Journal of Multivariate Analysis 178 (2020) 104627

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Testing for independence of high-dimensional variables: ρV-coefficient based approach

Masashi Hyodo [a,*], Takahiro Nishiyama [b], Tatjana Pavlenko [c]

[a] Faculty of Economics, Kanagawa University, 3–26–1 Rokkakubashi, Kanagawa-ku, Yokohama-shi, Kanagawa, 221–8624, Japan
[b] Department of Business Administration, Senshu University, 2–1–1 Higashimita, Tama-ku, Kawasaki-shi, Kanagawa, 214–8580, Japan
[c] Department of Mathematics, KTH Royal Institute of Technology, SE-100 44, Stockholm, Sweden

ARTICLE INFO

Article history:
Received 25 April 2019
Received in revised form 16 April 2020
Accepted 16 April 2020
Available online 25 April 2020

AMS 2010 subject classifications:
primary 62H15
secondary 62H20

Keywords:
RV-coefficient
Testing hypotheses
Multiple comparison procedure

ABSTRACT

We treat the problem of testing mutual independence of k high-dimensional random vectors when the data are multivariate normal and $k \geq 2$ is a fixed integer. For this purpose, we focus on the vector correlation coefficient, ρV and propose an extension of its classical estimator which is constructed to correct potential sources of inconsistency related to the high dimensionality. Building on the proposed estimator of ρV, we derive the new test statistic and study its limiting behavior in a general high-dimensional asymptotic framework which allows the vector's dimensionality arbitrarily exceed the sample size. Specifically, we show that the asymptotic distribution of the test statistic under the main hypothesis of independence is standard normal and that the proposed test is size and power consistent. Using our statistics, we further construct the step-down multiple comparison procedure based on the closed testing strategy for the simultaneous test for independence. Accuracy of the proposed tests in finite samples is shown through simulations for a variety of high-dimensional scenarios in combination with a number of alternative dependence structures. Real data analysis is performed to illustrate the utility of the test procedures.

© 2020 Published by Elsevier Inc.

1. Introduction

Testing independence of random variables is a standard task of statistical inference which naturally arises whenever it is needed to handle the dependence structures in multivariate data. Test of independence based on the product–moment correlation was initially explored in the classical seminar paper by Karl Pearson [19], followed by a substantial amount of research literature regarding this topic and its variants. One specific problem which emerges in contemporary applications is the test of independence of k p-dimensional random vectors, where $k \geq 2$ is an integer representing the number of underlying populations. In this study, we address this issue and propose the test of significance based on the high-dimensional extension of the ρV vector correlation, initially introduced by Escoufier [6] for characterizing the relationship of random vectors with a scalar measure of multivariate dependence. Based on the extended estimator of ρV and its asymptotic theory, we further develop two types of tests of independence of k random vectors in arbitrarily high dimensions, and show that both tests apply whether $p \geq n$ or $p < n$ settings, where n denotes the sample size.

* Corresponding author.
E-mail address: hyodoh_h@yahoo.co.jp (M. Hyodo).

https://doi.org/10.1016/j.jmva.2020.104627
0047-259X/© 2020 Published by Elsevier Inc.

1.1. Background and motivation

Extensive overview of the classical, large n and fixed p independence testing techniques is provided in the textbooks on multivariate statistical analysis, see e.g., Mardia et al. [18], Anderson [2], Fang and Zhang [7], and references there in. But, due to ever growing need of analyzing high- and ultra-high dimensional data, examples of applied areas include signal processing, astronomy, functional genomics and proteomics, just a few to name, the development of high-dimensional extensions of the classical testing procedures is of crucial importance. For instance, in functional genomics, multiple and high-dimensional data sets are frequently generated on the same samples of the biological system. This naturally calls for data fusion techniques which make it possible to extract the mutual information from all datasets simultaneously. The first step of the fusion strategy is to accurately identify whether certain similarities of the configuration of the samples (i.e., dependencies) occur between the datasets. Thus, it is necessary to develop novel testing methodologies suitable for testing the independence between such pairs of high-dimensional data sets. Another example motivating the research of this paper is discussed by Efron [5], who analyzed effects of the independence assumption for Cardio microarrays data comprising $n = 63$ arrays and $p = 20\,426$ genes. Starting with the presumption of independence across microarrays, which underlies most of conventional statistical inferential procedures, Efron demonstrated that the presence of dependence can invalidate the usual choice of a null hypothesis, leading to flawed assessments of significance. Hence, before conducting further high-dimensional statistical analyses such as classification, testing hypothesis of equality of mean vectors and covariance matrices, it is important to know when independence fails. For this purpose, testing procedures that are able to cope with nowadays $p \gg n$ data must be designed.

Our focus in this paper is on testing mutual independence of multivariate components building on the high-dimensional extension of ρV. As for the review of the existing literature on the subject of our study, we refer to Josse et al. [15] who considered ρV-based independence testing and argued for the permutation test strategy to approximate the distribution of the test statistic.

Further relevant approaches include Jiang et al. [13] who employed a high-dimensional correction of LRT to construct the test of independence of two vectors. However, the asymptotic theory of these corrected LRT statistic, such as its distribution under null hypothesis, is restricted to a bounded limiting ratio for both sub-vectors, i.e., to the high-dimensional case where $p_k/n \to c_k \in (0, 1]$, $k \in \{1, 2\}$. Testing the independence of two normal sub-vectors based on the structure of the covariance matrix was considered by Srivastava and Reid [21], and further generalized by Hyodo et al. [12] to test the independence of k sub-vectors. Yang and Pan [23] presented the independence test based on the sum of regularized sample canonical correlation coefficients. Testing of independence that does not require normality and is based on the distance correlation are presented by Székely and Rizzo [22]. Non-parametric approaches to the problem of testing independence can be found in e.g., Han and Liu [8] who treated the maxima of rank correlations measure, such as Kendall's tau, and Leung and Drton [16] who used the framework of U-statistics and propose a family of test statistics which is based on sum of squares of sample rank correlations such as Kendall's tau, Hoeffding's D statistics and a dominating term of Spearman's ρ.

1.2. Preliminaries and notations

In what follows, we focus on the more precise problem statement, after some prefatory notations are in place. Henceforth, for an integer $k \geq 2$, we will denote by $[k]$ the set $\{1, \ldots, k\}$. Let $\mathbf{x} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_k^\top)^\top$ denote a $(p \times k)$-dimensional random vector, in which $\mathbf{x}_g$ possesses a dimension p for each $g \in [k]$. Denote further by $\boldsymbol{\mu}_g$, by $\boldsymbol{\Sigma}_{gg}$, and by $\boldsymbol{\Sigma}_{gh}$, the mean vector of the gth sub-vector of $\mathbf{x}$, the covariance matrix of the gth sub-vector of $\mathbf{x}$, and the cross-covariance matrix of $\mathbf{x}_g$ and $\mathbf{x}_h$, respectively, for $g \neq h \in [k]$. Then $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_k^\top)^\top$ and $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_{gh})$, $g, h \in [k]$ are the mean vector and covariance matrix of $\mathbf{x}$, respectively. We are interested in testing the following hypothesis

$$\mathcal{H}: \quad \forall g, h \in [k] \quad \mathbf{x}_g \text{ and } \mathbf{x}_h \text{ are independent} \quad \text{vs.} \quad \mathcal{A}: \neg\mathcal{H}. \tag{1}$$

To this end, we draw a sample of independent observations of $\mathbf{x}$ using the following sampling scheme. Without loss of generality, we first assume that $1 \leq n_1 \leq \cdots \leq n_k$ and set $n_0 = 0$. Further, $\forall h \in [k]$ and $\forall j \in \{n_{h-1} + 1, \ldots, n_h\}$, we denote $p(k - h + 1)$ dimensional vectors by $\mathbf{x}_{\langle h\rangle j} = (\mathbf{x}_{hj}^\top, \ldots, \mathbf{x}_{kj}^\top)^\top$ and $\boldsymbol{\mu}_{\langle h\rangle} = (\boldsymbol{\mu}_h^\top, \ldots, \boldsymbol{\mu}_k^\top)^\top$, respectively. By considering a partition of $\boldsymbol{\Sigma}$ which is compatible with $\mathbf{x}_{\langle h\rangle}$ and $\boldsymbol{\mu}_{\langle h\rangle}$, we introduce a (positive definite) matrix

$$\boldsymbol{\Sigma}_{\langle h\rangle} = \begin{pmatrix} \boldsymbol{\Sigma}_{hh} & \cdots & \boldsymbol{\Sigma}_{hk} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{hk} & \cdots & \boldsymbol{\Sigma}_{kk} \end{pmatrix},$$

and assume that $\mathbf{x}_{\langle h\rangle j} \overset{i.i.d.}{\sim} \mathcal{N}_{p(k-h+1)}(\boldsymbol{\mu}_{\langle h\rangle}, \boldsymbol{\Sigma}_{\langle h\rangle})$. In addition, $\mathbf{x}_{\langle 1\rangle 1}, \ldots, \mathbf{x}_{\langle 1\rangle n_1}, \ldots, \mathbf{x}_{\langle k\,\rangle n_{k-1}+1}, \ldots, \mathbf{x}_{\langle k\,\rangle n_k}$ are assumed to be mutually independent across k populations, constituting thereby a sample of independent observations of $\mathbf{x}$ to be used for constructing the test procedure. A simple example of this sampling scheme is shown in Fig. 1.

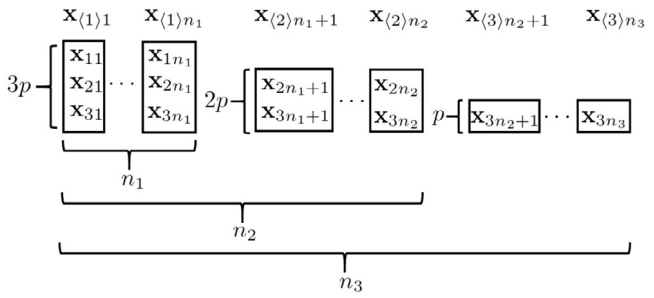

Fig. 1. An example of independent observations for $k = 3$ is illustrated in this figure. The independent observations $\mathbf{x}_{\langle 1\rangle 1}, \ldots, \mathbf{x}_{\langle 1\rangle n_1}, \mathbf{x}_{\langle 2\rangle n_1+1}, \ldots, \mathbf{x}_{\langle 2\rangle n_2}$, $\mathbf{x}_{\langle 3\rangle n_2+1}, \ldots, \mathbf{x}_{\langle 3\rangle n_3}$ are assumed to be mutually independent across 3 populations.

Observe that under the null hypothesis $\mathcal{H}$ of (1) stated in the multivariate normal setting, the population covariance matrix $\boldsymbol{\Sigma}_{\langle h\rangle}$ has all the cross-covariance components $\boldsymbol{\Sigma}_{gh} = \mathbf{O}$ which explicitly represents the classical inferential assumption of independence among k populations. To enhance the presentation, we use the notation

$$\mathbf{D}_{\langle h\rangle} = \operatorname{diag}(\boldsymbol{\Sigma}_{hh}, \ldots, \boldsymbol{\Sigma}_{kk}) = \begin{pmatrix} \boldsymbol{\Sigma}_{hh} & \cdots & \mathbf{O} \\ \vdots & \ddots & \vdots \\ \mathbf{O} & \cdots & \boldsymbol{\Sigma}_{kk} \end{pmatrix},$$

to denote the diagonal block matrix with blocks $\boldsymbol{\Sigma}_{hh}, \ldots, \boldsymbol{\Sigma}_{kk}$, i.e., all off-diagonal blocks of $\mathbf{D}_{\langle h\rangle}$ are $\mathbf{O}$. Here, $\mathbf{O}_{p\times p}$ will be used to denote the $p \times p$ null matrix and will be abbreviated to $\mathbf{O}$ when the dimensionality will be clear from the context. With the aid of these notations, the test of independence (1) can equivalently state as

$$\mathcal{H} : \boldsymbol{\Sigma}_{\langle 1\rangle} = \mathbf{D}_{\langle 1\rangle} \quad \text{vs.} \quad \mathcal{A} : \text{no restrictions on } \boldsymbol{\Sigma}_{\langle 1\rangle}. \tag{2}$$

The natural approach is to design a test statistic that measures the dependence among the components of $\mathbf{x}$ based on the sample, and reject $\mathcal{H}$ when its value is too large, where the critical value of rejection is set according to the asymptotic distribution of the test statistic under the null. Our focus in this paper is on the use of ρV vector correlations in settings where the dimension p is much larger than the sample size n_k. The new statistic we propose for testing $\mathcal{H}$ is constructed as a function of consistent estimators of the pairwise vector correlation coefficients and the corresponding asymptotic theory is developed to obtain the limit null distribution of this statistic with $p, n_1 \to \infty$. The test statistic is presented in the next section, beginning with the high-dimensional adjustment of the estimator of the vector correlation coefficient, followed by the characterization of the test's asymptotic behavior. A simultaneous test of independence is further constructed in Section 3, where the proposed statistic is incorporated into the step-down multiple comparison algorithm. A finite sample performance of the proposed tests is shown in Section 4 through a number of simulation scenarios and application. Finally, we summarize the main results. More technical details and proofs are gathered in the appendix.

Throughout the paper, $\operatorname{tr}(\mathbf{M})$ and $\|\mathbf{M}\|_F^2 = \operatorname{tr}(\mathbf{M}\mathbf{M}^\top)$ represent the trace of a square matrix $\mathbf{M}$ and its squared Frobenius norm, respectively. The symbol $\rightsquigarrow$ denotes convergence in distribution and $\otimes$ denotes Kronecker product.

2. Methodology and theory

Our proposed testing procedures will be studied under the high-dimensional or, as is frequently known as, (n, p)-asymptotic regime where $p = p(n_1)$ grows as a function of n_1 such that p also tends to infinity. Throughout, we denote this asymptotic regime by $p, n_1 \to \infty$.

2.1. Vector correlation coefficient in high-dimensional setting

For any indices $g \neq h \in [k]$, let ρV_{gh} denote the vector correlation coefficient between the two components of $\mathbf{x}$, $\mathbf{x}_g$ and $\mathbf{x}_h$, defined as (see Escoufier [6])

$$\rho V_{gh} = \frac{\|\boldsymbol{\Sigma}_{gh}\|_F^2}{\|\boldsymbol{\Sigma}_{gg}\|_F \|\boldsymbol{\Sigma}_{hh}\|_F}.$$

It is immediately clear that Pearson's product–moment correlation coefficient is a special case ρV_{gh} when $p = 1$. Furthermore, $\rho V_{gh} = \rho V_{hg}$, and $\rho V_{gh} = 0$ if and only if $\boldsymbol{\Sigma}_{gh} = \mathbf{O}$. Let a be any constant, $\mathbf{B}$ be a $p \times p$ matrix such that $\mathbf{B}\mathbf{B}^\top = \mathbf{I}_p$, and $\mathbf{c}$ be any p dimensional constant vector. If we set $\mathbf{x}_h = a\mathbf{B}\mathbf{x}_g + \mathbf{c}$, then $\rho V_{gh} = 1$.

In a view of this, if the joint distribution of $\mathbf{x}_g$ and $\mathbf{x}_h$ is normal, independence between $\mathbf{x}_g$ and $\mathbf{x}_h$ is equivalent to asserting that the population vector correlations all vanish, i.e., $\forall g \neq h \in [k]$, $\rho V_{gh} = 0$. Thus, the summation of these

measurements overall (g, h) pairs, subject to $g < h$, serves as an effective population measure of the overall dependency among k parts of $\mathbf{x}$ and the natural criteria for testing $\mathcal{H}$ should be based on a suitable statistic for $\sum_{1\le g<h\le k}\rho V_{gh}$.

The sample counterpart of ρV_{hg} can be obtained as

$$RV_{gh} = \frac{\|\mathbf{S}_{gh}\|_F^2}{\|\mathbf{S}_{gg}\|_F\|\mathbf{S}_{hh}\|_F},$$

where the sample covariance matrix of $\mathbf{x}_\ell$ and the cross-sample covariance matrix of $\mathbf{x}_g$ and $\mathbf{x}_h$ are constructed as

$$\forall \ell \in \{g,h\},\ \mathbf{S}_{\ell\ell} = \frac{1}{n_\ell - 1}\sum_{j=1}^{n_\ell}(\mathbf{x}_{\ell j} - \widetilde{\mathbf{x}}_\ell)(\mathbf{x}_{\ell j} - \widetilde{\mathbf{x}}_\ell)^\top,\ \mathbf{S}_{gh} = \frac{1}{n_g - 1}\sum_{j=1}^{n_g}(\mathbf{x}_{gj} - \overline{\mathbf{x}}_g)(\mathbf{x}_{hj} - \overline{\mathbf{x}}_h)^\top,\ \mathbf{S}_{hg} = \mathbf{S}_{gh}^\top \tag{3}$$

with $\overline{\mathbf{x}}_\ell = n_g^{-1}\sum_{j=1}^{n_g}\mathbf{x}_{\ell j}$ and $\widetilde{\mathbf{x}}_\ell = n_\ell^{-1}\sum_{j=1}^{n_\ell}\mathbf{x}_{\ell j}$ for $\ell \in \{g,h\}$. It is important to note that the empirical measure of the vector correlation, RV_{gh}, is invariant to location, rotation, and overall scaling, and consistent for the classical case of the sample size n_1 tending to infinity and the dimension p remaining fixed. The invariance property of RV_{gh} is of special advantage because it allows to discuss the asymptotic behavior of the test statistic constructed from RV_{gh} without knowing explicit information of the population mean vector and covariance matrix. The invariance property of RV_{gh} can be also confirmed by Josse and Holmes [14].

As an estimator of the dependency measure, RV_{gh} lacks consistency in the asymptotic regime when p tends to infinity along with n_1; this fact is justified by Lemma 1 and motivates us to search for appropriate alternatives to the "naive plug-in" approach in high dimensions.

Lemma 1. *Let RV_{gh} be as already defined. Then, for any indices $g \neq h \in [k]$, the following representation holds as $p, n_1 \to \infty$:*

$$RV_{gh} = \left(\rho V_{gh} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{gg})\mathrm{tr}(\boldsymbol{\Sigma}_{hh})}{n_g\|\boldsymbol{\Sigma}_{gg}\|_F\|\boldsymbol{\Sigma}_{hh}\|_F}\right)\prod_{\ell\in\{g,h\}}\left(1 + \frac{\{\mathrm{tr}(\boldsymbol{\Sigma}_{\ell\ell})\}^2}{n_\ell\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^2}\right)^{-1/2} + o_p(1). \tag{4}$$

Proof. First, we evaluate expectation and variance of $\|\mathbf{S}_{\ell\ell}\|_F^2$:

$$\begin{aligned}
\mathrm{E}(\|\mathbf{S}_{\ell\ell}\|_F^2) =& \frac{n_\ell\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^2}{n_\ell - 1} + \frac{\{\mathrm{tr}(\boldsymbol{\Sigma}_{\ell\ell})\}^2}{n_\ell - 1},\\
\mathrm{var}\left(\|\mathbf{S}_{\ell\ell}\|_F^2\right) =& \frac{8\{\mathrm{tr}(\boldsymbol{\Sigma}_{\ell\ell})\}^2\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^2}{(n_\ell-1)^3} + \frac{4n_\ell\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^4}{(n_\ell-1)^3} + \frac{16n_\ell\mathrm{tr}(\boldsymbol{\Sigma}_{\ell\ell})\mathrm{tr}(\boldsymbol{\Sigma}_{\ell\ell}^3)}{(n_\ell-1)^3} + \frac{4(2n_\ell^2 + n_\ell + 2)\mathrm{tr}(\boldsymbol{\Sigma}_{\ell\ell}^4)}{(n_\ell-1)^3}.
\end{aligned}$$

Thus, as $p, n_1 \to \infty$,

$$\frac{\|\mathbf{S}_{\ell\ell}\|_F^2}{\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^2} = 1 + \frac{\{\mathrm{tr}(\boldsymbol{\Sigma}_{\ell\ell})\}^2}{n_\ell\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^2} + O_p(n_\ell^{-1/2}). \tag{5}$$

Next, we evaluate expectation and variance of $\|\mathbf{S}_{gh}\|_F^2$:

$$\begin{aligned}
\mathrm{E}(\|\mathbf{S}_{gh}\|_F^2) =& \frac{n_g\|\boldsymbol{\Sigma}_{gh}\|_F^2}{n_g - 1} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{gg})\mathrm{tr}(\boldsymbol{\Sigma}_{hh})}{n_g - 1},\\
\mathrm{var}\left(\|\mathbf{S}_{gh}\|_F^2\right) =& \frac{2[\{\mathrm{tr}(\boldsymbol{\Sigma}_{gg})\}^2\|\boldsymbol{\Sigma}_{hh}\|_F^2 + \{\mathrm{tr}(\boldsymbol{\Sigma}_{hh})\}^2\|\boldsymbol{\Sigma}_{gg}\|_F^2 + 2\mathrm{tr}(\boldsymbol{\Sigma}_{gg})\mathrm{tr}(\boldsymbol{\Sigma}_{hh})\|\boldsymbol{\Sigma}_{gh}\|_F^2]}{(n_g-1)^3}\\
&+ \frac{2n_g(\|\boldsymbol{\Sigma}_{gh}\|_F^4 + \|\boldsymbol{\Sigma}_{gg}\|_F^2\|\boldsymbol{\Sigma}_{hh}\|_F^2)}{(n_g-1)^3} + \frac{8n_g\mathrm{tr}(\boldsymbol{\Sigma}_{gg})\mathrm{tr}(\boldsymbol{\Sigma}_{hh}\boldsymbol{\Sigma}_{hg}\boldsymbol{\Sigma}_{gh})}{(n_g-1)^3}\\
&+ \frac{8n_g\mathrm{tr}(\boldsymbol{\Sigma}_{hh})\mathrm{tr}(\boldsymbol{\Sigma}_{gg}\boldsymbol{\Sigma}_{gh}\boldsymbol{\Sigma}_{hg})}{(n_g-1)^3} + \frac{4(n_g^2 + n_g + 2)\mathrm{tr}(\boldsymbol{\Sigma}_{gg}\boldsymbol{\Sigma}_{gh}\boldsymbol{\Sigma}_{hh}\boldsymbol{\Sigma}_{hg})}{(n_g-1)^3} + \frac{4n_g^2\|\boldsymbol{\Sigma}_{gh}\boldsymbol{\Sigma}_{hg}\|_F^2}{(n_g-1)^3}.
\end{aligned}$$

Thus, as $p, n_1 \to \infty$,

$$\frac{\|\mathbf{S}_{gh}\|_F^2}{\|\boldsymbol{\Sigma}_{gg}\|_F\|\boldsymbol{\Sigma}_{hh}\|_F} = \rho V_{gh} + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{gg})\mathrm{tr}(\boldsymbol{\Sigma}_{hh})}{n_g\|\boldsymbol{\Sigma}_{gg}\|_F\|\boldsymbol{\Sigma}_{hh}\|_F} + O_p(n_g^{-1/2}). \tag{6}$$

Combining (5) and (6), the result is established. □

To realize the essence of Lemma 1, observe that for $\boldsymbol{\Sigma}_{gg} = \boldsymbol{\Sigma}_{hh} = \mathbf{I}_p$, $\boldsymbol{\Sigma}_{gh} = \mathbf{O}$ and $n_g = n_h = o(p)$,

$$R_{gh} = \left(1 + \frac{n_g}{p}\right)^{-1} + o_p(1) = 1 + o_p(1),$$

as $p \to \infty$, indicating that the RV_{gh} coefficient is not able to detect the absence of dependence i.e., the situation where $\rho V_{gh} = 0$.

By these arguments, the crucial step in our construction of test statistic for testing (1) and (2) is to obtain an estimator of ρV_{gh} suitable for high-dimensional settings. We first consider the following unbiased estimators of $\|\boldsymbol{\Sigma}_{gh}\|_F^2$ and $\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^2$, (see Srivastava and Reid [21])

$$\forall g < h \in [k],\ \widehat{\|\boldsymbol{\Sigma}_{gh}\|_F^2} = \frac{(n_g-1)^2}{(n_g-2)(n_g+1)}\left\{\|\mathbf{S}_{gh}\|_F^2 - \frac{\mathrm{tr}(\mathbf{S}_{gg})\mathrm{tr}(\mathbf{S}_{hh})}{n_g-1}\right\}, \tag{7}$$

$$\forall \ell \in [k],\ \widehat{\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^2} = \frac{(n_\ell-1)^2}{(n_\ell-2)(n_\ell+1)}\left[\|\mathbf{S}_{\ell\ell}\|_F^2 - \frac{\{\mathrm{tr}(\mathbf{S}_{\ell\ell})\}^2}{n_\ell-1}\right], \tag{8}$$

and then define the estimator of ρV_{gh} with the high dimensionality adjustment as

$$HRV_{gh} = \frac{\widehat{\|\boldsymbol{\Sigma}_{gh}\|_F^2}}{\widehat{\|\boldsymbol{\Sigma}_{gg}\|_F}\,\widehat{\|\boldsymbol{\Sigma}_{hh}\|_F}}. \tag{9}$$

The following lemma lists invariance properties of the proposed modification.

Lemma 2. *HRV_{gh} is invariant to location, rotation, and overall scaling.*

Proof. Fix the sample sizes n_g and n_h for $g \neq h$. For $i \in [n_g]$ and $j \in [n_h]$, define random vectors $\mathbf{y}_{gi} = a\mathbf{B}\mathbf{x}_{gi} + \mathbf{c}$ and $\mathbf{y}_{hj} = a\mathbf{B}\mathbf{x}_{hj} + \mathbf{c}$, where a is any constant, $\mathbf{B}$ is a $p \times p$ matrix such that $\mathbf{B}\mathbf{B}^\top = \mathbf{I}_p$, and $\mathbf{c}$ is any p dimensional constant vector. The key idea to prove the invariance is to show that the estimator of ρV_{gh}, composed by $\mathbf{y}_{gi}$ and $\mathbf{y}_{hj}$ coincides with that one composed by $\mathbf{x}_{gi}$ and $\mathbf{x}_{hj}$ for index sets $i \in [n_g]$ and $j \in [n_h]$.

Precisely, let $\widetilde{\mathbf{S}}_{\ell\ell}$ and $\widetilde{\mathbf{S}}_{gh}$ denote the sample and cross-sample covariance matrices of $\mathbf{y}_g$ and $\mathbf{y}_h$, constructed analogously to (3). Then the following representations hold $\widetilde{\mathbf{S}}_{\ell\ell} = a^2\mathbf{B}\mathbf{S}_{\ell\ell}\mathbf{B}^\top$ for $\ell \in \{g,h\}$, $\widetilde{\mathbf{S}}_{gh} = a^2\mathbf{B}\mathbf{S}_{gh}\mathbf{B}^\top$. We further denote by $\widehat{\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^2}$ and $\widehat{\|\boldsymbol{\Sigma}_{gh}\|_F^2}$ the unbiased estimators of the covariance and cross-covariance matrices constructed by $\mathbf{y}_g$ and $\mathbf{y}_h$ in the same way as (7) and (8), and notice that

$$\forall \ell \in \{g,h\},\ \widetilde{\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^2} = a^2\widehat{\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^2},\ \widetilde{\|\boldsymbol{\Sigma}_{gh}\|_F^2} = a^2\widehat{\|\boldsymbol{\Sigma}_{gh}\|_F^2}. \tag{10}$$

Let $\widetilde{HRV}_{gh}$ denote the estimator of ρV_{gh} constructed by $\widetilde{\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^2}$ and $\widetilde{\|\boldsymbol{\Sigma}_{gh}\|_F^2}$ similarly to (9). Then from (10), we obtain

$$\widetilde{HRV}_{gh} = \frac{a^2\widehat{\|\boldsymbol{\Sigma}_{gh}\|_F^2}}{(a\widehat{\|\boldsymbol{\Sigma}_{gg}\|_F})(a\widehat{\|\boldsymbol{\Sigma}_{hh}\|_F})} = HRV_{gh}$$

which finishes the proof. □

Now, to proceed further with the test statistic construction, we need one more result. The following theorem shows consistency of HRV_{gh} in the high-dimensional regime.

Theorem 1. *Let HRV_{gh} be as already defined. Then, as $p, n_1 \to \infty$, for any indices $g \neq h \in [k]$, it holds that $HRV_{gh} = \rho V_{gh} + o_p(1)$.*

Proof. First, we evaluate the variance of $\widehat{\|\boldsymbol{\Sigma}_{gh}\|_F^2}$, for which we obtain the following representation

$$\mathrm{var}\left(\widehat{\|\boldsymbol{\Sigma}_{gh}\|_F^2}\right) = \frac{2(\|\boldsymbol{\Sigma}_{gh}\|_F^4 + \|\boldsymbol{\Sigma}_{gg}\|_F^2\|\boldsymbol{\Sigma}_{hh}\|_F^2)}{(n_g-2)(n_g+1)} + \frac{4(n_g^2-5)\mathrm{tr}(\boldsymbol{\Sigma}_{gg}\boldsymbol{\Sigma}_{gh}\boldsymbol{\Sigma}_{hh}\boldsymbol{\Sigma}_{hg})}{(n_g-2)(n_g-1)(n_g+1)} + \frac{4\mathrm{tr}\{(\boldsymbol{\Sigma}_{gh}\boldsymbol{\Sigma}_{hg})^2\}}{n_g-1}. \tag{11}$$

From $\mathrm{tr}(\boldsymbol{\Sigma}_{gg}\boldsymbol{\Sigma}_{gh}\boldsymbol{\Sigma}_{hh}\boldsymbol{\Sigma}_{hg}) \leq \|\boldsymbol{\Sigma}_{gg}\|_F\|\boldsymbol{\Sigma}_{hh}\|_F\|\boldsymbol{\Sigma}_{gh}\|_F^2$, $\mathrm{tr}\{(\boldsymbol{\Sigma}_{gh}\boldsymbol{\Sigma}_{hg})^2\} \leq \|\boldsymbol{\Sigma}_{gh}\|_F^4$ and $\rho V_{gh} < 1$, we get

$$\frac{\mathrm{var}(\widehat{\|\boldsymbol{\Sigma}_{gh}\|_F^2})}{(\|\boldsymbol{\Sigma}_{gg}\|_F\|\boldsymbol{\Sigma}_{hh}\|_F)^2} = O\left(n_g^{-1}\right).$$

Using Chebyshev's inequality, we obtain

$$\frac{\widehat{\|\boldsymbol{\Sigma}_{gh}\|_F^2}}{\|\boldsymbol{\Sigma}_{gg}\|_F\|\boldsymbol{\Sigma}_{hh}\|_F} = \rho V_{gh} + o_p(1). \tag{12}$$

Next, we evaluate the variance of $\widehat{\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^2}$ for $\ell \in \{g,h\}$ for which we obtain

$$\mathrm{var}\left(\widehat{\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^2}\right) = \frac{4\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^4}{(n_\ell-2)(n_\ell+1)} + \frac{4(2n_\ell^2-n_\ell-7)\|\boldsymbol{\Sigma}_{\ell\ell}^2\|_F^2}{(n_\ell-2)(n_\ell-1)(n_\ell+1)}.$$

From $\|\boldsymbol{\Sigma}_{\ell\ell}^2\|_F^2 \leq \|\boldsymbol{\Sigma}_{\ell\ell}\|_F^4$, we get $\mathrm{var}(\widehat{\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^2})/\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^4 = O\left(n_\ell^{-1}\right)$. Thus, $\mathrm{var}\left(\widehat{\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^2}\right)/\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^4 = o(1)$ is established under p fixed $n_1 \to \infty$ or $p, n_1 \to \infty$. Using Chebyshev's inequality, we obtain

$$\frac{\widehat{\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^2}}{\|\boldsymbol{\Sigma}_{\ell\ell}\|_F^2} = 1 + o_p(1). \tag{13}$$

Combining (12) and (13) gives the claim. □

Remark 1. Theorem 1 remains valid in the large sample asymptotic setting, i.e. with p fixed and $n_1 \to \infty$.

2.2. The proposed test statistic and its asymptotic properties

In order to construct the test statistic, we first observe that the tests (1) and (2) can be restated in terms of ρV coefficient as

$$\mathcal{H} : \forall g \neq h \in [k] \quad \rho V_{gh} = 0 \quad \text{vs.} \quad \mathcal{A} : \rho V_{gh} > 0 \text{ for some } g \text{ and } h. \tag{14}$$

Further, with the high-dimensional adjustment HRV_{gh} at hand, we propose our test statistic for (1), (2) and (14), namely the vector correlation type statistics,

$$T = \sum_{1 \leq g < h \leq k} HRV_{gh},$$

which consistently estimates the population measure of the overall dependency, $\sum_{1 \leq g < h \leq k} \rho V_{gh}$ in the joint distribution of $\mathbf{x}_1, \ldots, \mathbf{x}_k$, and sums all pairwise sample correlations for a "one-sided" test. Note that under the null hypothesis, all of the population's ρV_{gh} should be zero corresponding to zero off-diagonal blocks $\boldsymbol{\Sigma}_{gh}$. Hence, as an immediate consequence of Theorem 1, the asymptotic behavior of proposed test statistic under the null $\mathcal{H}$ and alternative $\mathcal{A}$ when $n_1, p \to \infty$ is as follows

$$T = \begin{cases} o_p(1) & \text{under } \mathcal{H}, \\ \sum_{1 \leq g < h \leq k} \rho V_{gh} + o_p(1) & \text{under } \mathcal{A}, \end{cases}$$

i.e., the large values of T indicate departures from $\mathcal{H}$.

To state the size-α test of significance using T, we need to characterize its null asymptotic distribution. For this purpose, the following property will be assumed for our null asymptotic results.

Viewing both $\|\boldsymbol{\Sigma}_{gg}\|_F$ and $\|\boldsymbol{\Sigma}_{gg}^2\|_F$ as functions of p and using p as a driving asymptotic index, we assume that $\forall g \in [k]$,

(A1) $\|\boldsymbol{\Sigma}_{gg}^2\|_F^2 / \|\boldsymbol{\Sigma}_{gg}\|_F^4 = o(1)$ *as* $p \to \infty$.

To exemplify a covariance model family satisfying (A1), we consider a matrix $\boldsymbol{\Sigma}_{gg}$ which satisfy (A1) such that its rth largest eigen value of $\boldsymbol{\Sigma}_{gg}$ admits the representation

$$\lambda_r(\boldsymbol{\Sigma}_{gg}) = \begin{cases} \eta_{r(g)} p^{\theta_{r(g)}}, & r \in \{1, \ldots, t_g\}, \\ \eta_{r(g)}, & r \in \{t_g + 1, \ldots, p\}, \end{cases} \tag{15}$$

where $\eta_{r(g)} > 0$ and $\theta_{r(g)} > 0$ are fixed constants and t_g is a fixed positive integer. Now if $\theta_{1(g)} < 1/2$ the covariance structure of $\boldsymbol{\Sigma}_{gg}$ is called a non-strongly spiked eigenvalue model (NSSE-model). Accordingly, when these latter do not hold the structure is called a strongly spiked eigenvalue (SSE-model).

In what follows, we show that NSSE-model (15) assumed for $\boldsymbol{\Sigma}_{gg}$ is a sufficient condition for (A1). Indeed, assuming that $\forall g \in [k]$, $\boldsymbol{\Sigma}_g$ satisfies (15), we obtain

$$\mathrm{tr}\{(\boldsymbol{\Sigma}_{gg}\boldsymbol{\Sigma}_{hh})^2\} \leq \eta_{\max}^4 p^{4\theta_{\max}}, \ \mathrm{tr}(\boldsymbol{\Sigma}_{gg}\boldsymbol{\Sigma}_{hh}) \geq \eta_{\min}^2 p,$$

where $\theta_{\max} = \max\{\theta_{1(1)}, \ldots, \theta_{1(k)}\}$, $\eta_{\max} = \max\{\eta_{1(1)}, \ldots, \eta_{p(k)}\}$, $\eta_{\min} = \min\{\eta_{1(1)}, \ldots, \eta_{p(k)}\}$. Since $\theta_{\max} < 1/2$, the NSSE-model (15) for $\boldsymbol{\Sigma}_{gg}$ is the sufficient condition for (A1) to be fulfilled.

Now we are ready to state one of the main results of our study, the central limit theorem for T under $\mathcal{H}$.

Theorem 2. *Suppose that the null hypothesis* $\mathcal{H}$ *from* (14) *is true. Suppose further that (A1) is satisfied for all* $g \in [k]$ *and consider the asymptotic regime* $p, n_1 \to \infty$. *Then, after suitable rescaling,* T *is asymptotically normal, namely,* $\sigma^{-1}T \rightsquigarrow \mathcal{N}(0, 1)$ *with* $\sigma^2 = 2\sum_{g=1}^{k-1}(k-g)n_g^{-2}$.

Proof. From (13), under $p, n_1 \to \infty$, $\widehat{\|\boldsymbol{\Sigma}_{\ell\ell}\|_F} = \|\boldsymbol{\Sigma}_{\ell\ell}\|_F\{1 + o_p(1)\}$ for $\ell \in \{g, h\}$. Thus $T = \widetilde{T} + o_p(1)$, where $\widetilde{T} = \sum_{1 \leq g < h \leq k} \widehat{\|\boldsymbol{\Sigma}_{gh}\|_F^2} / (\|\boldsymbol{\Sigma}_{gg}\|_F \|\boldsymbol{\Sigma}_{hh}\|_F)$. Therefore, it is sufficient to show the asymptotic normality of $\widetilde{T}$. From (11), under $\mathcal{H}$, the asymptotic variance of $\widetilde{T}$ is obtained as $\sigma^2 = 2\sum_{g=1}^{k-1}(k-g)n_g^{-2}$.

Under $\mathcal{H}$, we have $\forall g \in [k],\ j \in [n_k]\ \mathbf{x}_{gj} = \boldsymbol{\Sigma}_{gg}^{1/2}\mathbf{z}_{gj} + \boldsymbol{\mu}_g$, where $\mathbf{z}_{gj} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$ and $\mathbf{z}_{gj}$ are mutually independent. We define $p \times \ell$ matrix, $\mathbf{Z}_{g(\ell)} = (\mathbf{z}_{g1}, \ldots, \mathbf{z}_{g\ell})$ for $\ell \leq n_g$. Note that each component of $\mathbf{Z}_{g(n_g)}$ independently $\mathcal{N}(0, 1)$ distributed. For $g < h$, i.e., $n_g \leq n_h$, under $\mathcal{H}$,

$$\widehat{\|\boldsymbol{\Sigma}_{gh}\|_F^2} = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{gg}\mathbf{Z}_{g(n_g)}\mathbf{Z}_{h(n_g)}^\top\boldsymbol{\Sigma}_{hh}\mathbf{Z}_{h(n_g)}\mathbf{Z}_{g(n_g)}^\top)}{n_g^2} - \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{gg}\mathbf{Z}_{g(n_g)}\mathbf{Z}_{g(n_g)}^\top)\mathrm{tr}(\boldsymbol{\Sigma}_{hh}\mathbf{Z}_{h(n_g)}\mathbf{Z}_{h(n_g)}^\top)}{n_g^3} + o_p\left(\frac{\|\boldsymbol{\Sigma}_{gg}\|_F\|\boldsymbol{\Sigma}_{hh}\|_F}{n_g}\right).$$

Let $\boldsymbol{\Gamma}_g$ be orthogonal matrix s.t. $\boldsymbol{\Gamma}_g^\top\boldsymbol{\Sigma}_{gg}\boldsymbol{\Gamma}_g = \mathrm{diag}(\lambda_{g1}, \ldots, \lambda_{gp})$. For $i \in [p]$, we set $\mathbf{u}_{gi} = (\mathbf{e}_i^\top\boldsymbol{\Gamma}_g^\top\mathbf{z}_{g1}, \ldots, \mathbf{e}_i^\top\boldsymbol{\Gamma}_g^\top\mathbf{z}_{gn_g})^\top$, where $\mathbf{e}_i$ denotes an ith unit vector which has all the elements equal to 0 except for one, namely its ith element which is equal to 1. Then $\mathbf{u}_{gi} \sim \mathcal{N}_{n_g}(\mathbf{0}, \mathbf{I}_{n_g})$ and $\mathbf{e}_j^\top\mathbf{u}_{gi} = \mathbf{z}_{gj}^\top\boldsymbol{\Gamma}_g\mathbf{e}_i$ are mutually independent whenever (g, i, j) are distinct indices. Let $\mathbf{u}_{gi(\ell)} = (\mathbf{e}_i^\top\boldsymbol{\Gamma}_g^\top\mathbf{z}_{g1}, \ldots, \mathbf{e}_i^\top\boldsymbol{\Gamma}_g^\top\mathbf{z}_{g\ell})^\top$. Then

$$\boldsymbol{\Gamma}_{gg}^\top\mathbf{Z}_{g(n_g)} = (\mathbf{u}_{g1(n_g)}, \ldots, \mathbf{u}_{gp(n_g)})^\top,\ \ \boldsymbol{\Gamma}_{hh}^\top\mathbf{Z}_{h(n_g)} = (\mathbf{u}_{h1(n_g)}, \ldots, \mathbf{u}_{hp(n_g)})^\top.$$

Using these variables, we rewrite

$$\widehat{\|\boldsymbol{\Sigma}_{gh}\|_F^2} = n_g^{-2}\sum_{i=1}^p\sum_{j=1}^p\lambda_{gi}\lambda_{hj}(\mathbf{u}_{gi(n_g)}^\top\mathbf{u}_{hj(n_g)})^2 - n_g^{-3}\sum_{i=1}^p\sum_{j=1}^p\lambda_{gi}\lambda_{hj}\mathbf{u}_{gi(n_g)}^\top\mathbf{u}_{gj(n_g)}\mathbf{u}_{hi(n_g)}^\top\mathbf{u}_{hj(n_g)} + o_p\left(\frac{\|\boldsymbol{\Sigma}_{gg}\|_F\|\boldsymbol{\Sigma}_{hh}\|_F}{n_g}\right),$$

yielding the representation $\widetilde{T}/\sigma = \sum_{i=1}^p\varepsilon_i + o_p(1)$, where

$$\begin{aligned}\varepsilon_i = \sum_{1\leq g<h\leq k}\frac{1}{\sigma n_g^2}\Bigg[&\frac{\lambda_{gi}\lambda_{hi}}{\|\boldsymbol{\Sigma}_{gg}\|_F\|\boldsymbol{\Sigma}_{hh}\|_F}\left\{(\mathbf{u}_{gi(n_g)}^\top\mathbf{u}_{hi(n_g)})^2 - \frac{\|\mathbf{u}_{gi(n_g)}\|^2\|\mathbf{u}_{hi(n_g)}\|^2}{n_g}\right\}\\ &+\sum_{j=0}^{i-1}\frac{\lambda_{gi}\lambda_{hj}}{\|\boldsymbol{\Sigma}_{gg}\|_F\|\boldsymbol{\Sigma}_{hh}\|_F}\left\{(\mathbf{u}_{gi(n_g)}^\top\mathbf{u}_{hj(n_g)})^2 - \frac{\|\mathbf{u}_{gi(n_g)}\|^2\|\mathbf{u}_{hj(n_g)}\|^2}{n_g}\right\}\\ &+\sum_{j=0}^{i-1}\frac{\lambda_{gj}\lambda_{hi}}{\|\boldsymbol{\Sigma}_{gg}\|_F\|\boldsymbol{\Sigma}_{hh}\|_F}\left\{(\mathbf{u}_{gj(n_g)}^\top\mathbf{u}_{hi(n_g)})^2 - \frac{\|\mathbf{u}_{gj(n_g)}\|^2\|\mathbf{u}_{hi(n_g)}\|^2}{n_g}\right\}\Bigg].\end{aligned}$$

Observe that for $i = 1$, both the second and third terms can be ignored being equal to zero. Define further $\mathcal{F}_0 = \{\emptyset, \Omega\}$, and let $\mathcal{F}_i$ for any natural number i be the σ-algebra generated by the random variables U_{i-1}, where

$$U_{i-1} = (\mathbf{u}_{11}, \ldots, \mathbf{u}_{1i-1}, \ldots, \mathbf{u}_{k1}, \ldots, \mathbf{u}_{ki-1}).$$

Then we find that $\mathcal{F}_0 \subseteq \cdots \subseteq \mathcal{F}_\infty$ and $\mathrm{E}(\varepsilon_i|\mathcal{F}_{i-1}) = 0$. Thus, $\{\varepsilon_i\}_{i=1}^\infty$ is a martingale difference sequence. We show the asymptotic normality of $\varepsilon_1 + \cdots + \varepsilon_p$ by adapting the martingale difference central limit theorem; see, e.g., Shiryaev [20]. Let $E_{i-1} = \mathrm{E}(\varepsilon_i^2|\mathcal{F}_{i-1})$. Then

$$\mathrm{E}\left(\sum_{i=1}^p E_{i-1}\right) = 1 + o(1),\ \ \mathrm{var}\left(\sum_{i=1}^p E_{i-1}\right) = O(n_1^{-1}).$$

Thus, $(\mathrm{I}) : \sum_{i=1}^p E_{i-1} = 1 + o_p(1)$ as $p, n_1 \to \infty$. Also

$$\sum_{i=1}^p\mathrm{E}(\varepsilon_i^4) = O\left(\sum_{g=1}^k\frac{\|\boldsymbol{\Sigma}_{gg}^2\|^2}{\|\boldsymbol{\Sigma}_{gg}\|^4}\right).$$

Thus, under (A1), $(\mathrm{II}) : \sum_{i=1}^p\mathrm{E}(\varepsilon_i^4) = o(1)$ as $p, n_1 \to \infty$. The above results (I) and (II) complete the proof. Note that (I) holds but (II) does not hold under p fixed $n_1 \to \infty$ since $\sum_{i=1}^p\mathrm{E}(\varepsilon_i^4) = O(1)$. □

Remark 2. Under $\mathcal{H}$ and (A1), $\mathrm{var}(T)/\sigma^2 = 1 + o(1)$ as $p, n_1 \to \infty$.

By Theorem 2, a critical value for the approximate size-α test can be calibrated based on the normal quantiles.

An alternative idea of how to express the test and show that we can control size is as follows. Using Theorem 2, our proposed approximate size-α test of the null hypothesis $\mathcal{H}$ can be based on the statistic

$$Q_\alpha = \mathbf{1}\left[\sum_{1\leq g<h\leq k}HRV_{gh}/\sigma \geq z_{1-\alpha}\right],$$

where $\mathbf{1}(\cdot)$ represents the indicator function and $z_\alpha = \Phi^{-1}(\alpha)$ denotes the upper α quantile of $\mathcal{N}(0, 1)$. The following corollary states that the test Q_α can efficiently control the size.

Corollary 1. *Suppose that the condition (A1) holds, then, as* $p, n_1 \to \infty$,

$$\Pr(Q_\alpha = 1 |\ \mathcal{H}) = \alpha + o(1).$$

We further evaluate the power of T under a kind of local alternative. Consider the alternative hypothesis

$$\mathcal{A}:\ \mathbf{x}_g \text{ and } \mathbf{x}_h \text{ are dependent for some } g, h \in [k]$$

satisfying condition (16) below. Draw n_i samples from such alternatives $\mathbf{x}_i$ following sampling scheme of Section 1 to form the respective analogues of HRV_{gh} and Q_α and denote them by $HRV_{gh}^{\mathcal{A}}$ and $Q_\alpha^{\mathcal{A}}$, respectively.

Theorem 3. *In addition to the assumptions in Theorem 2 let*

$$\Theta_{n_1} = \{\boldsymbol{\Sigma}_{\langle 1\rangle} :\ \max_{1\le g<h\le k} \rho V_{gh} \ge n_1^{-\delta}\} \tag{16}$$

be a set of alternatives $\boldsymbol{\Sigma}_{\langle 1\rangle}$ *such that* $\max_{1\le g<h\le k} \rho V_{gh} \ge n_1^{-\delta}$, *where* $0 < \delta < 1$. *Then, as* $p, n_1 \to \infty$,

$$\inf_{\Theta_{n_1}} \Pr(Q_\alpha^{\mathcal{A}} = 1 \,|\, \mathcal{A}) = 1 + o(1).$$

Proof. From (13), the power of our proposed test at $\boldsymbol{\Sigma}_{\langle 1\rangle}$ is $\Pr(Q_\alpha^{\mathcal{A}} = 1 \,|\, \mathcal{A}) = \Pr(\widetilde{T} \ge \sigma z_\alpha) + o(1)$ as p fixed $n_1 \to \infty$ or $p, n_1 \to \infty$. Thus it is sufficient to show that $\Pr(\widetilde{T} \ge \sigma z_\alpha) = 1 + o(1)$ for any $\boldsymbol{\Sigma}_{\langle 1\rangle} \in \Theta_{n_1}$.

We note that $\mathrm{E}(\widetilde{T}) = \sum_{1\le g<h\le k} \rho V_{gh} > 0$ for any $\boldsymbol{\Sigma}_{\langle 1\rangle} \in \Theta_{n_1}$, and

$$\Pr\left(\widetilde{T} \ge \sigma z_\alpha\right) \ge 1 - \Pr\left(|\widetilde{T} - \mathrm{E}(\widetilde{T}) - \sigma z_\alpha| \ge \mathrm{E}(\widetilde{T})\right).$$

Using Markov's inequality and Cauchy–Schwarz inequality in the context of expectation, we obtain

$$\Pr\left(\left|\widetilde{T} - \mathrm{E}(\widetilde{T}) - \sigma z_\alpha\right| \ge \mathrm{E}(\widetilde{T})\right) \le \mathrm{E}\left(\left|\widetilde{T} - \mathrm{E}(\widetilde{T}) - \sigma z_\alpha\right| / \mathrm{E}(\widetilde{T})\right) \le \mathrm{E}\left(\left|\widetilde{T} - \mathrm{E}(\widetilde{T}) - \sigma z_\alpha\right|^2\right) / \{\mathrm{E}(\widetilde{T})\}^2.$$

Since $\mathrm{E}(|\widetilde{T} - \mathrm{E}(\widetilde{T}) - \sigma z_\alpha|^2) = \mathrm{var}(\widetilde{T}) + \sigma^2 z_\alpha^2$, we obtain

$$\Pr(\widetilde{T} \ge \sigma z_\alpha) \ge 1 - \frac{\mathrm{var}(\widetilde{T}) + \sigma^2 z_\alpha^2}{\{\mathrm{E}(\widetilde{T})\}^2}. \tag{17}$$

We further evaluate $\mathrm{var}(\widetilde{T})$. For any $g < h, g, h \in [k]$, we define $A_{gh} = \widehat{HRV}_{gh} - \rho V_{gh}$, where $\widehat{HRV}_{gh} = \widehat{\|\boldsymbol{\Sigma}_{gh}\|_F^2} / (\|\boldsymbol{\Sigma}_{gg}\|_F \|\boldsymbol{\Sigma}_{hh}\|_F)$. Then

$$\frac{\mathrm{var}(\widetilde{T})}{\{\mathrm{E}(\widetilde{T})\}^2} = \frac{\mathrm{E}\{(\sum_{1\le g<h\le k} A_{gh})^2\}}{\{\mathrm{E}(\widetilde{T})\}^2} \le \frac{k(k-1)\sum_{1\le g<h\le k} \mathrm{E}(A_{gh}^2)}{2\{\mathrm{E}(\widetilde{T})\}^2}.$$

$\mathrm{E}(A_{gh}^2)$ is obtained by

$$\mathrm{E}(A_{gh}^2) = \frac{2(\rho V_{gh}^2 + 1)}{(n_g - 2)(n_g + 1)} + \frac{4(n_g^2 - 5)\mathrm{tr}(\boldsymbol{\Sigma}_{gg}\boldsymbol{\Sigma}_{gh}\boldsymbol{\Sigma}_{hh}\boldsymbol{\Sigma}_{hg})}{(n_g - 2)(n_g - 1)(n_g + 1)\|\boldsymbol{\Sigma}_{gg}\|_F^2 \|\boldsymbol{\Sigma}_{hh}\|_F^2} + \frac{4\mathrm{tr}\{(\boldsymbol{\Sigma}_{gh}\boldsymbol{\Sigma}_{hg})^2\}}{(n_g - 1)\|\boldsymbol{\Sigma}_{gg}\|_F^2 \|\boldsymbol{\Sigma}_{hh}\|_F^2}.$$

From $\mathrm{tr}(\boldsymbol{\Sigma}_{gg}\boldsymbol{\Sigma}_{gh}\boldsymbol{\Sigma}_{hh}\boldsymbol{\Sigma}_{hg}) \le \|\boldsymbol{\Sigma}_{gg}\|_F \|\boldsymbol{\Sigma}_{hh}\|_F \|\boldsymbol{\Sigma}_{gh}\|_F^2$ and $\mathrm{tr}\{(\boldsymbol{\Sigma}_{gh}\boldsymbol{\Sigma}_{hg})^2\} \le \|\boldsymbol{\Sigma}_{gh}\|_F^4$, we get

$$\frac{\mathrm{E}(A_{gh}^2)}{\{\mathrm{E}(\widetilde{T})\}^2} = O\left(\frac{\rho V_{gh} + \rho V_{gh}^2}{(\sum_{1\le g<h\le k} \rho V_{gh})^2 n_g} + \frac{\rho V_{gh}^2 + 1}{(\sum_{1\le g<h\le k} \rho V_{gh})^2 n_g^2}\right).$$

Note that $\mathrm{E}(\widetilde{T}) = \sum_{1\le g<h\le k} \rho V_{gh} \ge \max_{1\le g<h\le k} \rho V_{gh} \ge n_1^{-\delta}$. Thus, for any $\boldsymbol{\Sigma}_{\langle 1\rangle} \in \Theta_{n_1}$,

$$\frac{\mathrm{E}(A_{gh}^2)}{\{\mathrm{E}(\widetilde{T})\}^2} = O\left(\frac{1}{n_1^{1-\delta}} + \frac{1}{n_1} + \frac{1}{n_1^2}\right).$$

Since k is fixed, we obtain

$$\frac{\mathrm{var}(\widetilde{T})}{\{\mathrm{E}(\widetilde{T})\}^2} = O\left(\frac{1}{n_1^{1-\delta}} + \frac{1}{n_1} + \frac{1}{n_1^2}\right). \tag{18}$$

Next, we evaluate $\sigma^2 / \{\mathrm{E}(\widetilde{T})\}^2$. Since $\sigma^2 = O(n_1^{-2})$ and $\{\mathrm{E}(\widetilde{T})\}^2 \ge (\max_{1\le g<h\le k} \rho V_{gh})^2 \ge n_1^{-2\delta}$, we obtain

$$\frac{\sigma^2 z_\alpha^2}{\{\mathrm{E}(\widetilde{T})\}^2} = O\left(\frac{1}{n_1^{2(1-\delta)}}\right). \tag{19}$$

Substituting (18) and (19) to (17), for any $\boldsymbol{\Sigma}_{\langle 1\rangle} \in \Theta_{n_1}$,

$$\Pr(\widetilde{T} \geq \sigma z_\alpha) = 1 + O\left(\frac{1}{n_1^{1-\delta}} + \frac{1}{n_1^{2(1-\delta)}} + \frac{1}{n_1} + \frac{1}{n_1^2}\right).$$

Therefore, under p fixed $n_1 \to \infty$ or $p, n_1 \to \infty$, $\inf_{\Theta_{n_1}} \Pr(\widetilde{T} \geq \sigma z_\alpha) = 1 + o(1)$. □

Remark 3. Theorem 3 remains valid with p fixed and $n_1 \to \infty$.

3. Stepwise multiple significance test

In this section, we proceed to explore the proposed statistic T and construct the new step-down multiple comparison significance test for simultaneous testing of independence.

Let $\mathcal{M}_q$ be the family of subsets with cardinal number $q \geq 2$ of the set $[k]$. Also, let these subsets be denoted by $m = \{\ell_1, \ldots, \ell_q\} \in \mathcal{M}_q$ where $\ell_1 < \cdots < \ell_q$ and let $\boldsymbol{\Sigma}^{(q,m)}$ be the following $pq \times pq$ matrix for these m:

$$\boldsymbol{\Sigma}^{(q,m)} = \begin{pmatrix} \boldsymbol{\Sigma}_{\ell_1\ell_1} & \cdots & \boldsymbol{\Sigma}_{\ell_1\ell_q} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{\ell_q\ell_1} & \cdots & \boldsymbol{\Sigma}_{\ell_q\ell_q} \end{pmatrix}.$$

We wish to test the following hypothesis:

$$\mathcal{H}^{\{q,m\}}: \quad \forall g \neq h \in \{\ell_1, \ldots, \ell_q\}, \quad \boldsymbol{\Sigma}_{gh} = \mathbf{O} \quad \text{vs.} \quad \mathcal{A}^{\{q,m\}}: \quad \text{not } \mathcal{H}^{\{q,m\}},$$

and for this, we obtain the test statistic $T^{\{q,m\}}/\sigma_{\{q,m\}}$ based on results of Section 2, where $T^{\{q,m\}} = \sum_{1\leq g<h\leq \ell_q} HRV_{gh}$ and $g \neq h \in m$. Here, we consider the problem of testing family of hypotheses $\mathcal{J} = \{\mathcal{H}^{\{2,m\}} : \boldsymbol{\Sigma}_{\ell_1\ell_2} = \mathbf{O}, m \in \mathcal{M}_2\}$.

Let $\mathcal{G}_q$ be the set consisting of all hypothesis $\mathcal{H}^{\{q,m\}}$ and let $\mathcal{G} = \cup_{q=2}^k \mathcal{G}_q$. Then the family $\mathcal{G}$ is closed. Hence, we can derive a step-down multiple comparison procedure based on closed testing procedure for $\mathcal{G}$. We define

$$\alpha_q = \begin{cases} 1 - (1-\alpha)^{q/k}, & \text{for } q \in \{2, \ldots, k-2\}, \\ \alpha, & \text{for } q \in \{k-1, k\}, \end{cases}$$

and let $t_{\{q,m\}}(\alpha)$ be the upper α percentiles of the statistic $T^{\{q,m\}}$ under $\mathcal{H}^{\{q,m\}}$, that is, $t_{\{q,m\}}(\alpha)$ satisfies $\Pr\{T^{\{q,m\}} \geq t_{\{q,m\}}(\alpha)\} = \alpha$. Then we carry out the following Tukey–Welsch type step-down multiple test for all hypotheses in $\mathcal{G}$ by using the $T^{\{q,m\}}$:

Step 1. We test the hypothesis $\mathcal{H}^{\{k,m\}} = \mathcal{H}$.

(C1) If $T \geq t(\alpha_k)$, we reject $\mathcal{H}$ and go to Step 2.
(C2) If $T < t(\alpha_k)$, we retain all hypotheses in $\mathcal{G}$ and stop the test. Here, $t(\alpha_k)$ satisfies $\Pr\{T \geq t(\alpha_k)\} = \alpha_k$ under $\mathcal{H}$.

Step 2. We test all hypotheses $\mathcal{H}^{\{k-1,m\}}$ in $\mathcal{G}_{k-1}$.

(C1) If $T^{\{k-1,m\}} \geq t_{\{k-1,m\}}(\alpha_{k-1})$, we reject $\mathcal{H}^{\{k-1,m\}}$ and go to Step 3.
(C2) If $T^{\{k-1,m\}} < t_{\{k-1,m\}}(\alpha_{k-1})$, we retain $\mathcal{H}^{\{k-1,m\}}$ and all hypotheses in $\cup_{q=2}^{k-2}\mathcal{G}_q$ implied by $\mathcal{H}^{\{k-1,m\}}$. We finish the test.

Step 3. We test all hypotheses $\mathcal{H}^{\{k-2,m\}}$ in $\mathcal{G}_{k-2}$ which are not retained in Step 2.

(C1) If $T^{\{k-2,m\}} \geq t_{\{k-2,m\}}(\alpha_{k-2})$, we reject $\mathcal{H}^{\{k-2,m\}}$ and go to Step 4.
(C2) If $T^{\{k-2,m\}} < t_{\{k-2,m\}}(\alpha_{k-2})$, we retain $\mathcal{H}^{\{k-2,m\}}$ and all hypotheses in $\cup_{q=2}^{k-3}\mathcal{G}_q$ implied by $\mathcal{H}^{\{k-2,m\}}$. We finish the test.

We repeat similar judgments till Step $k-1$ at the maximum.

Remark 4. From a principle of closed testing procedure, we note that the maximum type-I FWE (family-wise error rate) of our proposed step-down multiple comparison procedure is not greater than α.

By the results of Theorem 2, the critical values $t_{\{q,m\}}(\alpha)$ for an approximate α-size test can be set as $\sigma_{\{q,m\}}z_{1-\alpha}$, where $\sigma_{\{q,m\}}z_{1-\alpha}$ satisfies $\Pr\{T^{\{q,m\}} \geq \sigma_{\{q,m\}}z_{1-\alpha}\} = \alpha + o(1)$ under $\mathcal{H}^{\{q,m\}}$ and assuming that (A1) holds.

A more detailed presentation of Tukey–Welsch step-down multiple testing procedure can be found in work of Hochberg and Tamhane [11]. Our suggested procedure will be demonstrated for $k = 4$ with real data analysis provided in Section 4.2. We also prepare Fig. 3 to illustrate the procedure in the step-by-step fashion.

Table 1
The empirical size of proposed test $\hat{\alpha}_T$ is listed in the column $\rho = 0.0$ for a variety of combinations of p and $\boldsymbol{n} = (n_1, n_2, n_3, n_4, n_5)^\top$. The data underlying the table are i.i.d. $p \times k$-variate normal with $k = 5$ and the covariance matrix having the following within-block structures $\boldsymbol{\Sigma}_{(1)} = \mathrm{diag}(\boldsymbol{\Sigma}_{11}, \ldots, \boldsymbol{\Sigma}_{55})$, where each $\boldsymbol{\Sigma}_{gg}$ has an $AR(1)$ structure, i.e., $\boldsymbol{\Sigma}_{gg} = g(0.5^{|i-j|})$. For each combination of p and $\boldsymbol{n}$, empirical size of the test is calculated from $\ell = 100,000$ independently generated data sets. The empirical power of proposed test $\hat{\beta}_T$ is listed in the columns $\rho = 0.4$ and 0.6 for a variety of combinations of p and $\boldsymbol{n} = (n_1, n_2, n_3, n_4, n_5)^\top$. The data underlying the table are i.i.d. $p \times k$-variate normal with $k = 5$ and the covariance matrix having $\mathcal{A}_1$ and $\mathcal{A}_2$. For each combination of p and $\boldsymbol{n}$, empirical power of proposed test is calculated from $\ell = 100,000$ independently generated data sets.

$\boldsymbol{n}^\top$	p	Size	Power ($\mathcal{A}_1$)		Power ($\mathcal{A}_2$)	
		$\rho = 0.0$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.4$	$\rho = 0.6$
(20,20,20,20,20)	50	0.060	0.159	0.361	0.333	0.783
	100	0.059	0.160	0.365	0.337	0.796
	200	0.060	0.156	0.367	0.336	0.801
	300	0.060	0.157	0.366	0.335	0.803
(10,15,20,25,30)	50	0.064	0.122	0.228	0.215	0.507
	100	0.064	0.124	0.231	0.215	0.518
	200	0.063	0.121	0.229	0.211	0.520
	300	0.063	0.122	0.229	0.215	0.519
(40,40,40,40,40)	50	0.055	0.313	0.788	0.728	0.998
	100	0.055	0.317	0.796	0.742	0.999
	200	0.055	0.312	0.798	0.745	0.999
	300	0.055	0.313	0.799	0.759	1.000
(20,30,40,50,60)	50	0.058	0.198	0.495	0.450	0.928
	100	0.056	0.196	0.496	0.455	0.937
	200	0.057	0.195	0.500	0.459	0.943
	300	0.055	0.194	0.498	0.456	0.943
(60,60,60,60,60)	50	0.053	0.522	0.976	0.947	1.000
	100	0.053	0.521	0.978	0.956	1.000
	200	0.054	0.524	0.979	0.961	1.000
	300	0.052	0.522	0.979	0.966	1.000
(40,50,60,70,80)	50	0.056	0.388	0.891	0.840	1.000
	100	0.053	0.387	0.895	0.853	1.000
	200	0.052	0.388	0.900	0.860	1.000
	300	0.053	0.390	0.901	0.862	1.000

4. Numerical study

We present results from numerical studies which are designed to evaluate the performance of the proposed statistic T for testing independence hypothesis $\mathcal{H}$ and for the multiple comparison procedure based on the simultaneous test of independence. Our simulations explore the size of the tests when critical values are selected using asymptotic normality of T and compare their power for a number of alternative scenarios. We also employ the proposed test to analyze data from Electroencephalograph (EEG) experiment to illustrate the application of our results.

4.1. Simulation experiments

We first assess the accuracy of the proposed test statistic for its size control. The test compares a rescaled statistic T to the limiting standard normal distribution from Theorem 2. Targeting the size of $\alpha = 0.05$, the null hypothesis (2) is rejected when the value of the rescaled statistic exceeds the 0.95th percentile of the standard normal distribution. With ℓ replications of the data set generated under the null hypothesis $\mathcal{H}$, we calculate the empirical size as

$$\hat{\alpha}_T = \frac{\#\{T^{\mathcal{H}}/\sigma \geq z_\alpha\}}{\ell},$$

where $T^{\mathcal{H}}$ represents the values of the test statistic T based on the data generated under the null hypothesis. The results are summarized in Table 1. As expected, in general, the tests have their sizes converging to the nominal level 0.05 as both p and n_k increase together. For certain combinations of p and n_k, the test sizes are not very satisfactory when n_k are very small, but they all become close to the nominal 0.05 level when n_k get above 40–50, indicating that the asymptotic properties of T described by Theorem 2 pitch in.

Next, we consider the power of the tests, as studied in Section 2. The empirical power is calculated as

$$\hat{\beta}_T = \frac{\#\{T^{\mathcal{A}}/\sigma \geq z_\alpha\}}{\ell},$$

where $T^{\mathcal{A}}$ represents the values of the test statistic T based on the data generated under the alternative hypothesis. The results are summarized in Table 1. For different combinations of $\boldsymbol{n} = (n_1, \ldots, n_k)$ and p, we generate data as a set of

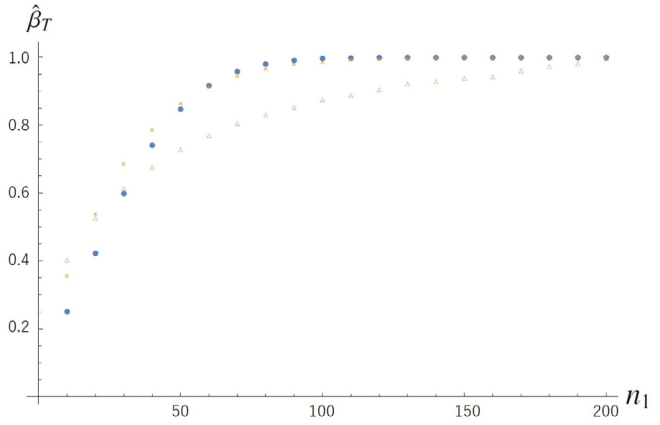

Fig. 2. The empirical powers for the case $(\delta, c) = (0.3, 1.0)$ (•), for the case $(\delta, c) = (0.5, 1.5)$ (×), and for the case $(\delta, c) = (0.7, 2.0)$ (△).

independent draws from two $p \times k$-variate normal distributions with different forms of alternatives of the covariance structure of $\boldsymbol{\Sigma}_{[1]}$. These are

$$\text{(i) } \mathcal{A}_1 : \boldsymbol{\Sigma}_{[1]} = \text{diag}(\boldsymbol{\Sigma}_{11}, \ldots, \boldsymbol{\Sigma}_{55}) + \begin{pmatrix} 0 & \eta & 0 & 0 & 0 \\ \eta & 0 & \eta & 0 & 0 \\ 0 & \eta & 0 & \eta & 0 \\ 0 & 0 & \eta & 0 & \eta \\ 0 & 0 & 0 & \eta & 0 \end{pmatrix} \otimes \boldsymbol{\Sigma}_{11}.$$

$$\text{(ii) } \mathcal{A}_2 : \boldsymbol{\Sigma}_{[1]} = \text{diag}(\boldsymbol{\Sigma}_{11}, \ldots, \boldsymbol{\Sigma}_{55}) + \begin{pmatrix} 0 & \eta & \eta & \eta & 0 \\ \eta & 0 & \eta & \eta & \eta \\ \eta & \eta & 0 & \eta & \eta \\ \eta & \eta & \eta & 0 & \eta \\ 0 & \eta & \eta & \eta & 0 \end{pmatrix} \otimes \boldsymbol{\Sigma}_{11}.$$

Further, for each distribution, two levels of η's are considered: 0.4 and 0.6. The power of the proposed test is largely dependent on (i) the sample size n_g, and (ii) the variation in η as it determines $\sum_{1\le g<h\le k} \rho V_{gh}$, the quantity which in turn determines the asymptotic power of the test as shown in Theorem 3. Specifically, if $\mathcal{A}_1$ holds, then $\sum_{1\le g<h\le k} \rho V_{gh} = 8\eta^2/5$, and if $\mathcal{A}_2$ holds, then $\sum_{1\le g<h\le k} \rho V_{gh} = 67\eta^2/20$. Observe that the value $\sum_{1\le g<h\le k} \rho V_{gh}$ does not depend on the dimension p, hence the power is expected to be mainly related to the value of η and number of η entries in the covariance structures under alternatives. In particular, the first alternative, $\mathcal{A}_1$, is designed to challenge the test procedure for some near block-diagonal structures with sparsely distributed non-zero off-blocks entries, whereas the second alternative, $\mathcal{A}_2$ represents a dense alternative. We find the powers for the alternative $\mathcal{A}_1$ are less affected by the increased dimensionality as compared to $\mathcal{A}_2$. Overall, the power of T under the second alternative increases systematically much faster than that under the first alternative, as the sample sizes and the dimension are increased. Taking into account that the value of $\sum_{1\le g<h\le k} \rho V_{gh}$ is systematically larger under $\mathcal{A}_2$ as compared to $\mathcal{A}_1$, this is a natural trend. And when η increases from 0.4 to 0.6 the power gets larger under both alternatives since the increase of η contributes to the increase of each ρV_{gh}, which measures the departure from the null hypothesis. With η increased under the second alternative, many entries of empirical powers of the test approach 1, which could be viewed as an empirical indication of the proposed test being consistent.

Next, we investigate the behavior of test power when the alternative hypothesis depends on n_1. According to Theorem 3, the power converges to 1 when $\delta < 1$. To reflect the conditions of Theorem 3, we further δ and c:

$$\mathcal{A}_3 : \boldsymbol{\Sigma}_{\langle 1\rangle} = \text{diag}(\boldsymbol{\Sigma}_{11}, \ldots, \boldsymbol{\Sigma}_{55}) + cn_1^{-\delta/2} \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \otimes \boldsymbol{\Sigma}_{11}.$$

We set $(\delta, c) \in \{(0.3, 1.0), (0.5, 1.5), (0.7, 2.0)\}$, $p = 5 \times n_1$, $n_1 = \cdots = n_5$, and $n_1 = 10 \times i$, where $i \in \{1, \ldots, 20\}$. The empirical powers of our proposed test are listed in Fig. 2. As the results in Fig. 2 show, T tends to have rather similar power converging to 1 across the set of alternatives generated by (δ, c). This convergence becomes slower with larger values of δ.

Table 2
The probability of selecting the correct model with the proposed multiple comparison procedure. We check whether the proposed procedure can correctly capture the covariance structure $\boldsymbol{\Sigma}_{\langle 1 \rangle}$. That is, we count the number of times that $\mathcal{H}^{\{2,\{1,3\}\}}$, $\mathcal{H}^{\{2,\{1,4\}\}}$ and $\mathcal{H}^{\{2,\{2,4\}\}}$ are retained by the procedure.

ρ	0.4			0.5			0.6		
$\boldsymbol{n}^\top \backslash p$	100	200	300	100	200	300	100	200	300
(40,40,40,40)	0.321	0.336	0.342	0.889	0.906	0.909	0.999	0.999	0.999
(30,35,45,50)	0.173	0.182	0.183	0.735	0.747	0.750	0.980	0.980	0.981
(50,50,50,50)	0.654	0.674	0.678	0.987	0.989	0.991	1.000	1.000	1.000
(40,45,55,60)	0.490	0.510	0.514	0.948	0.954	0.955	1.000	1.000	1.000
(60,60,60,60)	0.865	0.879	0.881	0.999	0.999	1.000	1.000	1.000	1.000
(50,55,65,70)	0.762	0.779	0.783	0.995	0.996	0.997	1.000	1.000	1.000
(70,70,70,70)	0.960	0.965	0.967	1.000	1.000	1.000	1.000	1.000	1.000
(60,65,75,80)	0.915	0.923	0.926	1.000	1.000	1.000	1.000	1.000	1.000

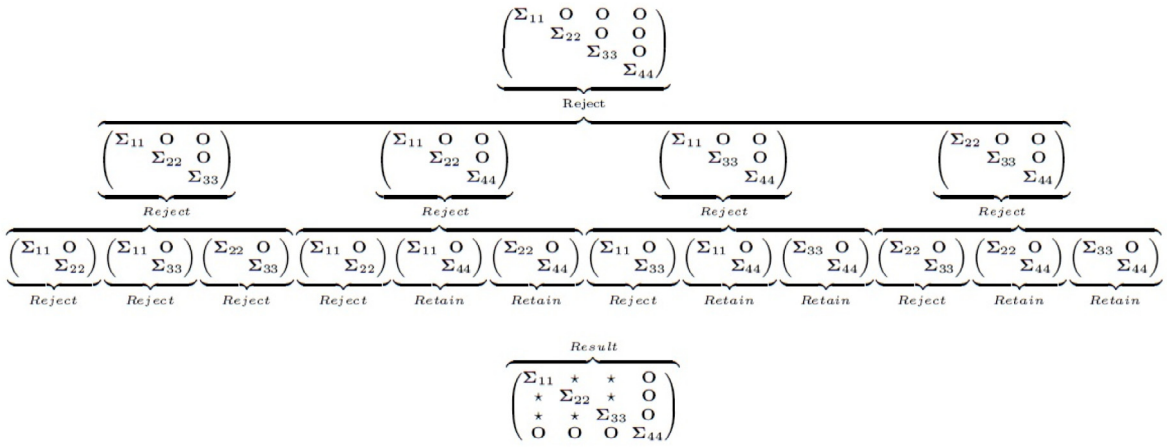

Fig. 3. Three steps of testing independence of the brain activity in four brain channels with EGG data. The symbol in a block of the resulting structure indicates the presence of dependence between the corresponding channels.

Thirdly, we investigate the probability of selecting the correct model with the proposed multiple comparison procedure. Let $\boldsymbol{\Sigma}_{\langle 1 \rangle}$ has the following structure

$$\boldsymbol{\Sigma}_{\langle 1 \rangle} = \mathrm{diag}(\boldsymbol{\Sigma}_{11}, \ldots, \boldsymbol{\Sigma}_{44}) + \begin{pmatrix} 0 & \sqrt{2}\eta & 0 & 0 \\ \sqrt{2}\eta & 0 & \sqrt{6}\eta & 0 \\ 0 & \sqrt{6}\eta & 0 & 2\sqrt{3}\eta \\ 0 & 0 & 2\sqrt{3}\eta & 0 \end{pmatrix} \otimes \boldsymbol{\Sigma}_{11}.$$

The results are summarized in Table 2. As η is larger, the selection probability tends to be larger. Also, as the dimension p increases, the selection probability slightly increases. When the total sample size is the same, the balance type has a higher selection probability than the unbalanced type (see Table 2).

Finally, for comparison purpose, we consider the test based on the asymptotic distribution of the weighted sum of chi-square random variables, that is asymptotic distribution of nRV, from Josse et al. [15]. The weighted sum of chi-squares is approximated with Hall-Buckley-Eagleson approach and unknown parameters of the test statistics are replaced with their large sample consistent counterparts. This method is fully described in Buckley and Eagleson [3]. As seen in Table 3, the size of the tests given by our proposed statistic T is similar across different data-generating distributions, being just slightly higher but stably very close to the nominal level. The stability of the size rendered by T points to our results on the high-dimensional consistency of this statistic which is obtained by the high-dimensionality adjustment suggested by Lemma 1. In contrast, the sizes of Josse-type asymptotic test are satisfactory only when the dimensionality p is small, by they decrease significantly below the target size of 0.05 as the dimension increases. As expected, since the large sample consistency of estimators of unknown parameters collapses in high dimensions, indicating that this type of asymptotic tests is not well adapted for high-dimensional settings.

Table 3
Size comparison of the proposed test and the test by Josse et al. [15] for the setting the nominal level to 0.05, $k = 2$ and $n_1 = n_2$. This table reports Monte-Carlo estimates of the finite-sample sizes for a variety of combinations of n_1 and p. The data underlying the table are i.i.d. $p \times 2$-variate normal and the covariance matrix having the following within-block structures $\boldsymbol{\Sigma}_{(1)} = \mathrm{diag}(\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{22})$, where each $\boldsymbol{\Sigma}_{gg}$ has an $AR(1)$ structure, i.e., $\boldsymbol{\Sigma}_{gg} = g(0.5^{|i-j|})$. For each combination of p and n_1, empirical sizes of the tests are calculated from $\ell = 100,000$ independently generated data sets.

$p \backslash n$	$n = 50$		$n = 100$		$n = 200$		$n = 400$	
	HNP	AT	HNP	AT	HNP	AT	HNP	AT
$p = 5$	0.066	0.041	0.065	0.046	0.064	0.047	0.064	0.049
$p = 50$	0.055	0.006	0.052	0.019	0.052	0.032	0.053	0.040
$p = 100$	0.054	0.000	0.053	0.006	0.053	0.019	0.051	0.031
$p = 200$	0.051	0.000	0.052	0.000	0.050	0.006	0.052	0.018
$p = 400$	0.050	0.000	0.051	0.000	0.050	0.001	0.051	0.008

Table 4
The table shows step, q, m, the total *HRV* coefficient $T^{\{q,m\}}$, standard deviation $\sigma_{\{q,m\}}$, test statistic $T^{\{q,m\}}/\sigma_{\{q,m\}}$, and critical value $z_{1-\alpha_q}$ for each hypothesis. If there is statistical evidence at the 5% level of significance, we denote value of test statistic with $*$.

Step	q	m	$T^{\{q,m\}}$	$\sigma_{\{q,m\}}$	$T^{\{q,m\}}/\sigma_{\{q,m\}}$	$z_{1-\alpha_q}$
1	4	{1, 2, 3, 4}	1.460	0.045	32.445$*$	1.645
2	3	{1, 2, 3}	1.418	0.032	44.562$*$	1.645
	3	{1, 2, 4}	0.586	0.032	18.428$*$	1.645
	3	{1, 3, 4}	0.507	0.032	15.941$*$	1.645
	3	{2, 3, 4}	0.408	0.032	12.836$*$	1.645
3	2	{1, 2}	0.540	0.018	29.406$*$	1.955
	2	{1, 3}	0.486	0.018	26.383$*$	1.955
	2	{1, 4}	0.027	0.018	1.450	1.955
	2	{2, 3}	0.393	0.018	21.394$*$	1.955
	2	{2, 4}	0.020	0.018	1.063	1.955
	2	{3, 4}	−0.004	0.018	−0.224	1.955

4.2. Applications : An example

For illustration, we employ the step-down multiple comparison significance testing to analyze the Electroencephalography (EEG) data publicly available at the University of California-Irvine Machine Learning Repository [1]. The data arose from a large study to examine Electroencephalograph (EEG) correlates of genetic predisposition to alcoholism. Monitoring of the brain electric activity is performed with 64 electrodes evenly distributed over subjects scalps and recording 256 measurements 1 s. The initial study involved two groups of subjects: alcoholics and controls. Each subject was exposed to either a single stimulus (S1) or to two stimuli (S1 and S2) which were pictures of objects chosen from a picture set. The outcome measurements are Event-Related Potentials (ERP) indicating the level of electrical activity in the region of the brain where each of the electrodes is placed.

This data set has been analyzed by several statisticians for various purposes, see, e.g., Harrar and Kong [9] whose main hypotheses of interest whether ERP profiles are similar between the alcoholic and control groups, and if different, to identify for which electrode (which part of the brain) dissimilarity occurred.

In this paper, we conduct the analysis the for the single stimulus (S1) exposure in the alcoholic group. We are interested in testing the independence of the level of electrical activity within the frontal regions of the brain. Specifically, the data set we focus on, consists of four channels (electrodes) FC1, FCz, FC2 and Cz where each channel has names identifying the location of the electrode on the scalp; F stands for frontal lobe, letter z (zero) is used for the mid-line and C identifies the central location between the frontal and parietal lobes. Combinations of two letters indicate intermediate locations, for example FC is in between frontal and central electrode locations (see Fig. 5 of Harrar and Kong [9] for illustration). In the notations of the paper, this data set comprises $k = 4$ sub-vectors (FC1 (1), FCz (2), FC2 (3), Cz (4)), each of dimensionality $p = 256$ with equal sample sizes, that is $n_1 = 77$. The multiple comparison procedure proposed in Section 3 is applied to clarify whether the levels of the brain activity at FC1, FCz, FC2 and Cz channels are mutually independent. Using the significance level $\alpha = 0.05$, the test results for each step are summarized in Table 4.

From Table 4, the hypotheses $\mathcal{H}^{\{2,\{1,2\}\}}$, $\mathcal{H}^{\{2,\{1,3\}\}}$, $\mathcal{H}^{\{2,\{2,3\}\}}$ are rejected, whereas $\mathcal{H}^{\{2,\{1,4\}\}}$, $\mathcal{H}^{\{2,\{2,4\}\}}$, $\mathcal{H}^{\{2,\{3,4\}\}}$ are retained. Hence, with the results above we have strong evidence to believe that the three channels, FC1, FCz, and FC2 correlate with each other, but there is no correlation between (FC1, FCz, FC2) and Cz. This suggests that the assumption on the cross-channel independence in such empirical studies may not be appropriate. The four steps of the testing model for this example are illustrated in Fig. 3.

5. Summary

A test statistic for mutual independence of k random vectors coming for multivariate normal populations is developed when the dimensionality is large, possibly much larger than the sample size. Such a test is usually carried out as a preliminary test in large scale multivariate inference; examples include discriminant analysis or model-based clustering, testing equality of mean vectors and covariance matrices where the ability to detect departures from independence is of crucial importance. Zhu et al. [24] points out the limitations of the usual distance covariance to quantify independence between two random vectors in the high-dimensional setup. Under the non-normal population, it seems there might be some intrinsic difficulties to completely characterize independence between two random vectors in the high-dimensional framework. This motivates us to develop the theory, which is based on the multivariate normality assumption, i.e., when ρV completely characterizes independence between two random vectors. In the context of testing the hypothesis of independence, the applicability of our proposed procedure can be checked first by applying a normality test for high-dimensional data such those proposed by e.g. Himeno and Yamada [10] or non-parametric procedures, proposed by e.g. Chen and Xia [4]. If normality does not hold, ρV simply represents a measure of correlation between two vectors and the developed procedure can be used to test whether this correlation is significant.

The reliance on exact multivariate normality is limiting and one approach to handle this limitation in practical applications is to exploit semi-parametric Gaussian copula developed in Liu et al. [17]. Specifically, the idea of this approach is to replace the random variable $\mathbf{x} = (x_1, \ldots, x_p)$ by the transformed variable $f(\mathbf{x}) = (f_1(x_1) \ldots f_p(x_p))$, and assume that $f(\mathbf{x})$ is multivariate normal. A set of univariate monotone functions $\{f_i\}_{i=1}^p$ is estimated from the data and being defined in this way, this semi-parametric copula results in a nonparametric extension of the normal distribution. The results of Liu et al. [17] also indicate that the nonparanormal distribution can be used as a safe replacement for normality-based estimators even if the data are truly normal.

With a view towards alternatives in which dependence is spread out over vector components, our proposed test statistic is formed as a sum of consistent estimators of the pairwise vector correlation coefficients. The corresponding asymptotic theory is then developed to derive the asymptotic normal limit of the proposed test when both sample size and dimensionality go to infinity. A step-down multiple comparison procedure that allows to control the family-wise error rate under independence is presented as a direct by-product.

Simulation results are used to demonstrate the finite sample performance of the test with respect to its size control and power, for large samples, arbitrary dimensions and a variety of dependence structure models often used in multivariate analysis. Our methodology is illustrated with the Electroencephalography (EEG) data where we applied the proposed step-down procedure for assessment of independence of electrical activity over certain regions of the human brain. This methodology is effective when the number of populations is relatively small, but problems arise when the number of populations is very large. If the number of populations is large, the number of steps becomes enormous, resulting in problems with computational complexity. These are one of the important issue for the future.

CRediT authorship contribution statement

Masashi Hyodo: Conceptualization, Methodology, Writing - original draft, Writing - review & editing. **Takahiro Nishiyama:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing. **Tatjana Pavlenko:** Project administration, Writing - review & editing.

Acknowledgments

We are grateful to the Editor-in-Chief, Associate Editor, and reviewers for many valuable comments and helpful suggestions, which have led to an improved version of this paper. The research of the first and second authors were supported by JSPS KAKENHI Grant Numbers 20K11712, 20K11714. The third author is in part supported by Grant 2013-45266 of the National Research Council of Sweden (VR).

References

[1] University of California-Irvine Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/EEG+Database.
[2] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, third ed., John Wiley and Sons, New York, 2003.
[3] M.J. Buckley, G.K. Eagleson, An approximation to the distribution of quadratic forms in normal random variables, Aust. J. Stat. 30 (1988) 150–159.
[4] H. Chen, Y. Xia, A nonparametric normality test for high-dimensional data, arXiv:1904.05289.
[5] B. Efron, Are a set of microarrays independent of each other? Ann. Appl. Stat. 3 (2009) 922–942.
[6] Y. Escoufier, Le Traitement des variables vectorielles, Biometrics 29 (1973) 751–760.
[7] K.T. Fang, Y.T. Zhang, Generalized Multivariate Analysis, Springer, New York, 1990.
[8] F. Han, H. Liu, Distribution-free tests of independence with applications to testing more structures, arXiv:1410.4179.
[9] S.W. Harrar, X. Kong, High-dimensional multivariate repeated measures analysis with unequal covariance matrices, J. Multivariate Anal. 145 (2016) 1–21.
[10] T. Himeno, T. Yamada, Estimations for some functions of covariance matrix in high dimension under non-normality and its applications, J. Multivariate Anal. 130 (2014) 27–44.
[11] Y. Hochberg, A.C. Tamhane, Multiple Comparison Procedures, John Wiley and Sons, New York, 1987.

[12] M. Hyodo, N. Shutoh, T. Nishiyama, T. Pavlenko, Testing block-diagonal covariance structure for high-dimensional data, Stat. Neerl. 69 (2015) 460–482.
[13] D. Jiang, Z. Bai, S. Zheng, Testing the independence of sets of large-dimensional variables, Sci. China Math. 56 (2013) 135–147.
[14] J. Josse, S. Holmes, Measuring multivariate association and beyond, Stat. Surv. 10 (2016) 132–167.
[15] J. Josse, J. Pagés, F. Husson, Testing the significance of the *RV* coefficient, Comput. Statist. Data Anal. 53 (2008) 82–91.
[16] D. Leung, M. Drton, Testing independence in high dimensions with sums of rank correlations, https://arxiv.org/abs/1501.01732.
[17] H. Liu, F. Han, M. Yuan, J. Lafferty, L. Wasserman, High dimensional semiparametric Gaussian copula graphical models, Ann. Statist. 40 (2012) 2293–2326.
[18] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, in: Probability and Mathematical Statistics: A Series of Monographs and Textbooks, Academic Press, New York, 1979.
[19] K. Pearson, On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Phil. Mag. 50 (1900) 157–175.
[20] A.N. Shiryaev, Probability, second ed., Springer-Verlag, New-York, 1984.
[21] M.S. Srivastava, N.M. Reid, Testing the structure of the covariance matrix with fewer observations than the dimension, J. Multivariate Anal. 112 (2012) 156–171.
[22] G.J. Szkely, M.L. Rizzo, The distance correlation t-test of independence in high dimension, J. Multivariate Anal. 112 (2013) 193–213.
[23] Y. Yang, G. Pan, Independence test for high dimensional data based on regularized canonical correlation coefficients, Ann. Statist. 43 (2015) 467–500.
[24] C. Zhu, S. Yao, X. Zhang, X. Shao, Distance-based and RKHS-based dependence metrics in high-dimension, arXiv:1902.03291v1.