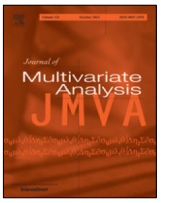
Journal of Multivariate Analysis 178 (2020) 104605

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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



Likelihood ratio tests for many groups in high dimensions

Holger Dette, Nina Dörnemann *

Fakultät für Mathematik, Ruhr-Universität Bochum, 44801 Bochum, Germany

ARTICLE INFO

Article history:
Received 19 June 2019
Received in revised form 27 February 2020
Accepted 27 February 2020
Available online 7 March 2020

AMS 2010 subject classifications:
primary 62H15
secondary 62H10

Keywords:
High-dimensional inference
Likelihood ratio test

ABSTRACT

In this paper, we investigate the asymptotic distribution of likelihood ratio tests in models with several groups, when the number of groups converges with the dimension and sample size to infinity. We derive central limit theorems for the logarithm of various test statistics and compare our results with the approximations obtained from a central limit theorem where the number of groups is fixed.

1. Introduction

Classical multivariate analysis tools as can be found in the text books [2] or [23] are developed under the paradigm that the dimension is substantially smaller than the sample size and do not yield to a reliable statistical inference if this assumption is not satisfied. Because modern datasets, as they occur in biostatistics, wireless communications and finance, are high-dimensional (see, e.g., [10,18] and references therein) there exists an enormous amount of literature developing statistical methods in the case where the dimension of the data is of comparable size (or even larger) than the sample size. Many authors, a few of them mentioned below, have worked on problems of this type and a large part of the literature investigates the asymptotic properties of "classical" test procedures under the assumption that the dimension p is proportional to the sample size n. In the papers [3,20,25], tests for the equality of high-dimensional covariance matrices are developed. The works of [13,31] address the question whether a high-dimensional covariance matrix admits a block-diagonal structure. Moreover, several authors investigate tests for independence of large-dimensional vectors (see [5,7,8,15]). A selection of high-dimensional testing problems using the likelihood ratio principle is considered in [16,17] .

In the case $p < n$ likelihood ratio tests are still well defined and it is shown in many papers that the asymptotic theory under the assumption $\lim_{n,p\to\infty} p/n = y \in (0, 1)$ yields a substantially better approximation for the nominal level of corresponding tests as classical asymptotic considerations keeping the dimension p fixed. In this paper we continue this discussion and investigate approximations for the likelihood ratio test statistics in cases where high-dimensional inference has to be performed for a large number of groups. Consider, for example, the problem of testing if the covariance matrix of a p-dimensional normal distributed vector has a block diagonal structure with q blocks. In Fig. 1 we show the p-values of the corresponding likelihood ratio test (see [29]) under the null hypothesis with 40 blocks of size 18 (thus the total dimension is 720) and sample size 800. The components of all vectors are independent identically standard normal distributed and thus the null hypothesis is obviously satisfied. The left panel shows the simulated p-values (based on

* Corresponding author.
E-mail address: nina.doernemann@ruhr-uni-bochum.de (N. Dörnemann).

https://doi.org/10.1016/j.jmva.2020.104605
0047-259X/

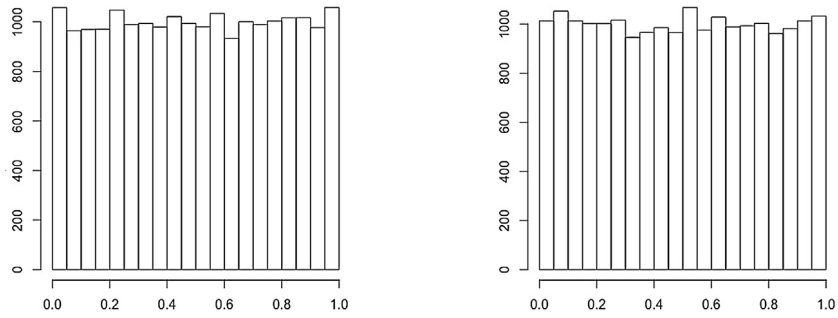


Fig. 1. Simulated p-values of the likelihood ratio test for the hypothesis of a block diagonal structure of 40 blocks of equal size 18 in a $p = 720$-dimensional normal distributed vector (sample size 800). Left panel: asymptotic level α test considering the number of groups as fixed; right panel: asymptotic level α test derived in this paper.

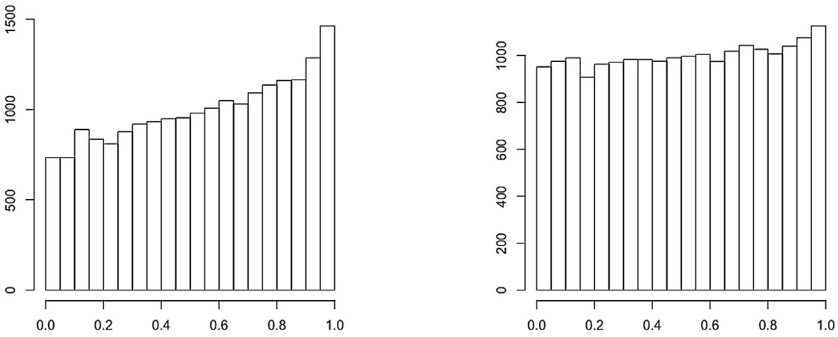


Fig. 2. Simulated p-values of the likelihood ratio test for the equality of q p-dimensional normal distributions, where $n_i = 80,\ p = 40,\ q = 300$. Left: asymptotic level α test ($p, n_i \to \infty$) considering the number of groups as fixed; right panel: asymptotic level α test derived in this paper for $p, n_i, q \to \infty$.

20 000 simulation runs) using the approximation provided by [17] considering the number of blocks as fixed, while the right panel shows the p-values using an approximation with $q \to \infty$ as derived in this paper. The two figures look very similarly.

On the other hand, in Fig. 2 we show the p-values of the likelihood ratio test for testing the equality of q normal distributions $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \ldots, \mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$ and the null hypothesis $\boldsymbol{\mu}_i = \boldsymbol{\mu}_j;\ \boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_j, i, j \in \{1, \ldots, q\}$ (see [30]). The sample size in each group is $n_i = 50$, $i \in \{1, \ldots, q\}$, the dimension is $p = 40$ and $q = 300$ different groups are considered (again all components of all vectors are independent identically standard normal distributed and 20 000 simulation runs have been performed). The left panel of the figure shows the results obtained using the quantiles for the asymptotic distribution obtained for fixed q (see Theorem 3 in [17]) while the right one corresponds to an asymptotic distribution derived in this paper under the assumption that $p, q, n_i \to \infty$ (see Theorem 4 for more details). In this case we observe that the latter approach provides a better approximation of the nominal level.

The present paper is devoted to give some (partial) explanation of observations of this type. We consider classical testing problems in high-dimensional statistical inference, where data can be decomposed in q groups, and investigate the asymptotic properties of likelihood ratio tests for various hypotheses if the dimension p and the number of groups q converge to infinity with increasing sample size. In all cases we establish the asymptotic normality of the log-likelihood ratio after appropriate standardization.

The work, which is most similar in spirit to our paper is the paper of Jiang and Yang [17], who considered the corresponding problems for a fixed number of groups. In contrast to these authors, who used the fact that the moment generating function of the log-likelihood ratio statistic can essentially be expressed as a product of ratios of Gamma functions, we use a central limit theorem for sums of a triangular array of independent random variables (see Theorem 5 in Section 4) to establish asymptotic normality. This approach is also applicable for other high-dimensional problems. As an example, we revisit the problem of testing a linear hypothesis about regression coefficients as considered in [4]. These authors showed the asymptotic normality of the (standardized) log-likelihood ratio test statistic by using recent results about linear spectral statistics of large dimensional F-matrices. With our approach we are able to extend their result and also provide a more handy representation of the asymptotic bias.

2. One sample problems

2.1. Testing for independence

A very prominent problem in high-dimensional data analysis is the problem of testing for the independence of sub-vectors of a multivariate normal distribution. To be precise, let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote a p-dimensional normal distributed vector with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and positive definite variance $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ and assume that $\mathbf{X}$ is decomposed as

$$\mathbf{X} = \left(\mathbf{X}^{(1)^\top}, \dots, \mathbf{X}^{(q_n)^\top}\right)^\top,$$

where $\mathbf{X}^{(i)}$ are vectors of dimension p_i, $i \in \{1, \dots, q_n\}$, such that $\sum_{i=1}^{q_n} p_i = p$. Let

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \dots & \boldsymbol{\Sigma}_{1q_n} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \dots & \boldsymbol{\Sigma}_{2q_n} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{\Sigma}_{q_n1} & \boldsymbol{\Sigma}_{q_n2} & \dots & \boldsymbol{\Sigma}_{q_nq_n} \end{pmatrix}. \tag{1}$$

denote the corresponding decomposition of the covariance matrix, where $\boldsymbol{\Sigma}_{ij} := \operatorname{Cov}(\mathbf{X}^{(i)}, \mathbf{X}^{(j)})$. The hypothesis of independent sub-vectors is formulated as

$$H_0 : \ \boldsymbol{\Sigma}_{ij} = \mathbf{0} \ \text{ for all } i \neq j. \tag{2}$$

Several authors have developed tests for the hypothesis (2) (see [5,7,8,13,15,16,31] among others), and in this section we focus on the likelihood ratio test based on a sample of independent identically distributed observations $\mathbf{X}_1, \dots, \mathbf{X}_n \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In [29], it is shown that the likelihood ratio statistic for the hypothesis (2) is given by

$$\Lambda_n = \frac{|\hat{\boldsymbol{\Sigma}}|^{n/2}}{\prod_{i=1}^{q_n} |\hat{\boldsymbol{\Sigma}}_{ii}|^{n/2}},$$

where

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n} (\mathbf{X}_k - \overline{\mathbf{X}})(\mathbf{X}_k - \overline{\mathbf{X}})^\top \ ,$$

is the common estimator of the covariance matrix, $\overline{\mathbf{X}} = (1/n)\sum_{k=1}^{n} \mathbf{X}_k$ the sample mean and $\hat{\boldsymbol{\Sigma}}_{ij}$ denotes the block in the ith row and jth column of the estimate $\hat{\boldsymbol{\Sigma}}$ corresponding to the decomposition (1). The following result specifies the asymptotic distribution of the likelihood ratio test under the null hypothesis of independent blocks, if the number of blocks q_n is increasing with the sample size. A proof can be found in Section 4. Here and throughout this paper the symbol $\xrightarrow{\mathcal{D}}$ denotes weak convergence.

Theorem 1. *If $q_n \to \infty$, $p_i/n \to \lambda_i \in (0, 1)$, $p/n \to c \in (0, 1)$ and $\sum_{i=1}^{\infty} \lambda_i = c$, then under the null hypothesis (2)*

$$\frac{2}{n} \ln \Lambda_n - s_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2),$$

where

$$\sigma^2 = 2\ln\Big((1-c)^{-1}\prod_{i=1}^{\infty}(1-\lambda_i)\Big), \tag{3}$$

$$s_n = \sum_{i=1}^{q_n}(n - p_i - 1)\ln\big(1 - \tfrac{p_i}{n-1}\big) - (n-p-1)\ln\big(1 - \tfrac{p}{n-1}\big) - \frac{\sigma^2}{4}. \tag{4}$$

Remark 1.

(a) Theorem 1 provides an asymptotic level α test for the null hypothesis (2) by rejecting H_0, whenever

$$-\frac{2}{n}\ln \Lambda_n > \sigma_{n,q_n} u_{1-\alpha} - s_n \ , \tag{5}$$

where $u_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of the standard normal distribution and

$$\sigma^2_{n,q} = 2\ln\Big(\frac{\prod_{i=1}^{q}(1 - \frac{p_i}{n})}{(1 - \frac{p}{n})}\Big). \tag{6}$$

(b) Jiang and Yang [17] derived the asymptotic distribution of the statistic $(2/n)\ln \Lambda_n$ in the case, where the number of groups is fixed, that is $q_n = q$ and the dimension is proportional to the sample size. In particular they showed that under the null hypothesis

$$\frac{\frac{2}{n}\ln \Lambda_n - s_{n,q}}{\sigma_{n-1,q}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \tag{7}$$

where $\sigma^2_{n-1,q}$ is defined in (6) and

$$s_{n,q} = \sum_{i=1}^{q}(n - p_i - \frac{3}{2})\ln(1 - \frac{p_i}{n-1}) - (n - p - \frac{3}{2})\ln(1 - \frac{p}{n-1})$$

(note that these authors use a slightly different notation). The corresponding asymptotic level α test for the null hypothesis (2) rejects H_0, whenever

$$-\frac{2}{n}\ln \Lambda_n > \sigma_{n-1,q}u_{1-\alpha} - s_{q,n}\ . \tag{8}$$

It is easy to see that under the assumptions of Theorem 1

$$\lim_{n\to\infty} \sigma^2_{n-1,q_n} = \sigma^2\ ,$$

where σ^2 is defined in (4). Moreover, recalling the definition of s_n in (4) we obtain by a straightforward calculation

$$\lim_{n\to\infty}\left(s_n - s_{n,q_n}\right) = \lim_{n\to\infty}\left(\frac{1}{2}\sum_{i=1}^{q_n}\ln(1 - \frac{p_i}{n-1}) - \frac{1}{2}\ln(1 - \frac{p}{n-1}) - \frac{\sigma^2}{4}\right) = 0.$$

These results explain why in Fig. 1 the simulated p-values of the likelihood ratio test (8) obtained by a central limit theorem with $p, n \to \infty$, q fixed and the likelihood ratio test (5) obtained by a central limit theorem using $p, n, q \to \infty$ are very similar.

Remark 2 (*Testing for Complete Independence*)**.** Besides, rather than testing for a block diagonal structure of the covariance matrix, one could be interested in testing for complete independence. More precisely, we denote by $\mathbf{R} \in \mathbb{R}^{p\times p}$ the correlation matrix of a p-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The hypothesis of complete independence can then be formulated as

$$H_0 : \mathbf{R} = \mathbf{I}, \tag{9}$$

where $\mathbf{I}$ denotes the identity matrix, in this case of size $p \times p$. Equivalently, one could test for a diagonal structure of the matrix $\boldsymbol{\Sigma}$. For vectors $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ and $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$ the Pearson correlation coefficient between $\mathbf{x}$ and $\mathbf{y}$ is given by

$$r_{\mathbf{x},\mathbf{y}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}},$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i,\ \ \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i.$$

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent normal distributed random variables with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and define the $n \times p$ data matrix by $(\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top = (\mathbf{Y}_1, \ldots, \mathbf{Y}_p)$. The likelihood ratio test statistic for the hypothesis (9) turns out to be (see [22])

$$\Lambda_n = |\mathbf{R}_n|^{\frac{n}{2}},$$

where $\mathbf{R}_n \in \mathbb{R}^{p\times p}$ denotes the sample correlation matrix, that is, the entries are given by $r_{\mathbf{Y}_i,\mathbf{Y}_j}$ for $i, j \in \{1, \ldots, p\}$. Jiang and Yang derived a CLT for the logarithm for this test statistic under the null hypothesis in a large p and large n context (see Corollary 1 in [17]). Their proof is based on an investigation of the moment generating function. Alternatively, one could also invoke the strategy proposed in this paper to prove the CLT. As a consequence of Corollary 1 in [12] and Theorem 4.2.1 in [2], we have for the distribution of the log-likelihood ratio test statistic

$$\frac{2}{n}\ln \Lambda_n \overset{\mathcal{D}}{=} \sum_{j=1}^{p-1}\ln B_j\ ,$$

where $B_1, \dots, B_{p-1}$ are independent Beta-distributed random variables, that is

$$B_j \sim \beta\left(\frac{1}{2}(n-p+j-1), \frac{1}{2}(p-j)\right), \quad j \in \{1, \dots, p-1\}.$$

Therefore, our approach proving CLTs as presented in Section 4 yields a similar result as [17]. In fact, using much the same arguments as given in Section 4 we are able to prove that under H_0 in (9)

$$\frac{2}{n}\ln \Lambda_n - s_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2),$$

whenever $p/n \to y \in (0, 1)$, where

$$s_n = \left(n - p - \frac{3}{2}\right)\ln\left(\frac{n-2}{n-p-1}\right) - (p-1), \ \sigma^2 = -2\left[y + \ln(1-y)\right] > 0.$$

Note that these quantities of the standardization are asymptotically equivalent to the mean and variance given in [17]. For the asymptotic behavior of $(2/n)\ln \Lambda_n$ under the alternative, that is, for a more general form of the true correlation matrix $\mathbf{R}$, we refer the reader to [14] and [21] which provide Gaussian approximations under both the null hypothesis and the alternative.

2.2. Testing a linear hypothesis about regression coefficients

A further problem appears if the p-dimensional (independent) random variables $\mathbf{X}_1, \dots, \mathbf{X}_n$ depend linearly on q-dimensional regressors, say $\mathbf{z}_1, \dots, \mathbf{z}_n$. To be precisely, assume $\mathbf{X}_k \sim \mathcal{N}(\boldsymbol{\beta}\mathbf{z}_k, \boldsymbol{\Sigma})$, $1 \le k \le n$, where the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ is positive definite and $\mathbf{z}_1, \dots, \mathbf{z}_n \in \mathbb{R}^q$ are known design vectors such that the matrix $(\mathbf{z}_1, \dots, \mathbf{z}_n)$ has rank q. Consider the decomposition

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2),$$

with $p \times q_1$ and $p \times q_2$ matrices $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, respectively, such that $q = q_1 + q_2$. We are interested in the hypothesis that the matrix $\boldsymbol{\beta}_1$ coincides with a given matrix $\boldsymbol{\beta}_{01} \in \mathbb{R}^{p\times q_1}$, that is

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{01}. \tag{10}$$

The likelihood ratio test statistic for this hypothesis is given by

$$\Lambda_n := \frac{|\hat{\boldsymbol{\Sigma}}|^{\frac{n}{2}}}{|\hat{\boldsymbol{\Sigma}}_0|^{\frac{n}{2}}},$$

where the $p \times p$ matrices $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Sigma}}_0$ are defined by

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{X}_i - \hat{\mathbf{B}}\mathbf{z}_i)(\mathbf{X}_i - \hat{\mathbf{B}}\mathbf{z}_i)^\top,$$

$$\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{n}\sum_{i=1}^n (\mathbf{X}_i - \boldsymbol{\beta}_{01}\mathbf{z}_{i,1} - \hat{\mathbf{B}}_{20}\mathbf{z}_{i,2})(\mathbf{X}_i - \boldsymbol{\beta}_{01}\mathbf{z}_{i,1} - \hat{\mathbf{B}}_{20}\mathbf{z}_{i,2})^\top,$$

respectively, and

$$\hat{\mathbf{B}} = \left(\sum_{i=1}^n \mathbf{X}_i\mathbf{z}_i^\top\right)\left(\sum_{i=1}^n \mathbf{z}_i\mathbf{z}_i^\top\right)^{-1},$$

$$\hat{\mathbf{B}}_{20} = \left(\sum_{i=1}^n (\mathbf{X}_i - \boldsymbol{\beta}_{01}\mathbf{z}_{i,1})\mathbf{z}_{i,2}^\top\right)\left(\sum_{i=1}^n \mathbf{z}_{i,2}\mathbf{z}_{i,2}^\top\right)^{-1}$$

are the maximum likelihood estimators of $\boldsymbol{\beta}$ under the null hypothesis and alternative, respectively (see [27] or [2]). Here we use the partition of the vector $\mathbf{z}_i^\top = (\mathbf{z}_{i,1}^\top, \mathbf{z}_{i,2}^\top)$ in vectors $\mathbf{z}_{i,1}^\top$ and $\mathbf{z}_{i,2}^\top$ of dimension q_1 and q_2, respectively. In the following theorem, we present the asymptotic null distribution of the likelihood ratio test statistic for a general linear hypothesis (10) in a high-dimensional regression model, where the dimensions $p = p_n$, $q = q_n$, $q_1 = q_{1,n}$ and $q_2 = q_{2,n}$ increase with the sample size. A part of this result, namely the case $p_n/q_{1,n} \to y_1 \in (0, 1)$, has been established by Bai et al. [4] using random matrix theory. In contrast to these authors we are also able to deal with the case $y_1 \in (0, \infty)$.

Theorem 2. *If $p \to \infty$, $q_{1,n} \to \infty$, $n - q_n \to \infty$, $p/q_{1,n} \to y_1 \in (0, \infty)$ and $p/(n-q_n) \to y_2 \in (0, 1)$, then under the null hypothesis (10)*

$$\frac{2}{n}\ln \Lambda_n - s_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2),$$

where

$$\sigma^2 := 2\Big\{\ln\Big(\frac{1}{1-y_2}\Big) - \ln\Big(\frac{y_1+y_2}{y_1+y_2-y_1y_2}\Big)\Big\}, \tag{11}$$

$$\begin{aligned} s_n :=& (n-q_{2,n}-1)\ln\Big(\frac{n-q_n-1}{n-q_{2,n}-1}\Big) + q_{1,n}\ln\Big(\frac{n-q_n-p-1}{n-q_n-1}\Big) \\ &+ (n-q_{2,n}-p-1)\ln\Big(\frac{n-q_{2,n}-p-1}{n-q_n-p-1}\Big) + \frac{\sigma^2}{4}. \end{aligned} \tag{12}$$

Remark 3. Bai et al. [4] considered the testing problem (10) in a similar high-dimensional framework. Note that the authors use the negative log likelihood ratio test statistic $-\ln \Lambda_n$, while Theorem 2 is formulated for $\ln \Lambda_n$. They made use of recent results about linear spectral statistics of large dimensional F-matrices and require a more restrictive condition on the ratio $p_n/q_{1,n}$ to apply this theory, that is $\lim_{n\to\infty} p_n/q_{1,n} = y_1 \in (0,1)$. To be more precise, in [4] it is proven that under the null hypothesis (10)

$$v(f)^{-\frac{1}{2}}\Big(-\frac{2}{n}\ln \Lambda_n - pF_{y_{n_1},y_{n_2}}(f) - m(f)\Big) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \tag{13}$$

whenever $p\to\infty$, $q_{1,n}\to\infty$, $n-q_n\to\infty$, $p/q_{1,n}\to y_1\in(0,1)$ and $p/(n-q_n)\to y_2\in(0,1)$. For an explicit definition of the expression $m(f)$, $v(f)$ and $F_{y_{n_1},y_{n_2}}$, we refer the reader to formulas (26), (27) and (29) in their paper. Theorem 2 extends the result in [4] to the case where $y_1 \geq 1$ and provides a simpler representation of the bias. Moreover, we have checked numerically that the standardizing terms in the central limit theorem stated in (13) and Theorem 2 behave similarly.

Consequently, the likelihood ratio tests based on the asymptotic distribution of Theorem 2 and Theorem 3.1 in [4] have very similar properties. Numerical results, which confirm this observation are not displayed for the sake of brevity.

3. Some q-sample problems

In this section we consider the comparison of q normal distributions $\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1),\dots,\mathcal{N}(\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q)$ with mean vectors $\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_q\in\mathbb{R}^p$ and covariance matrices $\boldsymbol{\Sigma}_1,\dots,\boldsymbol{\Sigma}_q\in\mathbb{R}^{p\times p}$, where for each group a sample of size n_j is available, $j\in\{1,\dots,q\}$, and the dimension and number of groups are increasing with the sample size.

3.1. Testing equality of several covariance matrices

An important assumption for multivariate analysis of variance (MANOVA) is that of equal covariances in the different groups. Thus we are interested in a test of the hypothesis

$$H_0 : \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_q. \tag{14}$$

This problem has been considered by several authors in the context of high-dimensional inference (see [24–26] or [17] among others).

In this section we add to this line of literature and investigate the asymptotic distribution of the likelihood ratio test based on samples of independent distributed observations $\mathbf{X}_{ji}\sim\mathcal{N}(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$, $1\leq i\leq n_j$, $1\leq j\leq q$, when the number of groups is large, i.e., $q\to\infty$. To be precise, let $n=\sum_{j=1}^q n_j$ be the total sample size, then the test statistic of the likelihood ratio test for the hypothesis (14) was derived by [28] and is given by

$$\Lambda_{n,1} = \frac{\prod_{j=1}^q |\mathbf{A}_j/n_j|^{\frac{1}{2}n_j}}{|\mathbf{A}/n|^{\frac{1}{2}n}}, \tag{15}$$

where the $p\times p$ matrices $\mathbf{A}_j$ and $\mathbf{A}$ are defined as

$$\mathbf{A}_j = \sum_{k=1}^{n_j}(\mathbf{X}_{jk}-\bar{\mathbf{X}}_j)(\mathbf{X}_{jk}-\bar{\mathbf{X}}_j)^\top, \quad \mathbf{A}=\sum_{j=1}^q \mathbf{A}_j. \tag{16}$$

As proposed in [6] we consider the modified likelihood ratio test statistic

$$\tilde{\Lambda}_{n,1} = \frac{\prod_{j=1}^q |\mathbf{A}_j/(n_j-1)|^{\frac{1}{2}(n_j-1)}}{|\mathbf{A}/(n-q)|^{\frac{1}{2}(n-q)}},$$

where each sample size n_j is substituted by its degree of freedom. Our next result deals with asymptotic distribution of the test statistic $\ln\tilde{\Lambda}_{n,1}$ for an increasing dimension and an increasing number of groups.

Theorem 3. *Let $n_j + 1 > p$ for all $1 \le j \le q$, $p \to \infty$, $q = o(\sqrt{p(n-q)})$, $p = o(n-q)$, assume that*

$$\limsup_{n\to\infty} \sup_{1\le j\le q} \frac{p}{n_j - 1} < 1$$

and that

$$\sigma_n^2 = \frac{1}{2}\sum_{j=1}^{q} \frac{(n_j-1)^2}{p(n-q)} \ln\Big(\frac{n_j-1}{n_j-p-1}\Big) - \frac{1}{2}$$

converges with a positive limit, say $\sigma^2 > 0$. Then, under the null hypothesis (14),

$$\frac{\ln \tilde{\Lambda}_{n,1} - \tilde{s}_n}{\sqrt{p(n-q)}} \to \mathcal{N}(0, \sigma^2),$$

where

$$\tilde{s}_n = \sum_{j=1}^{q} \frac{n_j-1}{2}\Big\{\Big(n_j - \frac{3}{2}\Big)\ln\Big(\frac{n_j-2}{n_j-p-2}\Big) - p\ln\Big(\frac{n_j-1}{n_j-p-2}\Big)\Big\} - \frac{n-q}{2}(n-q-p)\ln\Big(\frac{n-q}{n-q-p}\Big),$$

Remark 4.

(a) Note that under the assumptions of Theorem 3 the asymptotic distributions of $\ln \Lambda_{n,1}$ and $\ln \tilde{\Lambda}_{n,1}$ are not identical. In fact we have

$$\ln \Lambda_{n,1} - \ln \tilde{\Lambda}_{n,1} = \frac{1}{2}\sum_{j=1}^{q} pn_j \ln\Big(\frac{n_j-1}{n_j}\Big) - \frac{1}{2}pn\ln\Big(\frac{n-q}{n}\Big) - \sum_{j=1}^{q}\frac{1}{2}p\ln\Big(\frac{n_j-1}{n-q}\Big),$$

and in general this is not of order $o(\sqrt{p(n-q)})$ (consider for example the case $n_j = 2p+1$, $q = o(p^2)$, $q \to \infty$).

(b) Jiang and Yang [17] determined the asymptotic distribution of the statistic $\ln \tilde{\Lambda}_{n,1}$ for a fixed number q assuming that the limit p/n does not vanish. In particular, they showed that under the null hypothesis (14)

$$\frac{\ln \tilde{\Lambda}_{n,1} - \mu_n}{(n-q)\sigma_n^{(1)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \tag{17}$$

as $n_j, n, p \to \infty$, where the asymptotic bias and variance are given by

$$\mu_n = \frac{1}{4}\Big\{(n-q)(2n-2p-2q-1)\ln\Big(1-\frac{p}{n-q}\Big) - \sum_{j=1}^{q}(n_j-1)(2n_j-2p-3)\ln\Big(1-\frac{p}{n_j-1}\Big)\Big\},$$

$$(\sigma_n^{(1)})^2 = \frac{1}{2}\Big\{\ln\Big(1-\frac{p}{n-q}\Big) - \sum_{j=1}^{q}\Big(\frac{n_j-1}{n-q}\Big)^2\ln\Big(1-\frac{p}{n_j-1}\Big)\Big\},$$

respectively. As q being fixed, the authors assumed for their result that $p/n_j \to y_j \in (0,1]$ for all $j = 1,\dots,q$ and $\min_j n_j > p+1$. Note that the order of standardization in Theorem 3 is different than in (17). The standardization is of order $\sqrt{p(n-q)}$ which is, under the assumptions of Theorem 3, substantially smaller than $n-q$ as used in (17). Comparing the variance σ^2 in Theorem 3 with an adjusted version of $(\sigma_n^{(1)})^2$ (such that the different standardizations are canceled out) yields under the assumptions of Theorem 3

$$\sigma^2 - \frac{n-q}{p}(\sigma_n^{(1)})^2 = -\frac{1}{2} - \frac{1}{2}\frac{n-q}{p}\ln\Big(1-\frac{p}{n-q}\Big) = o(1).$$

On the other hand the difference of the means is given by

$$\tilde{s}_n - \mu_n = \frac{n-q}{4}\ln\Big(\frac{n-q-p}{n-q}\Big) + \sum_{j=1}^{q}\frac{n_j-1}{2}\Big\{p\ln\Big(\frac{n_j-p-2}{n_j-p-1}\Big) + \Big(n_j-\frac{3}{2}\Big)\Big[\ln\Big(\frac{n_j-2}{n_j-p-2}\Big) - \ln\Big(\frac{n_j-1}{n_j-p-1}\Big)\Big]\Big\}$$

Note that the first summand divided by the standardization $\sqrt{p(n-q)}$ vanishes under the assumptions of Theorem 3, while the other terms give a notable contribution to the expected value. Thus we expect that the corresponding likelihood ratio tests behave differently, if the number of groups is large.

3.2. Testing equality of several normal distributions

We consider the same setting as in Section 3.1 but this time we want to test whether q normal distributions are identical, that is,

$$H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_q, \ \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_q. \tag{18}$$

The test statistic of the likelihood ratio test for the hypothesis (18) is given by

$$\Lambda_n = \frac{\prod_{j=1}^q |\mathbf{A}_j/n_j|^{\frac{1}{2}n_j}}{|\mathbf{B}/n|^{\frac{1}{2}n}}, \tag{19}$$

where the $p \times p$ matrix $\mathbf{A}$ is defined in (16) and

$$\mathbf{B} = \sum_{j=1}^q \sum_{i=1}^{n_j} (\mathbf{X}_{ji} - \bar{\mathbf{X}})(\mathbf{X}_{ji} - \bar{\mathbf{X}})^\top = \mathbf{A} + \sum_{j=1}^q n_j (\bar{\mathbf{X}}_j - \bar{\mathbf{X}})(\bar{\mathbf{X}}_j - \bar{\mathbf{X}})^\top.$$

Note that Λ_n is the product of the likelihood ratio statistic $\Lambda_{n,1}$ in (15) for testing equality of covariance matrices and the likelihood ratio test statistic for testing equality of the means (see [11,32]). Several authors dealt with testing H_0 (see, e.g., [17,30]). The following result specifies the asymptotic distribution of the statistic $\ln \Lambda_n$ for increasing dimension and an increasing number of groups.

Theorem 4. *Let $n_j > p + 1$ for all $1 \le j \le q$, $p \to \infty$, $q = o(\sqrt{pn})$, $p = o(n)$, assume that*

$$\limsup_{n\to\infty} \sup_{1\le j\le q} \frac{p}{n_j} < 1$$

and that

$$\tilde{\sigma}_n^2 = \frac{1}{2} \sum_{j=1}^q \frac{n_j^2}{pn} \ln\Big(\frac{n_j - 1}{n_j - p - 1}\Big) - \frac{1}{2}$$

converges to a positive limit, say $\sigma^2 > 0$. Then, under the null hypothesis (18) we have

$$\frac{\ln \Lambda_n - \tilde{s}_n}{\sqrt{pn}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2),$$

where

$$\begin{aligned}
\tilde{s}_n := &\sum_{j=1}^q \frac{n_j}{2} \Big\{ \Big(n_j - p - \frac{3}{2}\Big) \ln\Big(\frac{n_j - 2}{n_j - p - 2}\Big) + p \ln\Big(\frac{n_j - 2}{n - q}\Big) - p \ln(n_j) \Big\} \\
&+ \frac{n}{2} \Big\{ p \ln\Big(\frac{\frac{1}{2}n - 1}{\frac{1}{2}n + \frac{1}{2}q - \frac{3}{2}}\Big) + (n - q - p) \ln\Big(\frac{n - q - p}{n - q}\Big) + p \ln(n) \\
&+ \Big(\frac{1}{2}n - p - 1\Big) \ln\Big(\frac{\frac{1}{2}n - 1}{\frac{1}{2}n - p - 1}\Big) + \Big(\frac{1}{2}n - p + \frac{1}{2}q - \frac{3}{2}\Big) \ln\Big(\frac{\frac{1}{2}n - p + \frac{1}{2}q - \frac{3}{2}}{\frac{1}{2}n + \frac{1}{2}q - \frac{3}{2}}\Big) \Big\}.
\end{aligned}$$

Remark 5. The asymptotic distribution of the statistic Λ_n in the case where q is fixed was determined in Theorem 3 of [17] who showed that under the null hypothesis (18)

$$\frac{\ln \Lambda_n - \tilde{\mu}_n}{n\tilde{\sigma}_n^{(1)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \tag{20}$$

if $\lim_{n\to\infty} \frac{p}{n_j} = y_j \in (0, 1]$, $1 \le j \le q$. Here the asymptotic bias and variance are given by

$$\tilde{\mu}_n = \frac{1}{4} \Big\{ -2qp - \sum_{j=1}^q y_j - n(2p - 2n + 3) \ln\Big(1 - \frac{p}{n}\Big) + \sum_{j=1}^q n_j (2p - 2n_j + 3) \ln\Big(1 - \frac{p}{n_j - 1}\Big) \Big\},$$

$$(\tilde{\sigma}_n^{(1)})^2 = \frac{1}{2} \Big\{ \ln\Big(1 - \frac{p}{n}\Big) - \sum_{j=1}^q \frac{n_j^2}{n^2} \ln\Big(1 - \frac{p}{n_j - 1}\Big) \Big\}$$

(note that these authors use a slightly different notation). It is important to note that the orders in the standardizations in both results are different. While the standardizing factor in (20) is of order n, it is of order $\sqrt{pn} = o(n)$ in Theorem 4.

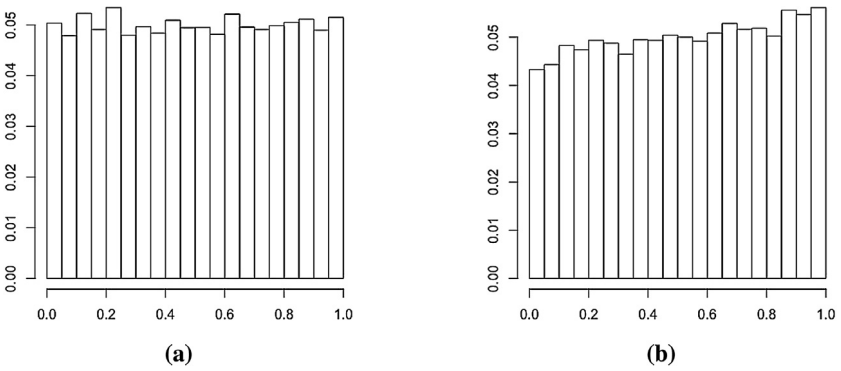


Fig. 3. Simulated p-values of the likelihood ratio test for the hypothesis (18) under the null hypothesis ($n_j = 200$, $p = 100$, $q = 50$). Left panel: asymptotic level α test considering the number of groups as fixed; right panel: asymptotic level α test derived from Theorem 4.

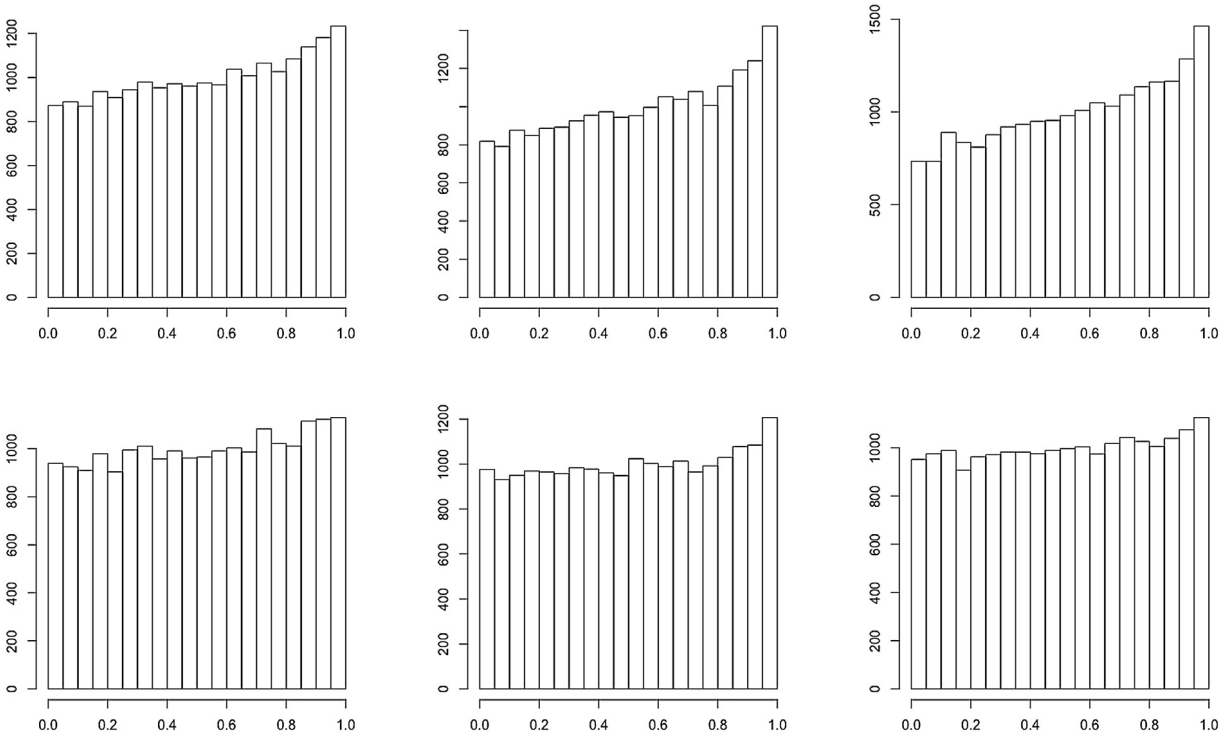


Fig. 4. Simulated p-values of the likelihood ratio test for the hypothesis (18) under the null hypothesis, where $n_j = 80$, $p = 40$ and $q = 100$ (left column), $q = 200$ (middle column), $q = 300$ (right column). Upper part: asymptotic level α test considering the number of groups as fixed; lower part: asymptotic level α test derived from Theorem 4.

Similarly, as in Remark 4 it can be shown that

$$\tilde{\sigma}_n^2 - \frac{n}{p}(\tilde{\sigma}_n^{(1)})^2 = o(1)$$

under the assumptions of Theorem 4, while in general the difference $\tilde{s}_n - \tilde{\mu}_n$ is not of order $o(\sqrt{pn})$ (consider, for example, the case $q = p$, $n_j = 2p + 1$, $n = p(2p + 1)$, $p \to \infty$). Based on these observations we expect differences in the likelihood ratio test, if the quantiles from the normal approximation for fixed q as derived in [17] or the quantiles from Theorem 4 are used as critical values. This is illustrated in Figs. 3 and 4, where we display the simulated p-values for the tests under the null hypothesis.

In Fig. 3 we consider the case $n_j = 200$, $p = 100$ and a relatively small number of groups $q = 50$. We observe that both approximations yield histograms close to the expected uniform distribution. On the other hand in Fig. 4 we consider the cases $n_j = 80$, $p = 40$, and $q = 100, 200$ and 300 and we observe larger differences in both approximations. In particular

the critical values derived from Theorem 4 yield a likelihood ratio test for the hypothesis (14) with a better performance than the test using the quantiles from fixed q asymptotics.

For a further evaluation of the performance of the tests based on the two approximation methods, in particular on the power, we display in Fig. 5 the empirical type one error (false detection rate under the null hypothesis — FDR) and the rejection proportion under the alternative. Each point of the figure represents the result of a test under the null hypothesis (first coordinate) and alternative (second coordinate). The corresponding procedure is said to have a good FDR control if the point is close to the vertical line defining the nominal level α (here 5%, 10%, 20% and 30%). The method is said to have a good performance in identifying true signals if the correct rejection proportion is high (i.e. the corresponding point has a large second coordinate).

For the alternative we considered three different models for the distribution of $\mathbf{X}_{ji} \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $1 \leq i \leq n_j$, $1 \leq j \leq q$, namely

$$\boldsymbol{\Sigma}_1 = \mathbf{I},\ \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_q = 2\mathbf{I};\ \boldsymbol{\mu}_1 = \mathbf{0},\ \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_q = (1, \ldots, 1)^\top, \tag{21}$$

$$\boldsymbol{\Sigma}_1 = 0.2\mathbf{J} + 0.8\mathbf{I},\ \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_q = \mathbf{I};\ \boldsymbol{\mu}_1 = (\underbrace{1, \ldots, 1}_{p/2}, \underbrace{0, \ldots, 0}_{p/2})^\top,\ \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_q = \mathbf{0}, \tag{22}$$

$$\boldsymbol{\Sigma}_1 = 0.2\mathbf{J} + 0.6\mathbf{I},\ \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_q = \mathbf{I};\ \boldsymbol{\mu}_1 = (\underbrace{1, \ldots, 1}_{p/4}, \underbrace{0, \ldots, 0}_{3p/4})^\top,\ \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_q = \mathbf{0}, \tag{23}$$

where each entry of the $p \times p$ matrix $\mathbf{J}$ is equal to 1. We observe that both methods control the FDR and yield reasonable power in all cases under consideration. While the results are very similar for $q = 100$ groups, in particular for tests with nominal level 5% and 10% (left column of Fig. 5), we observe more substantial differences for $q = 200$ and $q = 300$ groups (see the middle and right column of Fig. 5). Here, the approximation provided by Theorem 4 yields a test with a better approximation of the nominal level (i.e., the points are closer to the vertical axes) and with more power. Thus the test provided in Theorem 4 outperforms the test in [17] for large values of q.

4. Some proofs

In this section we present proofs of our results, where we restrict ourselves to the proofs of Theorems 1 and 2. The other statements are shown by similar arguments, which are omitted for the sake of brevity.

4.1. A central limit theorem

We begin stating a central limit theorem, which is used in the proofs of Theorems 1–4. We make extensive use of a central limit theorem for triangular array of independent random variables, which follows by similar arguments as given in [9]. Therefore the proof is omitted.

Theorem 5. *Let $(T_n)_{n\in\mathbb{N}}$ be a sequence of finite sets $\{(X_n(i))_{i\in T_n} | n \in \mathbb{N}\}$ denote an array of random variables and $\{(g_n(i))_{i\in T_n} | n \in \mathbb{N}\}$ an array of weights satisfying the following conditions:*

(i) The random variables $(X_n(i))_{i\in T_n}$ are independent for all $n \in \mathbb{N}$.
(ii) The random variables $X_n(i)$ are centered, that is, $\mathrm{E}[X_n(i)] = 0\ \forall i \in T_n,\ n \in \mathbb{N}$.
(iii) $\mathrm{E}\left[X_n^4(i)\right] \leq C\mathrm{E}\left[X_n^2(i)\right]^2$ for some universal constant $C > 0$ and for all $n \in \mathbb{N}$.
(iv) $\sup_{i\in T_n} g_n^2(i)\,\mathrm{Var}(X_n(i)) \to 0$ for $n \to \infty$.
(v) There exists a constant $\sigma^2 > 0$ such that

$$\sum_{i\in T_n} g_n^2(i)\,\mathrm{Var}(X_n(i)) \to \sigma^2 \text{ for } n \to \infty.$$

Then the random variable

$$Z_n := \sum_{i\in T_n} g_n(i) X_n(i)$$

converges in distribution to a normal distribution with mean 0 and variance σ^2.

Proof of Theorem 1. Define

$$V_n = \Lambda_n^{2/n} = \frac{|\hat{\boldsymbol{\Sigma}}|}{\prod_{i=1}^{q_n} |\hat{\boldsymbol{\Sigma}}_{ii}|}$$

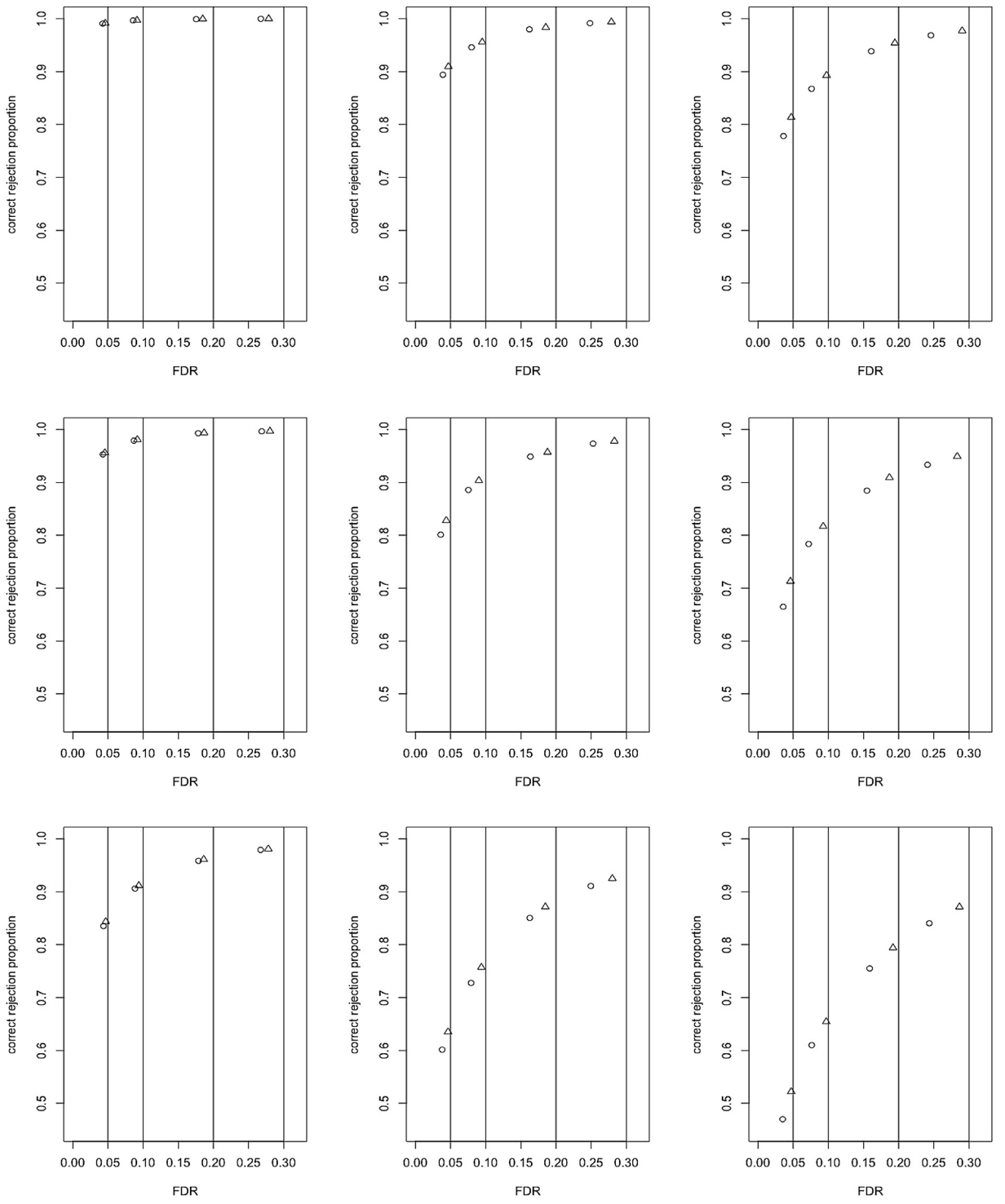


Fig. 5. Comparison of tests for the hypothesis (18) under the null hypothesis and under three alternatives, where $n_j = 80$, $p = 40$ and $q = 100$ (left column), $q = 200$ (middle column), $q = 300$ (right column). Upper part : alternative (21); middle part: alternative (22); bottom part: alternative (23). Circle: asymptotic level α test considering the number of groups as fixed; triangle: asymptotic level α test derived from Theorem 4. The vertical black lines represent the prescribed levels $\alpha \in \{0.05, 0.1, 0.2, 0.3\}$. The first coordinate of a point describes the empirical FDR and the second coordinate the rejection proportion under the alternative.

and note that under the null hypothesis (2) the distribution of V_n is given by a product of independent Beta-distributions (see [2]), that is

$$V_n \stackrel{\mathcal{D}}{=} \prod_{i=2}^{q_n}\prod_{j=1}^{p_i} V_{ij},$$

where the random variables $V_{i,j} \sim \beta((n-p_i^*-j)/2, p_i^*/2)$ are independent and $p_i^* = \sum_{l=1}^{i-1} p_l$. Consequently, with the notation

$$S_n := \ln V_n = \sum_{i=2}^{q_n}\sum_{j=1}^{p_i} \ln(V_{i,j})$$

the assertion follows from

$$S_n - s_n \stackrel{\mathcal{D}}{\to} \mathcal{N}(0,\sigma^2), \tag{24}$$

where σ^2 and s_n are defined in (3) and (4), respectively.

For a proof of this statement we use Theorem 5 and show that conditions (i)–(v) in this result are satisfied. We begin with a calculation of the variance of S_n noting that the variance of logarithm of a Beta distributed random variable $B \sim \beta(a,b)$ is given by

$$\mathrm{Var}(\ln(X)) = \psi_1(a) - \psi_1(a+b), \tag{25}$$

where $\psi_k(x) = \frac{d^{k+1}}{dx^{k+1}} \ln\Gamma(x)$ $(k \geq 0)$ denotes the polygamma function of order k (see [1]). This yields

$$\mathrm{Var}(S_n) = A_n^{(1)} - A_n^{(2)}, \tag{26}$$

where

$$A_n^{(1)} = \sum_{i=2}^{q_n}\sum_{j=1}^{p_i} \psi_1((n-p_i^*-j)/2),\ A_n^{(2)} = \sum_{i=2}^{q_n}\sum_{j=1}^{p_i} \psi_1((n-j)/2).$$

Observing the expansion $\psi_1(z) = \frac{1}{z} + O(z^{-2})$ for the logarithmic Gamma function of order 1, (see [1]) we obtain for the first term

$$A_n^{(1)} = \sum_{i=n-p}^{n-p_1-1} \psi_1(i/2) = \sum_{i=n-p}^{n-p_1-1} \frac{2}{i} + o(1) = 2\ln\Big(\frac{1-p_1/n-1/n}{1-p/n-1/n}\Big) + O(n^{-1}) = 2\ln\left(\frac{1-\lambda_1}{1-c}\right)(1+o(1)), \tag{27}$$

where we used the expansion

$$\sum_{k=1}^{n} \frac{1}{k} = \ln n + \gamma + \frac{1}{2n} + O(n^{-2}) \tag{28}$$

(here γ denotes the Euler–Mascheroni constant). For the second term we use the same expansion as above and obtain

$$A_n^{(2)} = \sum_{i=2}^{q_n}\sum_{j=1}^{p_i} \psi_1((n-j)/2) = \sum_{i=2}^{q_n}\sum_{j=1}^{p_i}\Big\{\frac{2}{n-j} + O((n-j)^{-2})\Big\} = \sum_{i=2}^{q_n}\sum_{j=1}^{p_i} \frac{2}{n-j} + O(n^{-1}), \tag{29}$$

where we used for last equality that $p_i \leq n-p_1$ for sufficiently large n and the fact that $p_i/n \to \lambda_i$ for all $i \geq 1$. Observing the expansion for the harmonic series in (28) it follows that

$$\begin{aligned}
A_n^{(2)} &= \sum_{i=2}^{q_n}\sum_{j=n-p_i}^{n-1} \frac{2}{j} + O(n^{-1}) = \sum_{i=2}^{q_n} 2\Big\{\ln(n-1) + \frac{1}{2(n-1)} + O(n^{-2}) \\
&\quad - 2\Big[\ln(n-p_i-1) + \frac{1}{2(n-p_i-1)} + O((n-p_i-1)^{-2})\Big]\Big\} + O(n^{-1}) \\
&= \sum_{i=2}^{q_n} 2\Big\{-\ln\Big(1-\frac{p_i}{n-1}\Big) + \frac{p_i}{2(n-1)(n-p_i-1)}\Big\} + O(n^{-1}) \\
&= \sum_{i=2}^{q_n} -2\ln\Big(1-\frac{p_i}{n-1}\Big) + O(n^{-1}) = \Big(-2\sum_{i=2}^{\infty}\ln(1-\lambda_i)\Big)\big(1+o(1)\big).
\end{aligned} \tag{30}$$

Here the last equality is a consequence of the theorem of the dominated convergence (see [19], Theorem 1.21). Combining (26), (27) and (30) finally shows

$$\mathrm{Var}(S_n) = 2\ln\Big((1-c)^{-1}\prod_{i=1}^{\infty}(1-\lambda_i)\Big) + o(1)\,,$$

which yields the asymptotic variance σ^2 in (3) and assumption (v) in Theorem 5.

Moreover, as the function ψ_1 is positive and decreasing we have

$$\sup_{i,j}\mathrm{Var}(\ln(V_{i,j})) = \sup_{i,j}\Big\{\psi_1\Big(\frac{n-p_i^\star-j}{2}\Big) - \psi_1\Big(\frac{n-j}{2}\Big)\Big\} \le \psi_1\Big(\frac{n-p-1}{2}\Big) = \frac{2}{n-p-1} + O\Big(\frac{1}{(n+p-1)^2}\Big) = o(1).$$

(note that $p/n \to c < 1$), which proves (iv). The conditions (i) and (ii) are obviously satisfied and the remaining inequality (iii) for the moments is a consequence of Lemma A.7 and Theorem A.8. in [9]. Consequently, we obtain from Theorem 5 the weak convergence

$$S_n - \mathrm{E}[S_n] \to \mathcal{N}(0,\sigma^2)\,,$$

and remains to calculate the representation of the expectation. For this purpose note that

$$\mathrm{E}[S_n] = \sum_{i=2}^{q_n}\sum_{j=1}^{p_i}\psi_0((n-p_i^*-j)/2) - \psi_0((n-j)/2) = B_n^{(1)} - B_n^2, \tag{31}$$

where

$$B_n^{(1)} = \sum_{i=n-p}^{n-p_1-1}\psi_0(i/2)\,,\ B_n^{(2)} = \sum_{i=2}^{q_n}\sum_{j=1}^{p_i}\psi_0((n-j)/2). \tag{32}$$

Observing the expansion $\psi_0(z) = \ln(z) - \frac{1}{2z} + O(z^{-2})$ and (28) we obtain

$$\begin{aligned}
B_n^{(1)} &= \sum_{i=n-p}^{n-p_1-1}\Big\{\ln(i/2) - \frac{1}{i} + O(i^{-2})\Big\} = \ln\Big(\frac{(n-p_1-1)!}{(n-p-1)!}\Big) - \ln(2)(p-p_1) - \ln\Big(\frac{n-p_1-1}{n-p-1}\Big)\\
&\quad - \frac{p_1-p}{(n-p_1-1)(n-p-1)} + O((n-p_1-1)^{-2}) + O((n-p-1)^{-2})\\
&= \ln\Big(\frac{(n-p_1-1)!}{(n-p-1)!}\Big) - \ln(2)(p-p_1) - \ln\Big(\frac{n-p_1-1}{n-p-1}\Big) + o(1)
\end{aligned}$$

and

$$\begin{aligned}
B_n^{(2)} &= \sum_{i=2}^{q_n}\sum_{j=1}^{p_i}\psi_0((n-j)/2) = \sum_{i=2}^{q_n}\sum_{j=1}^{p_i}\Big\{\ln(n-j) - \ln(2) - \frac{1}{n-j} + O((n-j)^{-2})\Big\}\\
&= \Big\{\ln\Big(\prod_{i=2}^{q_n}\frac{(n-1)!}{(n-p_i-1)!}\Big) - \ln(2)(p-p_1) + \sum_{i=2}^{\infty}\ln(1-\lambda_i)\Big\} + o(1),
\end{aligned}$$

where we used similar arguments as in the derivation of (29) and (30). Combining these results with (31) finally yields

$$\mathrm{E}[S_n] = \ln\Big(\prod_{i=1}^{q_n}(n-p_i-1)!\Big) - \ln((n-p-1)!(n-1)!^{q_n-1}) - \frac{\sigma^2}{2} + o(1).$$

Finally an application of Stirling's formula

$$\ln(n!) = n\ln n - n + \frac{1}{2}\ln(2\pi n) + \frac{1}{12n} + O(n^{-3})$$

yields for the first term

$$\begin{aligned}
&\ln\Big(\prod_{i=1}^{q_n}(n-p_i-1)!\Big) - \ln((n-p-1)!(n-1)!^{q_n-1})\\
&= \sum_{i=1}^{q_n}\Big\{(n-p_i-1)\ln(n-p_i-1) - (n-p_i-1) + \frac{\ln(2\pi(n-p_i-1))}{2} + \frac{1}{12(n-p_i-1)}\\
&\qquad + O((n-p_i-1)^{-3})\Big\} - \Big\{(n-p-1)\ln(n-p-1) + (q_n-1)(n-1)\ln(n-1)
\end{aligned}$$

$$
\begin{aligned}
&- (n-p-1) - (q_n-1)(n-1) + \frac{\ln(2\pi(n-p-1)) + (q_n-1)\ln(2\pi(n-1))}{2} \\
&\quad + \frac{1}{12(n-p-1)} + \frac{q_n-1}{12(n-1)} + O((n-p-1)^{-3}) + q_n O((n-1)^{-3}) \Big\} \\
&= \sum_{i=1}^{q_n} \Big\{ (n-p_i-1)\ln(1-p_i/(n-1)) + \frac{\ln(1-p_i/(n-1))}{2} + \frac{1}{12(n-p_i-1)} - \frac{1}{12(n-1)} \Big\} \\
&\quad - \Big\{ (n-p-1)\ln(1-p/(n-1)) + \frac{\ln(1-p/(n-1))}{2} + \frac{1}{12(n-p-1)} - \frac{1}{12(n-1)} \Big\} + o(1) \\
&= \Big\{ \sum_{i=1}^{q_n} (n-p_i-1)\ln(1-p_i/(n-1)) \Big\} - (n-p-1)\ln(1-p/(n-1)) + \frac{\sigma^2}{4} + o(1),
\end{aligned}
$$

and the representation for the expectation in (4) follows, completes the proof Theorem 1.

Proof of Theorem 2. Define $\tilde{U}_n = \Lambda_n^{2/n}$ and note that under the null hypothesis, the distribution of $\tilde{U}_n$ is given by a product of independent Beta distributions (see [2]), that is,

$$\tilde{U}_n = \prod_{i=1}^{p} U_i,$$

where the random variables

$$U_i \sim \beta\left(\frac{1}{2}(n-q_n+1-i), \frac{1}{2}q_{1,n}\right).$$

Now consider the transformation

$$S_n := \frac{2}{n}\ln \Lambda_n = \sum_{i=1}^{p} \ln(U_i),$$

then the assertion follows from

$$S_n - s_n \xrightarrow{\mathcal{D}_{H_0}} \mathcal{N}(0, \sigma^2),$$

where s_n and σ^2 are defined in (11) and (12). In order to prove the asymptotic normality of $S_n - s_n$, we show that conditions (i)–(v) hold, beginning with a derivation of the variance.

$$
\begin{aligned}
\sum_{i=1}^{p} \mathrm{Var}(\ln(U_i)) &= \sum_{i=1}^{p} \Big\{ \psi_1\Big(\frac{1}{2}(n-q_n+1-i)\Big) - \psi_1\Big(\frac{1}{2}(n-q_{2,n}+1-i)\Big) \Big\} \\
&= \sum_{i=1}^{p} \Big\{ \frac{2}{n-q_n+1-i} + \mathcal{O}\Big(\frac{1}{(n-q_n+1-i)^2}\Big) - \frac{2}{n-q_{2,n}+1-i} + \mathcal{O}\Big(\frac{1}{(n-q_{2,n}+1-i)^2}\Big) \Big\} \\
&= 2\Big\{ \sum_{i=n-q_n+1-p}^{n-q_n} \frac{1}{i} - \sum_{i=n-q_{2,n}+1-p}^{n-q_{2,n}} \frac{1}{i} \Big\} + o(1) = 2\Big\{ \ln\Big(\frac{n-q_n}{n-q_n-p}\Big) - \ln\Big(\frac{n-q_{2,n}}{n-q_{2,n}-p}\Big) \Big\} \\
&\quad + \mathcal{O}\Big(\frac{1}{n-q_n-p}\Big) \\
&= 2\Big\{ \ln\Big(\frac{1}{1-y_2}\Big) - \ln\Big(\frac{y_1+y_2}{y_1+y_2-y_1y_2}\Big) \Big\} + o(1).
\end{aligned}
$$

Regarding the error terms, note that the assumptions of Theorem 2 imply $n-q_n-p \to \infty$ and $\frac{p}{n-q_n-p} = \mathcal{O}(1)$. We continue expanding the expected value

$$
\begin{aligned}
\sum_{i=1}^{p} \mathrm{E}(\ln(U_i)) &= \sum_{i=1}^{p} \Big\{ \psi_0\Big(\frac{1}{2}(n-q_n+1-i)\Big) - \psi_0\Big(\frac{1}{2}(n-q_{2,n}+1-i)\Big) \Big\} \\
&= \sum_{i=1}^{p} \Big\{ \ln\Big(\frac{1}{2}(n-q_n+1-i)\Big) - \frac{1}{n-q_n+1-i} - \Big[\ln\Big(\frac{1}{2}(n-q_{2,n}+1-i)\Big) - \frac{1}{n-q_{2,n}+1-i} \Big] \Big\} \\
&\quad + o(1)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=n-q_n+1-p}^{n-q_n} \left\{ \ln\Big(\frac{i}{2}\Big) - \frac{1}{i} \right\} - \sum_{i=n-q_{2,n}+1-p}^{n-q_{2,n}} \left\{ \ln\Big(\frac{i}{2}\Big) - \frac{1}{i} \right\} + o(1) \\
&= \ln\Big(\frac{(n-q_n)!}{(n-q_n-p)!}\Big) - \ln\Big(\frac{n-q_n}{n-q_n-p}\Big) - \ln\Big(\frac{(n-q_{2,n})!}{(n-q_{2,n}-p)!}\Big) + \ln\Big(\frac{n-q_{2,n}}{n-q_{2,n}-p}\Big) + o(1) \\
&= \ln\Big(\frac{(n-q_n-1)!}{(n-q_n-p-1)!}\Big) - \ln\Big(\frac{(n-q_{2,n}-1)!}{(n-q_{2,n}-p-1)!}\Big) + o(1) \\
&= (n-q_n-1)\ln(n-q_n-1) + (n-q_{2,n}-p-1)\ln(n-q_{2,n}-p-1) \\
&\quad - (n-q_n-p-1)\ln(n-q_n-p-1) - (n-q_{2,n}-1)\ln(n-q_{2,n}-1) + \frac{\sigma^2}{4} + o(1).
\end{aligned}
$$

Furthermore, we have

$$
\begin{aligned}
\sup_{1\le i\le p} \mathrm{Var}(\ln(V_i)) &= \sup_{1\le i\le p} \left\{ \psi_1\Big(\frac{1}{2}(n-q_n+1-i)\Big) - \psi_1\Big(\frac{1}{2}(n-q_{2,n}+1-i)\Big) \right\} \\
&\le \psi_1\Big(\frac{1}{2}(n-q_n+1-p)\Big) = \mathcal{O}\Big(\frac{2}{n-q_n+1-p}\Big) = o(1),
\end{aligned}
$$

which is condition (iv). Obviously, (i) and (ii) are also satisfied. The inequality for the moments in (iii) follows from Lemma A.7 and Theorem A.8 in [9]. Therefore, all conditions (i)–(v) are satisfied and the assertion follows from Theorem 5.

CRediT authorship contribution statement

Holger Dette: Writing - original draft. **Nina Dörnemann:** Writing - original draft.

Acknowledgments

The authors would like to thank M. Stein who typed this manuscript with considerable technical expertise. The work of H. Dette and N. Dörnemann was partially supported by the Deutsche Forschungsgemeinschaft, Germany (DFG Research Unit 1735, DE 502/26-2, RTG 2131). The authors would like to thank the reviewers for very constructive comments on an earlier version of this paper.

References

[1] M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, ninth Dover printing, tenth GPO printing, Dover, New York, 1964.
[2] T.W. Anderson, An introduction to multivariate statistical analysis, second ed., in: Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, Inc., New York, 1984, p. xviii+675.
[3] Z. Bai, D. Jiang, J.-F. Yao, S. Zheng, Corrections to LRT on large-dimensional covariance matrix by RMT, Ann. Statist. 37 (2009) 3822–3840.
[4] Z. Bai, D. Jiang, J.-F. Yao, S. Zheng, Testing linear hypotheses in high-dimensional regressions, J. Theoret. Appl. Statist. 47 (2013) 1207–1223.
[5] Z. Bao, J. Hu, G. Pan, W. Zhou, Test of independence for high-dimensional random vectors based on freeness in block correlation matrices, Electron. J. Stat. 11 (1) (2017) 1527–1548.
[6] M.S. Bartlett, Properties of sufficiency and statistical tests, Proc. R. Soc. Lond. Ser. A-Math. Phys. Sci. 160 (901) (1937) 268–282.
[7] T. Bodnar, H. Dette, N. Parolya, Testing for independence of large dimensional vectors, The Annals of Statistics 47 (5) (2019) 2977–3008.
[8] X. Chen, W. Liu, Testing independence with high-dimensional correlated samples, Ann. Statist. 46 (2) (2018) 866–894.
[9] H. Dette, D. Tomecki, Determinants of block hankel matrices for random matrix-valued measures, Stochastic Processes and their Applications 129 (12) (2019) 5200–5235.
[10] J. Fan, R. Li, Statistical challenges with high dimensionality: Feature selection in knowledge discovery, in: Proceedings of the International Congress of Mathematicians, Vol. 3, Madrid, 2006.
[11] K.B. Gregory, R.J. Caroll, V. Baladandayuthapani, S.N. Lahiri, A two-sample test for equality of means in high dimension, J. Amer. Statist. Assoc. 110 (2015) 837–849.
[12] A.M. Hanea, G.F. Nane, The asymptotic distribution of the determinant of a random correlation matrix, Stat. Neerl. 72 (1) (2018) 14–33.
[13] M. Hyodo, N. Shutoh, T. Nishiyama, T. Pavlenko, Testing block-diagonal covariance structure for high-dimensional data, Stat. Neerl. 69 (4) (2015) 460–482.
[14] T. Jiang, Determinant of sample correlation matrix with application, Ann. Appl. Probab. 29 (3) (2019) 1356–1397.
[15] D. Jiang, Z. Bai, S. Zheng, Testing the independence of sets of large-dimensional variables, Sci. China Math. 56 (1) (2013) 135–147.
[16] T. Jiang, Y. Qi, Likelihood ratio tests for high-dimensional normal distributions, Scand. J. Stat. 42 (4) (2015) 988–1009.
[17] T. Jiang, F. Yang, Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions, Ann. Statist. 41 (2013) 2029–2074.
[18] I.M. Johnstone, High dimensional statistical inference and random matrices, in: Proceedings of the International Congress of Mathematicians, Madrid, 2006.
[19] O. Kallenberg, Foundations of Modern Probability, Probability and its Applications, Springer New York, 1997.
[20] W. Li, Y. Qin, Hypothesis testing for high-dimensional covariance matrices, J. Multivariate Anal. 128 (2014) 108–119.
[21] X. Mestre, P. Vallet, Correlation tests and linear spectral statistics of the sample correlation matrix, IEEE Trans. Inform. Theory 63 (7) (2017) 4585–4618.
[22] D. Morrison, Multivariate Statistical Methods, Cengage Learning, Duxbury Press, 2004.

[23] R.J. Muirhead, Aspects of Multivariate Statistical Theory, John Wiley & Sons, New York, 1982.
[24] P. O'Brien, Robust procedures for testing equality of covariance matrices, Biometrics 48 (3) (1992) 819–827.
[25] J.R. Schott, A test for the equality of covariance matrices when the dimension is large relative to the sample sizes, Comput. Statist. Data Anal. 51 (2007) 6535–6542.
[26] M.S. Srivastava, H. Yanagihara, Testing the equality of several covariance matrices with fewer observations than the dimension, J. Multivariate Anal. 101 (2010) 1319–1329.
[27] N. Sugiura, Y. Fujikoshi, Asymptotic expansions of the non-null distribution of the likelihood ratio criteria for multivariate linear hypothesis and independence, Ann. Math. Stat. 40 (3) (1969) 942–952.
[28] S.S. Wilks, Certain generalizations in the analysis of variance, Biometrika 24 (1932) 471–494.
[29] S.S. Wilks, On the independence of k sets of normally distributed statistical variables, Econometrica 3 (1935) 309–326.
[30] S.S. Wilks, Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution, Ann. Math. Stat. 17 (3) (1946) 257–281.
[31] Y. Yamada, M. Hyodo, N. Shutoh, T. Nishiyama, Testing block-diagonal covariance structure for high-dimensional data under non-normality, J. Multivariate Anal. 155 (2017) 305–316.
[32] J. Yao, Z. Bai, S. Zheng, Large Sample Covariance Matrices and High-Dimensional Data Analysis (No. 39), Cambridge University Press, New York, 2015.