061	0.538	0.552	0.569	0.493

3.2. Real data analysis

Efron et al. [7] provided a diabetes dataset that included 442 diabetic patients with 10 essential variables, namely age, sex, body mass index, average blood pressure and six blood serum measurements. There are no missing observations in the sample. These data are available from R package 'lars'. The development of a prediction system for diabetic patients requires a set of variables and a statistical model trained by a structured database. As a result, it is only normal to test their total independence prior to train the model. With the methods of this paper, the p-values of the proposed statistics are much less than 0.05, which is strong evidence that the variables considered here are correlated. From a medical point of view, Wilsgaard et al. [22] suggested that there is a correlation between sex, body mass index, and blood pressure, i.e., the effect of weight on blood pressure may differ between men and women. Huang et al. [14] studied the relationship between body mass index, blood pressure and age, and the relative risk of high blood pressure associated with being overweight decreases with age. Thus, our result on the complete correlation between the variables is reasonable.

4. Conclusion

In this paper, a powerful and computationally tractable procedure for testing the mutual independence of random vectors in high-dimensional data was presented that is based on rank method. The resulting test statistic is distribution-free, so that the corresponding critical value or p-value can be obtained by Monte Claro simulation. Simulation results show that our proposed RAT algorithm can be adaptive to the underlying data to test independence compared to the L_2-type and L_∞-type test statistics, and is significantly more computationally efficient compared to a permutation test. The results of the real data analysis are generally consistent with those of the simulation study, demonstrating the feasibility and validity of the proposed RAT algorithm.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets in the paper are public.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 11971045 and 12271014), Young Talent program of Beijing Municipal Education Commission (No. CIT&TCD201904021), and the Science and Technology Project of Beijing Municipal Education Commission (No. KM202210005012).

References

[1] Anderson, T. W., Anderson, T. W., Anderson, T. W., Anderson, T. W., 1958. An introduction to multivariate statistical analysis. Vol. 2. Wiley New York.

[2] Bao, Z., 2019. Tracy–Widom limit for Kendall's tau. Ann. Statist. 47 (6), 3504–3532.

[3] Blum, J. R., Kiefer, J., Rosenblatt, M., 1961. Distribution free tests of independence based on the sample distribution function. Vol. 32. Sandia Corporation.

[4] Cai, Z., Lei, J., Roeder, K., 2023. Asymptotic distribution-free independence test for high dimension data. J. Amer. Statist. Assoc., 1–20.

[5] Chen, D., Feng, L., 2022. Asymptotic independence of the quadratic form and maximum of independent random variables with applications to high-dimensional tests. arXiv preprint arXiv:2204.08628.

[6] Drton, M., Han, F., Shi, H., 2020. High-dimensional consistent independence testing with maxima of rank correlations. Ann. Statist. 48 (6), 3206–3227.

[7] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Statist. 32 (2), 407–451.

[8] Fan, J., Liao, Y., Yao, J., 2015. Power enhancement in high-dimensional cross-sectional tests. Econometrica 83 (4), 1497–1541.

[9] Feng, L., Jiang, T., Liu, B., Xiong, W., 2022. Max-sum tests for cross-sectional independence of high-dimensional panel data. Ann. Statist. 50 (2), 1124–1143.

[10] Han, F., Chen, S., Liu, H., 2017. Distribution-free tests of independence in high dimensions. Biometrika 104 (4), 813–828.

[11] Han, F., Xu, S., Zhou, W.-X., 2018. On Gaussian comparison inequality and its application to spectral analysis of large random matrices. Bernoulli 24 (3), 1787–1833.

[12] Hoeffding, W., 1948. A non-parametric test of independence. Ann. Math. Statist. 19 (4), 546–557.

[13] Hotelling, H., Pabst, M. R., 1936. Rank correlation and tests of significance involving no assumption of normality. Ann. Math. Stat. 7 (1), 29–43.

[14] Huang, Z., Willett, W. C., Manson, J., Rosner, B. A., Stampfer, M. J., Speizer, F. E., Colditz, G. A., 1998. Body Weight, Weight Change, and Risk for Hypertension in Women. Ann. Intern. Med. 128 (2), 81–88.

[15] Kendall, M. G., 1938. A new measure of rank correlation. Biometrika 30 (1/2), 81–93.

[16] Mao, G., 2017. Robust test for independence in high dimensions. Comm. Statist. Theory Methods 46 (20), 10036–10050.

[17] Mao, G., 2018. Testing independence in high dimensions using Kendall's tau. Comput. Stat. Data Anal. 117, 128–137.

[18] Nagao, H., 1973. On some test criteria for covariance matrix. Ann. Statist. 1 (4), 700–709.

[19] Roy, S. N., 1957. Some aspects of multivariate analysis. Statistical Publishing Society, Kolkata.

[20] Schott, J. R., 2005. Testing for complete independence in high dimensions. Biometrika 92 (4), 951–956.

[21] Shi, X., Xu, M., Du, J., 2023. Max-sum test based on Spearman's footrule for high-dimensional independence tests. Comput. Stat. Data Anal. 185, 107768.

[22] Wilsgaard, T., Schirmer, H., Arnesen, E., 2000. Impact of body weight on blood pressure with a focus on sex differences: the Tromso Study, 1986-1995. Arch. Intern. Med. 160 (18), 2847–2853.

[23] Xu, G., Lin, L., Wei, P., Pan, W., 2016. An adaptive two-sample test for high-dimensional means. Biometrika 103 (3), 609–624.

[24] Yanagimoto, T., 1970. On measures of association and a related problem. Ann. Inst. Statist. Math. 22 (1), 57–63.

[25] Yao, S., Zhang, X., Shao, X., 2018. Testing mutual independence in high dimension via distance covariance. J. R. Stat. Soc. Ser. B. Stat. Methodol. 80 (3), 455–480.