Statistica Sinica **31** (2021), 723-747
doi:https://doi.org/10.5705/ss.202018.0279

LEAST FAVORABLE DIRECTION TEST FOR MULTIVARIATE ANALYSIS OF VARIANCE IN HIGH DIMENSION

Rui Wang and Xingzhong Xu

Beijing Institute of Technology

Abstract: This study considers multivariate analysis of variance for normal samples in a high-dimensional medium sample size setting. When the sample dimension is larger than the sample size, the classical likelihood ratio test is not defined, because the likelihood function is unbounded. Based on this unboundedness, we propose a new test called the least favorable direction test. The asymptotic distributions of the test statistic are derived under both nonspiked and spiked covariances. The local asymptotic power function of the test is also given. The results for the asymptotic power function and simulations show that the proposed test is particularly powerful under the spiked covariance.

Key words and phrases: High-dimensional data, least favorable direction test, multivariate analysis of variance, principal component analysis, spiked covariance.

1. Introduction

Suppose there are k ($k \geq 2$) independent samples of p-dimensional data. Within the ith sample ($1 \leq i \leq k$), the observations $\{X_{ij}\}_{j=1}^{n_i}$ are independent and identically distributed (i.i.d.) as $\mathcal{N}_p(\theta_i, \mathbf{\Sigma})$, which is a p-dimensional normal distribution with mean vector θ_i and common variance matrix $\mathbf{\Sigma}$.

We test the following hypotheses:

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_k \quad \text{vs.} \quad H_1 : \theta_i \neq \theta_j, \text{ for some } i \neq j. \tag{1.1}$$

This testing problem is known as the one-way multivariate analysis of variance (MANOVA), and has been well studied when p is small relative to N, where $N = \sum_{i=1}^{k} n_i$ is the total sample size.

Let $\mathbf{H} = \sum_{i=1}^{k} n_i(\bar{\mathbf{X}}_i - \bar{\mathbf{X}})(\bar{\mathbf{X}}_i - \bar{\mathbf{X}})^\top$ be the sum-of-squares between groups, and let $\mathbf{G} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{\mathbf{X}}_i)(X_{ij} - \bar{\mathbf{X}}_i)^\top$ be the sum-of-squares within groups, where $\bar{\mathbf{X}}_i = n_i^{-1}\sum_{j=1}^{n_i} X_{ij}$ is the sample mean of group i, and $\bar{\mathbf{X}} =$

Corresponding author: Xingzhong Xu, Beijing Key Laboratory on MCAACI, Beijing Institute of Technology, Beijing 100081, China. E-mail: xuxz@bit.edu.cn.

Table 1

Wilks' Lambda:	$\lvert\mathbf{G}+\mathbf{H}\rvert/\lvert\mathbf{G}\rvert$
Pillai trace:	$\mathrm{tr}[\mathbf{H}(\mathbf{G}+\mathbf{H})^{-1}]$
Hotelling–Lawley trace:	$\mathrm{tr}[\mathbf{H}\mathbf{G}^{-1}]$
Roy's maximum root:	$\lambda_1(\mathbf{H}\mathbf{G}^{-1})$

$N^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}$ is the pooled sample mean. There are four classical test statistics for hypotheses (1.1), all of which are based on the eigenvalues of $\mathbf{H}\mathbf{G}^{-1}$.

In some modern scientific applications, researchers would like to test hypotheses (1.1) in the high-dimensional setting, that is, where p is greater than N; see, for example, Verstynen et al. (2005) and Tsai and Chen (2009). However, none of the four classical test statistics are defined when $p \geq N$. As a result, extensive research has been done on the testing problem (1.1) in high-dimensional settings. Thus far, numerous tests have been proposed for the case $k = 2$; see, for example, Bai and Saranadasa (1996), Srivastava (2007), Chen and Qin (2010), Cai, Liu and Xia (2014), and Feng et al. (2015). Tests have also been proposed for the general case of $k \geq 2$. Schott (2007) modified the Hotelling–Lawley trace and proposed the following test statistic:

$$T_{Sc} = \frac{1}{\sqrt{N-1}}\left(\frac{1}{k-1}\,\mathrm{tr}\left(\mathbf{H}\right) - \frac{1}{N-k}\,\mathrm{tr}\left(\mathbf{G}\right)\right).$$

Here, T_{Sc} is a member of the so-called sum-of-squares statistics, because it is based on an estimation of the squared Euclidean norm $\sum_{i=1}^{k} n_i\|\theta_i - \bar{\theta}\|^2$, where $\bar{\theta} = N^{-1}\sum_{i=1}^{k} n_i\theta_i$. See Srivastava and Kubokawa (2013), Yamada and Himeno (2015), Hu et al. (2017), Zhang, Guo and Zhou (2017), Zhou, Guo and Zhang (2017), and Cao, Park and He (2019) for other sum-of-squares test statistics for $k \geq 2$. Sum-of-squares tests are known to be particularly powerful in the case of dense alternatives. In another work, Cai and Xia (2014) proposed the test statistic

$$T_{CX} = \max_{1\leq i\leq p} \sum_{1\leq j<l\leq k} \frac{n_j n_l}{n_j + n_l} \frac{(\Omega(\bar{\mathbf{X}}_j - \bar{\mathbf{X}}_l))_i^2}{\omega_{ii}},$$

where $\Omega = (\omega)_{ij} = \mathbf{\Sigma}^{-1}$ is the precision matrix. When Ω is unknown, it is substituted by an estimator. Unlike T_{Sc}, T_{CX} is an extreme-value test statistic, and is powerful in the case of sparse alternatives.

Most existing sum-of-squares test procedures require the condition $\mathrm{tr}(\mathbf{\Sigma}^4)/\mathrm{tr}^2(\mathbf{\Sigma}^2) \to 0$, which is equivalent to

$$\frac{\boldsymbol{\lambda}_1}{\sqrt{\operatorname{tr}(\boldsymbol{\Sigma}^2)}} \to 0, \tag{1.2}$$

where $\boldsymbol{\lambda}_i$ is the ith largest eigenvalue of $\boldsymbol{\Sigma}$, for $i = 1, \ldots, p$. In fact, the equivalence of these two conditions can be seen from the following inequalities:

$$\frac{\boldsymbol{\lambda}_1^4}{\operatorname{tr}^2(\boldsymbol{\Sigma}^2)} \leq \frac{\operatorname{tr}(\boldsymbol{\Sigma}^4)}{\operatorname{tr}^2(\boldsymbol{\Sigma}^2)} \leq \frac{\boldsymbol{\lambda}_1^2 \operatorname{tr}(\boldsymbol{\Sigma}^2)}{\operatorname{tr}^2(\boldsymbol{\Sigma}^2)} = \frac{\boldsymbol{\lambda}_1^2}{\operatorname{tr}(\boldsymbol{\Sigma}^2)}.$$

Condition (1.2) is reasonable if $\boldsymbol{\Sigma}$ is nonspiked, in the sense that it does not have significantly large eigenvalues. However, in practice, variables may be heavily correlated with common factors, in which case, the covariance matrix $\boldsymbol{\Sigma}$ is spiked, in the sense that a few eigenvalues of $\boldsymbol{\Sigma}$ are significantly larger than the others (Fan, Yuan and Mincheva (2013); Cai, Ma and Wu (2015); Wang and Fan (2017)). In such cases, condition (1.2) can be violated and, consequently, existing sum-of-squares tests may not have the correct level. Adjusted sum-of-squares test procedures have been proposed to solve this problem; see, for example, Katayama, Kano and Srivastava (2013), Ma, Lan and Wang (2015), Zhang, Guo and Zhou (2017), and Wang and Xu (2019). However, the power behavior of these corrected tests may not be satisfactory.

Recently, Aoshima and Yata (2018) and Wang and Xu (2018) considered a two-sample mean testing problem under the spiked covariance model. These tests have better power behavior than that of sum-of-squares tests. However, both studies imposed strong conditions on the magnitude of p. For example, under the approximate factor model in Fan, Yuan and Mincheva (2013), the test in Aoshima and Yata (2018) requires $p/N \to 0$, whereas the test in Wang and Xu (2018) requires that $p/N^2 \to 0$ and that the small eigenvalues of $\boldsymbol{\Sigma}$ are all equal.

The likelihood ratio test (LRT) method has been very successful in leading to satisfactory procedures in many specific problems. However, the LRT statistic for hypotheses (1.1), that is, Wilks' Lambda statistic, is not defined for $p > N-k$. In the high-dimensional setting, neither the sum-of-squares nor the extreme-value statistics are based on the likelihood function. This motivates us to construct a likelihood-based test in the high-dimensional setting. In a recent work, Zhao and Xu (2016) proposed a generalized likelihood ratio test in the context of the one-sample mean vector test. They used a least favorable argument to construct a generalized likelihood ratio test statistic. Their simulation results showed that their test exhibits good power performance, especially when the variables are correlated. However, they do not provide a theoretical proof.

We propose a generalized likelihood ratio test statistic for hypotheses (1.1), called the least favorable direction (LFD) test statistic, which is a generalization of the test in Zhao and Xu (2016). We give the asymptotic distributions of the test statistic under both nonspiked and spiked covariances. An adaptive LFD test procedure is constructed by consistently detecting the unknown covariance structure and estimating the unknown parameters. The asymptotic local power function of the LFD test is also given. Our theoretical results show that the LFD test is particularly powerful under the spiked covariance. This explains the simulation results of Zhao and Xu (2016). Extending the work of Zhao and Xu (2016), our main contribution is that we provide a thorough theoretical analysis of the LFD test. This analysis falls within the high-dimensional medium sample size setting, where both $N, p \to \infty$, but $p/N \to \infty$ (see Aoshima et al. (2018), Sec. 5). To prove our main results, we carefully study the high-order asymptotic behavior of the eigenvalues and eigenspaces of the sample covariance matrix. These results are also of independent interest. We further compare the proposed test procedure with existing tests using simulations. Here, we show that the LFD test exhibits behavior comparable with that of existing sum-of-squares tests under the nonspiked covariance, while significantly outperforming competing tests under the spiked covariance.

The rest of the paper is organized as follows. In Section 2, we propose the LFD test statistic and derive its explicit forms. The asymptotic distributions of the LFD test statistic under nonspiked and spiked covariances are given in Section 3. Based on these theoretical results, an adaptive LFD test procedure is proposed. Section 4 complements our study with numerical simulations. Section 5 concludes the paper. Finally, the proofs are gathered in the Supplementary Material.

2. Least Favorable Direction Test

We first introduce some necessary notation. Define the $p \times N$ pooled sample matrix $\mathbf{X}$ as

$$\mathbf{X} = (X_{11}, X_{12}, \ldots, X_{1n_1}, X_{21}, X_{22}, \ldots, X_{2n_2}, \ldots, X_{k1}, X_{k2}, \ldots, X_{kn_k}).$$

The sum-of-squares within groups $\mathbf{G}$ can be written as $\mathbf{G} = \mathbf{X}(\mathbf{I}_N - \mathbf{J}\mathbf{J}^\top)\mathbf{X}^\top$, where

$$\mathbf{J} = \begin{pmatrix} \frac{1}{\sqrt{n_1}}\mathbf{1}_{n_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\sqrt{n_2}}\mathbf{1}_{n_2} & \mathbf{0} \\ \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \frac{1}{\sqrt{n_k}}\mathbf{1}_{n_k} \end{pmatrix}$$

is an $N \times k$ matrix, and $\mathbf{1}_{n_i}$ is an n_i-dimensional vector with all elements equal to one, for $i = 1, \dots, k$. Let $n = N - k$ be the degrees of freedom of $\mathbf{G}$. Construct an $N \times n$ matrix $\tilde{\mathbf{J}}$ as

$$\tilde{\mathbf{J}} = \begin{pmatrix} \tilde{\mathbf{J}}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{J}}_2 & \mathbf{0} \\ \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \tilde{\mathbf{J}}_k \end{pmatrix},$$

where $\tilde{\mathbf{J}}_i$ is an $n_i \times (n_i - 1)$ matrix defined as

$$\tilde{\mathbf{J}}_i = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \cdots & \frac{1}{\sqrt{(n_i-2)(n_i-1)}} & \frac{1}{\sqrt{(n_i-1)n_i}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \cdots & \frac{1}{\sqrt{(n_i-2)(n_i-1)}} & \frac{1}{\sqrt{(n_i-1)n_i}} \\ 0 & -\frac{2}{\sqrt{6}} & \cdots & \vdots & \vdots \\ \vdots & \vdots & \cdots & -\frac{n_i-2}{\sqrt{(n_i-2)(n_i-1)}} & \frac{1}{\sqrt{(n_i-1)n_i}} \\ 0 & 0 & \cdots & 0 & -\frac{n_i-1}{\sqrt{(n_i-1)n_i}} \end{pmatrix}.$$

The matrix $\tilde{\mathbf{J}}$ is a column orthogonal matrix satisfying $\tilde{\mathbf{J}}^\top\tilde{\mathbf{J}} = \mathbf{I}_n$ and $\tilde{\mathbf{J}}\tilde{\mathbf{J}}^\top = \mathbf{I}_N - \mathbf{J}\mathbf{J}^\top$. Define $\mathbf{Y} = \mathbf{X}\tilde{\mathbf{J}}$. Then, $\mathbf{G}$ can be written as

$$\mathbf{G} = \mathbf{Y}\mathbf{Y}^\top.$$

The sum-of-squares between groups $\mathbf{H}$ can be written as

$$\mathbf{H} = \mathbf{X}\left(\mathbf{J}\mathbf{J}^\top - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top\right)\mathbf{X}^\top = \mathbf{X}\mathbf{J}\left(\mathbf{I}_k - \frac{1}{N}\mathbf{J}^\top\mathbf{1}_N\mathbf{1}_N^\top\mathbf{J}\right)\mathbf{J}^\top\mathbf{X}^\top.$$

By some matrix algebra, we have $\mathbf{I}_k - N^{-1}\mathbf{J}^\top\mathbf{1}_N\mathbf{1}_N^\top\mathbf{J} = \mathbf{C}\mathbf{C}^\top$, where $\mathbf{C}$ is a

$k \times (k-1)$ matrix defined as $\mathbf{C} = \mathbf{C}_1\mathbf{C}_2$, and

$$\mathbf{C}_1 = \begin{pmatrix} \sqrt{n_1} & \sqrt{n_1} & \cdots & \sqrt{n_1} & \sqrt{n_1} \\ -\frac{n_1}{\sqrt{n_2}} & \sqrt{n_2} & \cdots & \sqrt{n_2} & \sqrt{n_2} \\ 0 & -\frac{n_1+n_2}{\sqrt{n_3}} & \cdots & \vdots & \vdots \\ \vdots & \vdots & \cdots & -\frac{\sum_{i=1}^{k-2} n_i}{\sqrt{n_{k-1}}} & \sqrt{n_{k-1}} \\ 0 & 0 & \cdots & 0 & -\frac{\sum_{i=1}^{k-1} n_i}{\sqrt{n_k}} \end{pmatrix},$$

$$\mathbf{C}_2 = \begin{pmatrix} \frac{n_1(n_1+n_2)}{n_2} & 0 & \cdots & 0 \\ 0 & \frac{(\sum_{i=1}^{2} n_i)(\sum_{i=1}^{3} n_i)}{n_3} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \frac{(\sum_{i=1}^{k-1} n_i)(\sum_{i=1}^{k} n_i)}{n_k} \end{pmatrix}^{-1/2}.$$

Then, $\mathbf{H}$ can be written as

$$\mathbf{H} = \mathbf{XJCC}^\top\mathbf{J}^\top\mathbf{X}^\top.$$

Define $\boldsymbol{\Theta} = (\sqrt{n_1}\theta_1, \ldots, \sqrt{n_k}\theta_k)$. Then, the null hypothesis H_0 is equivalent to $\boldsymbol{\Theta}\mathbf{C} = \mathbf{O}_{p\times(k-1)}$, where $\mathbf{O}_{p\times(k-1)}$ is a $p \times (k-1)$ matrix with all entries zero. Thus, the hypotheses (1.1) are equivalent to

$$H_0 : \boldsymbol{\Theta}\mathbf{C} = \mathbf{O}_{p\times(k-1)} \quad \text{vs.} \quad H_1 : \boldsymbol{\Theta}\mathbf{C} \neq \mathbf{O}_{p\times(k-1)}.$$

The testing problem (1.1) is well studied for low-dimensional settings. A classical test statistic is Roy's maximum root, constructed by Roy (1953) using his well-known union intersection principle. The key idea is to decompose $\mathbf{X}$ into a set of univariate data $\{\mathbf{X}_a = a^\top\mathbf{X} : a \in \mathbb{R}^p, a^\top a = 1\}$. This induces the following decompositions of the null and alternative hypotheses:

$$H_0 = \bigcap_{a\in\mathbb{R}^p, a^\top a=1} H_{0a} \quad \text{vs.} \quad H_1 = \bigcup_{a\in\mathbb{R}^p, a^\top a=1} H_{1a},$$

where $H_{0a} : a^\top\boldsymbol{\Theta}\mathbf{C} = \mathbf{O}_{1\times(k-1)}$ and $H_{1a} : a^\top\boldsymbol{\Theta}\mathbf{C} \neq \mathbf{O}_{1\times(k-1)}$. Let $L_0(a)$ and $L_1(a)$ be the maximum likelihood of $\mathbf{X}_a$ under H_{0a} and H_{1a}, respectively. For each a satisfying $a^\top a = 1$, the component LRT statistic

$$\frac{L_1(a)}{L_0(a)} = \Big(\frac{a^\top(\mathbf{G}+\mathbf{H})a}{a^\top\mathbf{G}a}\Big)^{N/2}$$

can be used to test H_{0a} versus H_{1a}. Using the union intersection principle, Roy proposed the test statistic $\max_{a^\top a=1} L_1(a)/L_0(a) = (1+\lambda_1(\mathbf{H}\mathbf{G}^{-1}))^{N/2}$, where $\lambda_i(\cdot)$ denotes the ith largest eigenvalue. This statistic is an increasing function of Roy's maximum root.

From a likelihood point of view, the log likelihood ratio is an estimator of the Kullback–Leibler divergence between the true distribution and the null distribution. Hence, the component LRT statistic $L_1(a)/L_0(a)$ characterizes the discrepancy between the true and the null distribution along the direction a. This motivates us to consider the direction

$$a^* = \operatorname*{argmax}_{a^\top a=1} \frac{L_1(a)}{L_0(a)}, \tag{2.1}$$

which hopefully yields the largest discrepancy between the true and the null distribution. Thus, H_{0a^*} is the component null hypothesis least likely to be true. We call a^* the least favorable direction. Note that Roy's maximum root is the component LRT statistic along the least favorable direction.

Unfortunately, Roy's maximum root can only be defined when $n \geq p$, and hence cannot be used in the high-dimensional setting. In what follows, we assume $p > n$. In this case, the set

$$\mathcal{A} \overset{def}{=} \{a : L_1(a) = +\infty,\, a^\top a = 1\} = \{a : a^\top \mathbf{G} a = 0,\, a^\top a = 1\}$$

is not empty because $\mathbf{G}$ is singular. Consequently, the right-hand side of (2.1) is not well defined because the ratio involves infinity. Hence, we need a new definition for the LFD in the high-dimensional setting. Define

$$\mathcal{B} = \{a : L_0(a) = +\infty,\, a^\top a = 1\} = \{a : a^\top(\mathbf{G}+\mathbf{H})a = 0,\, a^\top a = 1\}.$$

Note that $\mathcal{B} \subset \mathcal{A}$. Moreover, by the independence of $\mathbf{G}$ and $\mathbf{H}$, with probability one, we have $\mathcal{A} \cap \mathcal{B}^c \neq \emptyset$. Then, for any direction a, there are three possible scenarios: $L_1(a) < +\infty$ and $L_0(a) < +\infty$; $L_1(a) = +\infty$ and $L_0(a) < +\infty$; and $L_1(a) = +\infty$ and $L_0(a) = +\infty$. To maximize the discrepancy between $L_1(a)$ and $L_0(a)$, one may consider the direction a such that $L_1(a) = +\infty$ and $L_0(a) < +\infty$. This suggests that the least favorable direction a^*, which hopefully maximizes the discrepancy between $L_1(a)$ and $L_0(a)$, should be defined as $a^* = \operatorname{argmin}_{a \in \mathcal{A}\cap\mathcal{B}^c} L_0(a)$. Equivalently,

$$a^* = \operatorname*{argmin}_{a \in \mathcal{A}\cap\mathcal{B}^c} L_0(a) = \operatorname*{argmax}_{a^\top a=1, a^\top G a=0} a^\top \mathbf{H} a.$$

Based on a^* and the likelihood $L_0(a)$, we propose a new test statistic,

$$T(\mathbf{X}) = a^{*T}\mathbf{H}a^* = \max_{a^\top a=1, a^\top \mathbf{G}a=0} a^\top \mathbf{H}a.$$

The null hypothesis is rejected when $T(\mathbf{X})$ is sufficiently large. We call $T(\mathbf{X})$ the LFD test statistic. Because the least favorable direction a^* is obtained from the component likelihood function, the statistic $T(\mathbf{X})$ is also a generalized likelihood ratio test statistic.

Now, we derive the explicit forms of the LFD test statistic. Let $\mathbf{Y} = \mathbf{U_Y D_Y V_Y^\top}$ be the singular value decomposition of $\mathbf{Y}$, where $\mathbf{U_Y}$ and $\mathbf{V_Y}$ are $p \times \min(n, p)$ and $n \times \min(n, p)$ column orthogonal matrices, respectively, and $\mathbf{D_Y}$ is a $\min(n, p) \times \min(n, p)$ diagonal matrix, with diagonal elements comprising the non-increasingly ordered singular values of $\mathbf{Y}$. If $p > n$, let $\mathbf{P_Y} = \mathbf{U_Y U_Y^\top}$ be the projection matrix onto the column space of $\mathbf{Y}$. Then, Lemma 1 in the Supplementary Material implies that, for $p > n$,

$$T(\mathbf{X}) = \lambda_1\big(\mathbf{C}^\top \mathbf{J}^\top \mathbf{X}^\top (\mathbf{I}_p - \mathbf{P_Y})\mathbf{XJC}\big). \tag{2.2}$$

Although (2.2) is convenient for the theoretical analysis, it is not convenient for computation. When $p > N$, another simple form of $T(\mathbf{X})$ can be used for computation. If $p > N$, then $\mathbf{X}^\top\mathbf{X}$ is invertible. By the relationship

$$\begin{aligned}\begin{pmatrix} \mathbf{J}^\top\mathbf{X}^\top\mathbf{XJ} & \mathbf{J}^\top\mathbf{X}^\top\mathbf{X}\tilde{\mathbf{J}} \\ \tilde{\mathbf{J}}^\top\mathbf{X}^\top\mathbf{XJ} & \tilde{\mathbf{J}}^\top\mathbf{X}^\top\mathbf{X}\tilde{\mathbf{J}} \end{pmatrix}^{-1} &= \left(\begin{pmatrix} \mathbf{J}^\top \\ \tilde{\mathbf{J}}^\top \end{pmatrix} \mathbf{X}^\top\mathbf{X} \begin{pmatrix} \mathbf{J} & \tilde{\mathbf{J}} \end{pmatrix} \right)^{-1} \\ &= \begin{pmatrix} \mathbf{J}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{J} & \mathbf{J}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\tilde{\mathbf{J}} \\ \tilde{\mathbf{J}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{J} & \tilde{\mathbf{J}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\tilde{\mathbf{J}} \end{pmatrix}\end{aligned}$$

and the matrix inverse formula, we have that

$$\begin{aligned}\big(\mathbf{J}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{J}\big)^{-1} =& \mathbf{J}^\top\mathbf{X}^\top\mathbf{XJ} - \mathbf{J}^\top\mathbf{X}^\top\mathbf{X}\tilde{\mathbf{J}}(\tilde{\mathbf{J}}^\top\mathbf{X}^\top\mathbf{X}\tilde{\mathbf{J}})^{-1}\tilde{\mathbf{J}}^\top\mathbf{X}^\top\mathbf{XJ} \\ =& \mathbf{J}^\top\mathbf{X}^\top(\mathbf{I}_p - \mathbf{P_Y})\mathbf{XJ}.\end{aligned}$$

Thus,

$$T(\mathbf{X}) = \lambda_1\Big(\mathbf{C}^\top\big(\mathbf{J}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{J}\big)^{-1}\mathbf{C}\Big). \tag{2.3}$$

Compared with (2.2), the expression in (2.3) does not involve $\mathbf{P_Y}$ and, thus, is more convenient for computation.

In the case of $k = 2$, it can be seen that the least favorable direction is proportional to $(\mathbf{I}_p - \mathbf{P}_{\mathbf{Y}})(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$, and the LFD test statistic has expression

$$T(\mathbf{X}) = \frac{n_1 n_2}{n_1 + n_2} \|(\mathbf{I}_p - \mathbf{P}_{\mathbf{Y}})(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)\|^2.$$

In this case, the least favorable direction coincides with the maximal data piling direction proposed by Ahn and Marron (2010).

3. Theoretical Analysis

We now analyze the asymptotic distributions of the LFD test statistic. The normality of the observations is an important assumption for our results, and is assumed throughout this section. We present theoretical results under both nonspiked and spiked covariances. Based on these results, we construct an adaptive test with an asymptotically correct level. In addition, these results allow us to derive the local asymptotic power function of the LFD test.

3.1. Nonspiked covariance

In this subsection, we establish the asymptotic distribution of $T(\mathbf{X})$ under the nonspiked covariance. Let $\mathbf{W}_{k-1}$ be a $(k-1) \times (k-1)$ symmetric random matrix in which the entries above the main diagonal are i.i.d. $\mathcal{N}(0,1)$ random variables, and the entries on the diagonal are i.i.d. $\mathcal{N}(0,2)$ random variables. The following theorem establishes the asymptotic distribution of the LFD test statistic.

Theorem 1. *Suppose as $n, p \to \infty$, condition* (1.2) *holds. Furthermore, suppose $n\boldsymbol{\lambda}_1/\operatorname{tr}(\boldsymbol{\Sigma}) \to 0$ and $\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_p = O(n^{-1}\sqrt{\operatorname{tr}(\boldsymbol{\Sigma}^2)})$. Then, under the local alternative hypothesis $\|\mathbf{C}^\top \boldsymbol{\Theta}^\top \boldsymbol{\Theta} \mathbf{C}\| = O(\sqrt{\operatorname{tr}(\boldsymbol{\Sigma}^2)})$,*

$$\frac{T(\mathbf{X}) - \big(\operatorname{tr}(\boldsymbol{\Sigma}) - n\operatorname{tr}(\boldsymbol{\Sigma}^2)/\operatorname{tr}(\boldsymbol{\Sigma})\big)}{\sqrt{\operatorname{tr}(\boldsymbol{\Sigma}^2)}} \sim \lambda_1 \left(\mathbf{W}_{k-1} + \frac{\mathbf{C}^\top \boldsymbol{\Theta}^\top \boldsymbol{\Theta} \mathbf{C}}{\sqrt{\operatorname{tr}(\boldsymbol{\Sigma}^2)}} \right) + o_P(1),$$

where $\sim$ means having the same distribution.

Remark 1. The condition $n\boldsymbol{\lambda}_1/\operatorname{tr}(\boldsymbol{\Sigma}) \to 0$ implies $p/n \to \infty$. Hence, $T(\mathbf{X})$ is well defined for large n. The condition $\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_p = O(n^{-1}\sqrt{\operatorname{tr}(\boldsymbol{\Sigma}^2)})$ requires that the range of the eigenvalues of $\boldsymbol{\Sigma}$ not be too large.

To centralize $T(\mathbf{X})$ under the conditions of Theorem 1, we need to estimate the parameters $\operatorname{tr}(\boldsymbol{\Sigma})$ and $\operatorname{tr}(\boldsymbol{\Sigma}^2)$. Let $\hat{\boldsymbol{\Sigma}} = n^{-1}\mathbf{G} = n^{-1}\mathbf{Y}\mathbf{Y}^\top$ be the sample covariance matrix. We use the following simple estimators:

$$\widehat{\operatorname{tr}(\boldsymbol{\Sigma})} = \operatorname{tr}(\hat{\boldsymbol{\Sigma}}), \quad \widehat{\operatorname{tr}(\boldsymbol{\Sigma}^2)} = \operatorname{tr}(\hat{\boldsymbol{\Sigma}}^2) - n^{-1} \operatorname{tr}^2(\hat{\boldsymbol{\Sigma}}).$$

Define

$$Q_1 = \frac{T(\mathbf{X}) - \left(\widehat{\operatorname{tr}(\boldsymbol{\Sigma})} - n\widehat{\operatorname{tr}(\boldsymbol{\Sigma}^2)}/\widehat{\operatorname{tr}(\boldsymbol{\Sigma})}\right)}{\sqrt{\widehat{\operatorname{tr}(\boldsymbol{\Sigma}^2)}}}.$$

Let $F_1(x)$ be the cumulative distribution function of $\lambda_1(\mathbf{W}_{k-1})$. Then, we reject the null hypothesis if $Q_1 > F_1^{-1}(1-\alpha)$. The following corollary gives the asymptotic local power function of the proposed test under the nonspiked covariance.

Corollary 1. *Under the conditions of Theorem* 1*,*

$$\Pr\left(Q_1 > F_1^{-1}(1-\alpha)\right) = \Pr\left(\lambda_1\left(\mathbf{W}_{k-1} + \frac{\mathbf{C}^\top\boldsymbol{\Theta}^\top\boldsymbol{\Theta}\mathbf{C}}{\sqrt{\operatorname{tr}(\boldsymbol{\Sigma}^2)}}\right) > F_1^{-1}(1-\alpha)\right) + o(1).$$

Corollary 1 shows that under the nonspiked covariance, the LFD test exhibits power behavior similar to that of existing sum-of-squares tests. In fact, if $k = 2$, the asymptotic local power function given by Corollary 1 is equal to the asymptotic local power function of the tests in Bai and Saranadasa (1996) and Chen and Qin (2010).

3.2. Spiked covariance

Now, we derive the asymptotic results under the spiked covariance, which is more involved than the nonspiked case. Let $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ denote the eigenvalue decomposition of $\boldsymbol{\Sigma}$, where $\boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p)$ and $\mathbf{U}$ is an orthogonal matrix. Suppose that $\boldsymbol{\Sigma}$ has r spiked eigenvalues, where $1 \le r \le p$ can also vary as $n, p \to \infty$. We first assume the spiked number r is known. We latter consider the adaptation to unknown r. Denote $\boldsymbol{\Lambda}_1 = \operatorname{diag}(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_r)$ and $\boldsymbol{\Lambda}_2 = \operatorname{diag}(\boldsymbol{\lambda}_{r+1}, \ldots, \boldsymbol{\lambda}_p)$. Correspondingly, we denote $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2)$, where $\mathbf{U}_1$ and $\mathbf{U}_2$ are the first r columns and the last $p - r$ columns, respectively, of $\mathbf{U}$. Then, $\boldsymbol{\Sigma} = \mathbf{U}_1\boldsymbol{\Lambda}_1\mathbf{U}_1^\top + \mathbf{U}_2\boldsymbol{\Lambda}_2\mathbf{U}_2^\top$.

First, we derive the asymptotic properties of the eigenvalues and eigenspaces of the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$, because these play a key role in our later analysis. The following proposition gives the asymptotic behavior of $\lambda_1(\hat{\boldsymbol{\Sigma}}), \ldots, \lambda_r(\hat{\boldsymbol{\Sigma}})$ and $\sum_{i=r+1}^n \lambda_i(\hat{\boldsymbol{\Sigma}})$.

Proposition 1. *Suppose $r \le n$. Then, uniformly for $i = 1, \ldots, r$,*

$$\lambda_i(\hat{\mathbf{\Sigma}}) = \boldsymbol{\lambda}_i + n^{-1}\operatorname{tr}(\mathbf{\Lambda}_2) + O_P\left(\boldsymbol{\lambda}_i\sqrt{\frac{r}{n}} + \sqrt{\frac{\operatorname{tr}(\mathbf{\Lambda}_2^2)}{n}} + \boldsymbol{\lambda}_{r+1}\right)$$

and

$$\sum_{i=r+1}^{n} \lambda_i(\hat{\mathbf{\Sigma}}) = \left(1 - \frac{r}{n}\right)\operatorname{tr}(\mathbf{\Lambda}_2) + O_P\left(r\sqrt{\frac{\operatorname{tr}(\mathbf{\Lambda}_2^2)}{n}} + r\boldsymbol{\lambda}_{r+1}\right).$$

Remark 2. Recent works have examined the asymptotic behavior of the spiked eigenvalues of the sample covariance matrix; see, for example, Yata and Aoshima (2013), Shen, Shen and Marron (2016), Wang and Fan (2017), and Cai, Han and Pan (2019). An important improvement of Proposition 1 over existing results is that Proposition 1 does not impose any conditions on the structure of $\mathbf{\Sigma}$, but still gives the correct convergence rate.

Based on Proposition 1, we propose the following estimators of $\operatorname{tr}(\mathbf{\Lambda}_2)$ and $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_r$:

$$\widehat{\operatorname{tr}(\mathbf{\Lambda}_2)} = \left(1 - \frac{r}{n}\right)^{-1} \sum_{i=r+1}^{n} \lambda_i(\hat{\mathbf{\Sigma}}), \quad \hat{\boldsymbol{\lambda}}_i = \lambda_i(\hat{\mathbf{\Sigma}}) - n^{-1}\widehat{\operatorname{tr}(\mathbf{\Lambda}_2)}, \quad i = 1, \ldots, r.$$

Moreover, we propose the following estimator of $\operatorname{tr}(\mathbf{\Lambda}_2^2)$, which we use in our later analysis:

$$\widehat{\operatorname{tr}(\mathbf{\Lambda}_2^2)} = \sum_{i=r+1}^{n} \left(\lambda_i(\hat{\mathbf{\Sigma}}) - n^{-1}\widehat{\operatorname{tr}(\mathbf{\Lambda}_2)}\right)^2.$$

The following proposition gives the convergence rate of these estimators.

Proposition 2. *Suppose $r = o(n)$. Then, uniformly for $i = 1, \ldots, r$,*

$$\hat{\boldsymbol{\lambda}}_i = \boldsymbol{\lambda}_i + O_P\left(\boldsymbol{\lambda}_i\sqrt{\frac{r}{n}} + \sqrt{\frac{\operatorname{tr}(\mathbf{\Lambda}_2^2)}{n}} + \boldsymbol{\lambda}_{r+1}\right)$$

and

$$\begin{aligned}
\widehat{\operatorname{tr}(\mathbf{\Lambda}_2)} &= \operatorname{tr}(\mathbf{\Lambda}_2) + O_P\left(r\sqrt{\frac{\operatorname{tr}(\mathbf{\Lambda}_2^2)}{n}} + r\boldsymbol{\lambda}_{r+1}\right),\\
\widehat{\operatorname{tr}(\mathbf{\Lambda}_2^2)} &= \operatorname{tr}(\mathbf{\Lambda}_2^2) + O_P\left(\frac{r\operatorname{tr}(\mathbf{\Lambda}_2^2)}{n} + r\boldsymbol{\lambda}_{r+1}^2\right).
\end{aligned}$$

Remark 3. Our estimators of $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_r$ and $\operatorname{tr}(\mathbf{\Lambda}_2)$ are similar to some existing estimators, including the noise-reduction estimators of Yata and Aoshima (2012) and the estimators of Wang and Fan (2017). However, their theoretical results require that r is fixed, p is not large, and $\mathbf{\Sigma}$ satisfies certain spiked covariance models.

Remark 4. The estimation of $\mathrm{tr}(\mathbf{\Lambda}_2^2)$ is relatively unexplored. Recently, Aoshima and Yata (2018) proposed an estimator of $\mathrm{tr}(\mathbf{\Lambda}_2^2)$ based on the cross-data-matrix methodology. They also proved the consistency of their estimator. However, their method relies on an arbitrary split of the data into two samples of equal size.

Next, we consider the asymptotic behavior of the eigenspaces of $\hat{\mathbf{\Sigma}}$. Let $\mathbf{U}_{\mathbf{Y},1}$ denote the first r columns of $\mathbf{U}_{\mathbf{Y}}$. Then, the columns of $\mathbf{U}_{\mathbf{Y},1}$ are the principal eigenvectors of $\hat{\mathbf{\Sigma}}$, and $\mathbf{P}_{\mathbf{Y},1} = \mathbf{U}_{\mathbf{Y},1}\mathbf{U}_{\mathbf{Y},1}^\top$ is the projection matrix onto the rank r principal subspace of $\hat{\mathbf{\Sigma}}$. The properties of $\mathbf{P}_{\mathbf{Y},1}$ and the individual principal eigenvectors have been studied extensively. See Cai, Ma and Wu (2015), Shen, Shen and Marron (2016), and Wang and Fan (2017), and the references therein. Existing results include the consistency of the principal subspace and the high-order asymptotic behavior of the individual principal eigenvectors. However, these results are not sufficient for our analysis. The following proposition gives the high-order asymptotic behavior of $\mathbf{P}_{\mathbf{Y},1}$. To the best of our knowledge, this is a novel result in the literature.

Write $\mathbf{Y} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{Z}$, where $\mathbf{Z}$ is a $p \times n$ random matrix with i.i.d. $\mathcal{N}(0,1)$ entries. Then, $\mathbf{Y} = \mathbf{U}_1\mathbf{\Lambda}_1^{1/2}\mathbf{Z}_1 + \mathbf{U}_2\mathbf{\Lambda}_2^{1/2}\mathbf{Z}_2$, where $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are the first r rows and the last $p-r$ rows, respectively, of $\mathbf{Z}$.

Proposition 3. *Suppose $r = o(n)$, $\mathrm{tr}(\mathbf{\Lambda}_2)/(n\boldsymbol{\lambda}_r) \to 0$, and $r\boldsymbol{\lambda}_{r+1}/\mathrm{tr}(\mathbf{\Lambda}_2) \to 0$. Then,*

$$\left\|\mathbf{P}_{\mathbf{Y},1} - \mathbf{P}_{\mathbf{Y},1}^\dagger\right\| = O_P\left(\frac{\mathrm{tr}(\mathbf{\Lambda}_2)}{n\boldsymbol{\lambda}_r} + \frac{\boldsymbol{\lambda}_{r+1}}{\boldsymbol{\lambda}_r}\right),$$

where $\|\cdot\|$ is the spectral norm, $\mathbf{P}_{\mathbf{Y},1}^\dagger = \mathbf{U}_1\mathbf{U}_1^\top + \mathbf{U}_1\mathbf{Q}^\top\mathbf{U}_2^\top + \mathbf{U}_2\mathbf{Q}\mathbf{U}_1^\top$, and $\mathbf{Q} = \mathbf{\Lambda}_2^{1/2}\mathbf{Z}_2\mathbf{Z}_1^\top(\mathbf{Z}_1\mathbf{Z}_1^\top)^{-1}\mathbf{\Lambda}_1^{-1/2}$.

Remark 5. The condition $\mathrm{tr}(\mathbf{\Lambda}_2)/(n\boldsymbol{\lambda}_r) \to 0$ is commonly adopted in studies on principal subspaces. In fact, when this condition is violated, the principal subspace loses its relation to the rank-r eigenspace of $\mathbf{\Sigma}$; see, for example, Nadler (2008).

Remark 6. Several high-order Davis–Kahan theorems have been established, for example, Lemma 2 in Koltchinskii and Lounici (2016) and Lemma 2 in Fan et al.

(2019). These general results explicitly characterize the linear term and the high-order error on the rank-r eigenspace due to matrix perturbation. Applying these results to $\hat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}$, we can obtain similar results to that given in Proposition 3; however, the above results are slightly weaker and require stronger conditions.

If $p > n$, let $\mathbf{U}_{\mathbf{Y},2}$ be the $r+1$ to nth columns of $\mathbf{U}_{\mathbf{Y}}$. Then, $\mathbf{P}_{\mathbf{Y},2} = \mathbf{U}_{\mathbf{Y},2}\mathbf{U}_{\mathbf{Y},2}^\top$ is the projection matrix onto the eigenspace spanned by the $r+1$ to nth eigenvectors of $\hat{\mathbf{\Sigma}}$. Our later analysis also requires the asymptotic properties of $\mathbf{P}_{\mathbf{Y},2}$, which have not been considered in the literature. Let $\mathbf{V}_{\mathbf{Z}_1} = \mathbf{Z}_1^\top(\mathbf{Z}_1\mathbf{Z}_1^\top)^{-1/2}$. Then, $\mathbf{V}_{\mathbf{Z}_1}\mathbf{V}_{\mathbf{Z}_1}^\top = \mathbf{Z}_1^\top(\mathbf{Z}_1\mathbf{Z}_1^\top)^{-1}\mathbf{Z}_1$ is the projection matrix onto the row space of $\mathbf{Z}_1$. Let $\tilde{\mathbf{V}}_{\mathbf{Z}_1}$ be an $n \times (n-r)$ column orthogonal matrix that satisfies $\tilde{\mathbf{V}}_{\mathbf{Z}_1}\tilde{\mathbf{V}}_{\mathbf{Z}_1}^\top = \mathbf{I}_n - \mathbf{V}_{\mathbf{Z}_1}\mathbf{V}_{\mathbf{Z}_1}^\top$. The following proposition gives the asymptotic behavior of $\mathbf{P}_{\mathbf{Y},2}$.

Proposition 4. *Suppose $r = o(n)$, $\operatorname{tr}(\mathbf{\Lambda}_2)\boldsymbol{\lambda}_1/(n\boldsymbol{\lambda}_r^2) \to 0$, and $n\boldsymbol{\lambda}_{r+1}/\operatorname{tr}(\mathbf{\Lambda}_2) \to 0$. Then,*

$$\left\|\mathbf{P}_{\mathbf{Y},2} - \mathbf{P}_{\mathbf{Y},2}^\dagger\right\| = O_P\left(\sqrt{\frac{\operatorname{tr}(\mathbf{\Lambda}_2)\boldsymbol{\lambda}_1}{n\boldsymbol{\lambda}_r^2}} + \sqrt{\frac{n\boldsymbol{\lambda}_{r+1}}{\operatorname{tr}(\mathbf{\Lambda}_2)}}\right),$$

where $\mathbf{P}_{\mathbf{Y},2}^\dagger = (\operatorname{tr}(\mathbf{\Lambda}_2))^{-1}\mathbf{U}_2\mathbf{\Lambda}_2^{1/2}\mathbf{Z}_2\tilde{\mathbf{V}}_{\mathbf{Z}_1}\tilde{\mathbf{V}}_{\mathbf{Z}_1}^\top\mathbf{Z}_2^\top\mathbf{\Lambda}_2^{1/2}\mathbf{U}_2^\top$.

Remark 7. The condition $\operatorname{tr}(\mathbf{\Lambda}_2)\boldsymbol{\lambda}_1/(n\boldsymbol{\lambda}_r^2) \to 0$ is stronger than the condition $\operatorname{tr}(\mathbf{\Lambda}_2)/(n\boldsymbol{\lambda}_r) \to 0$ in Proposition 3. These two conditions are equivalent if $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_r$ are of the same order.

Now, we are ready to derive the asymptotic properties of $T(\mathbf{X})$ under the spiked covariance. Let $\mathbf{W}_{k-1}^*$ be a $(k-1)\times(k-1)$ symmetric random matrix, distributed as $\text{Wishart}(r, \mathbf{I}_{k-1})$ and independent of $\mathbf{W}_{k-1}$, where $\text{Wishart}(m, \mathbf{\Psi})$ is the Wishart distribution with parameter $\mathbf{\Psi}$ and m degrees of freedom. The following theorem gives the asymptotic distribution of $T(\mathbf{X})$ under the null and local alternative hypotheses.

Theorem 2. *Suppose $r = o(\sqrt{n})$, $r\operatorname{tr}(\mathbf{\Lambda}_2)\boldsymbol{\lambda}_1/(n\boldsymbol{\lambda}_r^2) \to 0$, $rn\boldsymbol{\lambda}_{r+1}/\operatorname{tr}(\mathbf{\Lambda}_2) \to 0$, $r\boldsymbol{\lambda}_{r+1}/\sqrt{\operatorname{tr}(\mathbf{\Lambda}_2^2)} \to 0$, and $\boldsymbol{\lambda}_{r+1} - \boldsymbol{\lambda}_p = O(n^{-1}\sqrt{\operatorname{tr}(\mathbf{\Lambda}_2^2)})$. Then,*

(i) under the null hypothesis $\mathbf{\Theta}\mathbf{C} = \mathbf{O}_{p\times(k-1)}$,

$$\begin{aligned}&\frac{T(\mathbf{X}) - \left((1+r/n)\operatorname{tr}(\mathbf{\Lambda}_2) - n\operatorname{tr}(\mathbf{\Lambda}_2^2)/\operatorname{tr}(\mathbf{\Lambda}_2)\right)}{\sqrt{rn^{-2}\operatorname{tr}^2(\mathbf{\Lambda}_2) + \operatorname{tr}(\mathbf{\Lambda}_2^2)}}\\ &\sim \lambda_1\left(\frac{n^{-1}\operatorname{tr}(\mathbf{\Lambda}_2)}{\sqrt{rn^{-2}\operatorname{tr}^2(\mathbf{\Lambda}_2) + \operatorname{tr}(\mathbf{\Lambda}_2^2)}}(\mathbf{W}_{k-1}^* - r\mathbf{I}_{k-1})\right.\end{aligned}$$

$$+ \frac{\sqrt{\operatorname{tr}(\boldsymbol{\Lambda}_2^2)}}{\sqrt{rn^{-2}\operatorname{tr}^2(\boldsymbol{\Lambda}_2)+\operatorname{tr}(\boldsymbol{\Lambda}_2^2)}}\mathbf{W}_{k-1}\Bigg) + o_P(1);$$

(ii) *if* $r \to \infty$ *or* $\operatorname{tr}(\boldsymbol{\Lambda}_2)/(n\sqrt{\operatorname{tr}(\boldsymbol{\Lambda}_2^2)}) \to 0$, *then under the local alternative hypothesis* $\|\mathbf{C}^\top\boldsymbol{\Theta}^\top\boldsymbol{\Theta}\mathbf{C}\| = O(\sqrt{rn^{-2}\operatorname{tr}^2(\boldsymbol{\Lambda}_2)+\operatorname{tr}(\boldsymbol{\Lambda}_2^2)})$,

$$\frac{T(\mathbf{X}) - \big((1+r/n)\operatorname{tr}(\boldsymbol{\Lambda}_2) - n\operatorname{tr}(\boldsymbol{\Lambda}_2^2)/\operatorname{tr}(\boldsymbol{\Lambda}_2)\big)}{\sqrt{rn^{-2}\operatorname{tr}^2(\boldsymbol{\Lambda}_2)+\operatorname{tr}(\boldsymbol{\Lambda}_2^2)}}$$
$$\sim \lambda_1\Bigg(\frac{n^{-1}\operatorname{tr}(\boldsymbol{\Lambda}_2)}{\sqrt{rn^{-2}\operatorname{tr}^2(\boldsymbol{\Lambda}_2)+\operatorname{tr}(\boldsymbol{\Lambda}_2^2)}}(\mathbf{W}_{k-1}^* - r\mathbf{I}_{k-1})$$
$$+ \frac{\sqrt{\operatorname{tr}(\boldsymbol{\Lambda}_2^2)}}{\sqrt{rn^{-2}\operatorname{tr}^2(\boldsymbol{\Lambda}_2)+\operatorname{tr}(\boldsymbol{\Lambda}_2^2)}}\mathbf{W}_{k-1} + \frac{\mathbf{C}^\top\boldsymbol{\Theta}^\top\mathbf{U}_2\mathbf{U}_2^\top\boldsymbol{\Theta}\mathbf{C}}{\sqrt{rn^{-2}\operatorname{tr}^2(\boldsymbol{\Lambda}_2)+\operatorname{tr}(\boldsymbol{\Lambda}_2^2)}}\Bigg) + o_P(1).$$

Remark 8. Suppose the approximate factor model in Fan, Yuan and Mincheva (2013) holds. That is, r is fixed, $\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_r$ diverge at rate $O(p)$, and $\boldsymbol{\lambda}_{r+1},\ldots,\boldsymbol{\lambda}_p$ are bounded. Then, the conditions of Theorem 2 become $p/n \to \infty$ and $\boldsymbol{\lambda}_{r+1} - \boldsymbol{\lambda}_p = O(\sqrt{p}/n)$. Hence, Theorem 2 holds for ultrahigh-dimensional data. In contrast, recent tests under the spiked covariance model can only be used for lower-dimensional data. In fact, under the approximate factor model in Fan, Yuan and Mincheva (2013), Aoshima and Yata (2018) requires $p/n \to 0$, and Wang and Xu (2018) requires $p/n^2 \to 0$ and $\boldsymbol{\lambda}_{r+1} = \cdots = \boldsymbol{\lambda}_p$. Note that if $k = 2$ and $p/n^2 \to 0$, then the coefficient of $\mathbf{W}_{k-1}^* - r\mathbf{I}_{k-1}$ is negligible, and, as a result, $T(\mathbf{X})$ is asymptotically normally distributed. Thus, Theorem 2 gives the high-order behavior of $T(\mathbf{X})$.

Now, we formulate a test procedure with an asymptotically correct level. Define the standardized statistic as

$$Q_2 = \frac{T(\mathbf{X}) - \Big((1+r/n)\widehat{\operatorname{tr}(\boldsymbol{\Lambda}_2)} - n\widehat{\operatorname{tr}(\boldsymbol{\Lambda}_2^2)}/\widehat{\operatorname{tr}(\boldsymbol{\Lambda}_2)}\Big)}{\sqrt{rn^{-2}(\widehat{\operatorname{tr}(\boldsymbol{\Lambda}_2)})^2 + \widehat{\operatorname{tr}(\boldsymbol{\Lambda}_2^2)}}}.$$

Let $F_2(x;\operatorname{tr}(\boldsymbol{\Lambda}_2),\operatorname{tr}(\boldsymbol{\Lambda}_2^2))$ be the cumulative distribution function of

$$\lambda_1\Bigg(\frac{n^{-1}\operatorname{tr}(\boldsymbol{\Lambda}_2)}{\sqrt{rn^{-2}\operatorname{tr}^2(\boldsymbol{\Lambda}_2)+\operatorname{tr}(\boldsymbol{\Lambda}_2^2)}}(\mathbf{W}_{k-1}^* - r\mathbf{I}_{k-1}) + \frac{\sqrt{\operatorname{tr}(\boldsymbol{\Lambda}_2^2)}}{\sqrt{rn^{-2}\operatorname{tr}^2(\boldsymbol{\Lambda}_2)+\operatorname{tr}(\boldsymbol{\Lambda}_2^2)}}\mathbf{W}_{k-1}\Bigg).$$

Then, we reject the null hypothesis if

$$Q_2 > F_2^{-1}\left(1-\alpha; \widehat{\mathrm{tr}(\mathbf{\Lambda}_2)}, \widehat{\mathrm{tr}(\mathbf{\Lambda}_2^2)}\right).$$

The following corollary shows that this test procedure has an asymptotically correct level, as well as giving the asymptotic local power function.

Corollary 2. *Suppose the conditions of Theorem* 2 *hold. Then,*

(i) under the null hypothesis $\mathbf{\Theta C} = \mathbf{O}_{p\times(k-1)}$,

$$\Pr\left(Q_2 > F_2^{-1}\left(1-\alpha; \widehat{\mathrm{tr}(\mathbf{\Lambda}_2)}, \widehat{\mathrm{tr}(\mathbf{\Lambda}_2^2)}\right)\right) = \alpha + o(1);$$

(ii) if $r \to \infty$ *or* $\mathrm{tr}(\mathbf{\Lambda}_2)/(n\sqrt{\mathrm{tr}(\mathbf{\Lambda}_2^2)}) \to 0$, *then under the local alternative hypothesis* $\|\mathbf{C}^\top\mathbf{\Theta}^\top\mathbf{\Theta C}\| = O(\sqrt{rn^{-2}\,\mathrm{tr}^2(\mathbf{\Lambda}_2) + \mathrm{tr}(\mathbf{\Lambda}_2^2)})$,

$$\begin{aligned}
&\Pr\left(Q_2 > F_2^{-1}\left(1-\alpha; \widehat{\mathrm{tr}(\mathbf{\Lambda}_2)}, \widehat{\mathrm{tr}(\mathbf{\Lambda}_2^2)}\right)\right)\\
&= \Pr\Bigg(\lambda_1\Bigg(\frac{n^{-1}\,\mathrm{tr}(\mathbf{\Lambda}_2)}{\sqrt{rn^{-2}\,\mathrm{tr}^2(\mathbf{\Lambda}_2)+\mathrm{tr}(\mathbf{\Lambda}_2^2)}}(\mathbf{W}_{k-1}^* - r\mathbf{I}_{k-1})\\
&\quad + \frac{\sqrt{\mathrm{tr}(\mathbf{\Lambda}_2^2)}}{\sqrt{rn^{-2}\,\mathrm{tr}^2(\mathbf{\Lambda}_2)+\mathrm{tr}(\mathbf{\Lambda}_2^2)}}\mathbf{W}_{k-1} + \frac{\mathbf{C}^\top\mathbf{\Theta}^\top\mathbf{U}_2\mathbf{U}_2^\top\mathbf{\Theta C}}{\sqrt{rn^{-2}\,\mathrm{tr}^2(\mathbf{\Lambda}_2)+\mathrm{tr}(\mathbf{\Lambda}_2^2)}}\Bigg)\\
&\quad > F_2^{-1}\left(1-\alpha; \mathrm{tr}(\mathbf{\Lambda}_2), \mathrm{tr}(\mathbf{\Lambda}_2^2)\right)\Bigg) + o(1).
\end{aligned}$$

To gain some insight into the asymptotic behavior of $T(\mathbf{X})$, we consider $k=2$ and compare the power of the LFD test with that of Bai and Saranadasa (1996) and Chen and Qin (2010). Corollary 2 implies that if

$$\liminf_{n\to\infty} \frac{\mathbf{C}^\top\mathbf{\Theta}^\top\mathbf{U}_2\mathbf{U}_2^\top\mathbf{\Theta C}}{\sqrt{rn^{-2}\,\mathrm{tr}^2(\mathbf{\Lambda}_2)+\mathrm{tr}(\mathbf{\Lambda}_2^2)}} > 0,$$

then the LFD test has nontrivial power, asymptotically. In contrast, if

$$\limsup_{n\to\infty} \frac{\mathbf{C}^\top\mathbf{\Theta}^\top\mathbf{\Theta C}}{\sqrt{\mathrm{tr}(\mathbf{\Sigma}^2)}} = 0,$$

then the tests in Bai and Saranadasa (1996) and Chen and Qin (2010) exhibit trivial power, asymptotically. To compare $\mathbf{C}^\top\mathbf{\Theta}^\top\mathbf{U}_2\mathbf{U}_2^\top\mathbf{\Theta C}$ and $\mathbf{C}^\top\mathbf{\Theta}^\top\mathbf{\Theta C}$, we temporarily place a prior on $\mathbf{\Theta}$. Suppose $\sqrt{n_i}\theta_i$ has prior distribution $\mathcal{N}_p(\mathbf{0}_p, \psi\mathbf{I}_p)$, for $i=1,2$. Then, $\psi^{-1}\mathbf{C}^\top\mathbf{\Theta}^\top\mathbf{\Theta C}$ follows a χ^2 distribution with p degrees of freedom. On the other hand, $\psi^{-1}\mathbf{C}^\top\boldsymbol{\Theta}^\top\mathbf{U}_2\mathbf{U}_2^\top\boldsymbol{\Theta}\mathbf{C}$ follows a χ^2 distribution with $p-r$ degrees of freedom. Thus, we have

$$\frac{\mathbf{C}^\top\boldsymbol{\Theta}^\top\mathbf{U}_2\mathbf{U}_2^\top\boldsymbol{\Theta}\mathbf{C}}{\mathbf{C}^\top\boldsymbol{\Theta}^\top\boldsymbol{\Theta}\mathbf{C}} \xrightarrow{P} 1.$$

Therefore, on average, the signal contained in $\mathbf{C}^\top\boldsymbol{\Theta}^\top\mathbf{U}_2\mathbf{U}_2^\top\boldsymbol{\Theta}\mathbf{C}$ is roughly the same as that in $\mathbf{C}^\top\boldsymbol{\Theta}^\top\boldsymbol{\Theta}\mathbf{C}$. Now, we compare the asymptotic variance. It is not hard to see that under the conditions of Theorem 2, we have $rn^{-2}\operatorname{tr}^2(\boldsymbol{\Lambda}_2)/\operatorname{tr}(\boldsymbol{\Sigma}^2) \to 0$. Also, if $\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_r$ are sufficiently large, then $\operatorname{tr}(\boldsymbol{\Lambda}_2^2)/\operatorname{tr}(\boldsymbol{\Sigma}^2) \to 0$. Hence, it can be expected that

$$\frac{rn^{-2}\operatorname{tr}^2(\boldsymbol{\Lambda}_2)+\operatorname{tr}(\boldsymbol{\Lambda}_2^2)}{\operatorname{tr}(\boldsymbol{\Sigma}^2)} \to 0.$$

That is, the asymptotic variance of $T(\mathbf{X})$ is typically much smaller than those of the tests in Bai and Saranadasa (1996) and Chen and Qin (2010). To appreciate this, note that in the expression (2.2), $(\mathbf{I}_p-\mathbf{P}_\mathbf{Y})\mathbf{XJC}|\mathbf{P}_\mathbf{Y} \sim \mathcal{N}_p(\mathbf{0}_p,(\mathbf{I}_p-\mathbf{P}_\mathbf{Y})\boldsymbol{\Sigma}(\mathbf{I}_p-\mathbf{P}_\mathbf{Y}))$. However, $\mathbf{I}_p-\mathbf{P}_\mathbf{Y}$ tends to be orthogonal to $\mathbf{U}_1\mathbf{U}_1^\top$, which is the projection matrix onto the eigenspace corresponding to the leading eigenvalues of $\boldsymbol{\Sigma}$. Hence, the projection by $\mathbf{I}_p-\mathbf{P}_\mathbf{Y}$ helps reduce the variance of $\mathbf{XJC}$.

Thus, if $\boldsymbol{\Theta}$ satisfies

$$\liminf_{n\to\infty}\frac{\mathbf{C}^\top\boldsymbol{\Theta}^\top\boldsymbol{\Theta}\mathbf{C}}{\sqrt{rn^{-2}\operatorname{tr}^2(\boldsymbol{\Lambda}_2)+\operatorname{tr}(\boldsymbol{\Lambda}_2^2)}} > 0, \quad \limsup_{n\to\infty}\frac{\mathbf{C}^\top\boldsymbol{\Theta}^\top\boldsymbol{\Theta}\mathbf{C}}{\sqrt{\operatorname{tr}(\boldsymbol{\Sigma}^2)}} = 0,$$

then the LFD test has nontrivial power, whereas the tests in Bai and Saranadasa (1996) and Chen and Qin (2010) exhibit trivial power. Hence, the LFD test tends to be more powerful than those of Bai and Saranadasa (1996) and Chen and Qin (2010).

In practice, we may not know whether the covariance matrix is spiked. Furthermore, even if we know that it is spiked, the spike number r may be unknown. Therefore, we propose an adaptive test procedure. Note that Theorem 1 requires $n\boldsymbol{\lambda}_1/\operatorname{tr}(\boldsymbol{\Sigma}) \to 0$, and Theorem 2 requires $\operatorname{tr}(\boldsymbol{\Lambda}_2)/n\boldsymbol{\lambda}_r \to 0$ and $n\boldsymbol{\lambda}_{r+1}/\operatorname{tr}(\boldsymbol{\Lambda}_2) \to 0$. This motivates us to consider the following adaptive test procedure. Let $\tau > 1$ be a hyperparameter. If

$$\frac{n\lambda_1(\hat{\boldsymbol{\Sigma}})}{\operatorname{tr}(\hat{\boldsymbol{\Sigma}})} < \tau,$$

then we reject the null hypothesis if $Q_1 > F^{-1}(1-\alpha)$. Otherwise, we reject the null hypothesis if $Q_2 > F_2^{-1}(1-\alpha; \widehat{\mathrm{tr}(\mathbf{\Lambda}_2,)}, \widehat{\mathrm{tr}(\mathbf{\Lambda}_2^2)})$, where the unknown r is substituted by the estimator

$$\hat{r} = \min\left\{1 \le i < n : \frac{n\lambda_{i+1}(\hat{\mathbf{\Sigma}})}{\sum_{j=i+1}^{n} \lambda_j(\hat{\mathbf{\Sigma}})} < \tau\right\}.$$

We have the following proposition.

Proposition 5. *Let $\tau > 1$ be a constant.*

(i) Under the conditions of Theorem 1*,*

$$\Pr\left(\frac{n\lambda_1(\hat{\mathbf{\Sigma}})}{\mathrm{tr}(\hat{\mathbf{\Sigma}})} < \tau\right) \to 1;$$

(ii) Under the conditions of Theorem 2*,*

$$\Pr\left(\frac{n\lambda_1(\hat{\mathbf{\Sigma}})}{\mathrm{tr}(\hat{\mathbf{\Sigma}})} < \tau\right) \to 0, \quad \Pr(\hat{r} = r) \to 1.$$

Proposition 5 implies that the spiked covariance structure can be detected consistently. Therefore, the proposed adaptive LFD test procedure can indeed adapt to the unknown covariance structure.

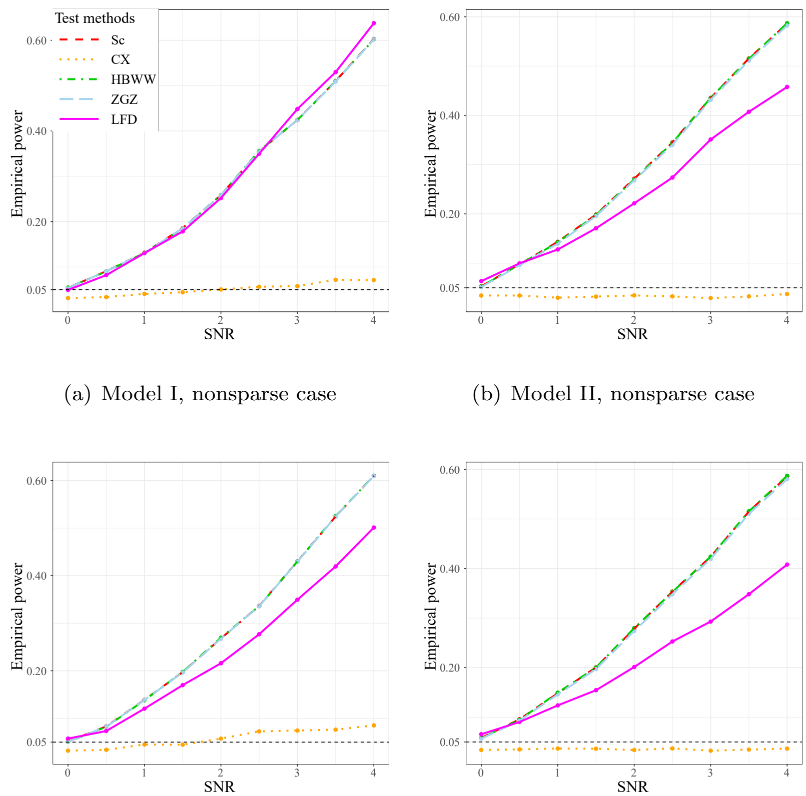
4. Numerical Study

In this section, we compare the numerical performance of the adaptive LFD test procedure with that of the MANOVA tests in Schott (2007), Cai and Xia (2014), Hu et al. (2017), and Zhang, Guo and Zhou (2017). These competing tests are denoted by Sc, CX, HBWW, and ZGZ, respectively. Throughout the simulations, we take the nominal test level $\alpha = 0.05$ and the group number $k = 3$. For the adaptive LFD test, we take $\tau = 5$. For CX, we use their oracle procedure. All simulation results are based on 5,000 replications.

First, we simulate the empirical level and power under various models of $\mathbf{\Sigma}$ and $\mathbf{\Theta}$. To characterize the signal strength, we define the signal-to-noise ratio (SNR) as

$$\mathrm{SNR} = \frac{\mathbf{C}^\top \mathbf{\Theta}^\top \mathbf{\Theta} \mathbf{C}}{\sqrt{\mathrm{tr}(\mathbf{\Sigma}^2)}}.$$

We consider four models for $\mathbf{\Sigma}$, where the first two are nonspiked, and the last



(a) Model I, nonsparse case

(b) Model II, nonsparse case

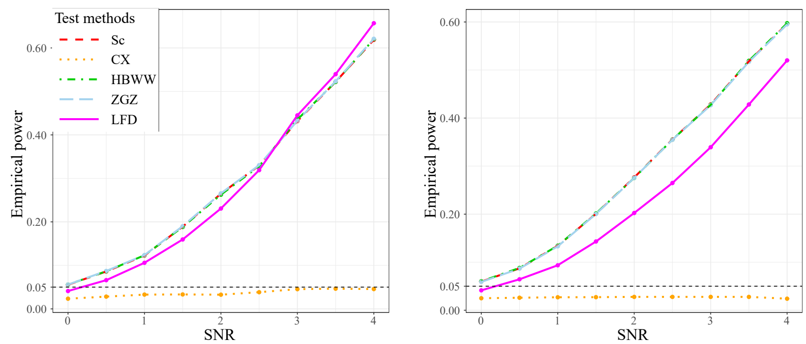
(c) Model I, sparse case

(d) Model II, sparse case

Figure 1. Empirical size and power of tests under model I and model II; $n_1 = n_2 = n_3 = 20$, $p = 300$.

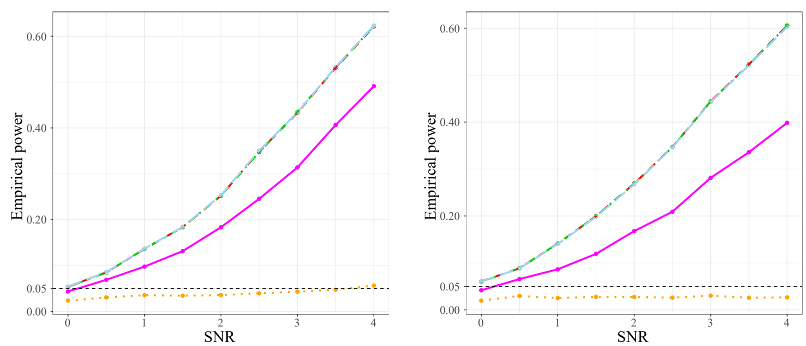
two are spiked.

- Model I: $\mathbf{\Sigma} = \mathbf{I}_p$.
- Model II: $\mathbf{\Sigma} = (\sigma_{ij})$, where $\sigma_{ij} = 0.6^{|i-j|}$.
- Model III: $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{U}$ is a $p \times p$ orthogonal matrix generated from the Haar distribution and $\mathbf{\Lambda} = \mathrm{diag}(3p, 2p, p, 1, \ldots, 1)$.
- Model IV: $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top + \mathbf{A}\mathbf{A}^\top$, where $\mathbf{U}$ is a $p \times p$ orthogonal matrix generated from the Haar distribution, $\mathbf{\Lambda} = \mathrm{diag}(p, p, 1, \ldots, 1)$, and $\mathbf{A}$ is a $p \times p$ matrix, the elements of which are independently generated from the Bernoulli distribution with success probability 0.01.



(a) Model I, nonsparse case

(b) Model II, nonsparse case

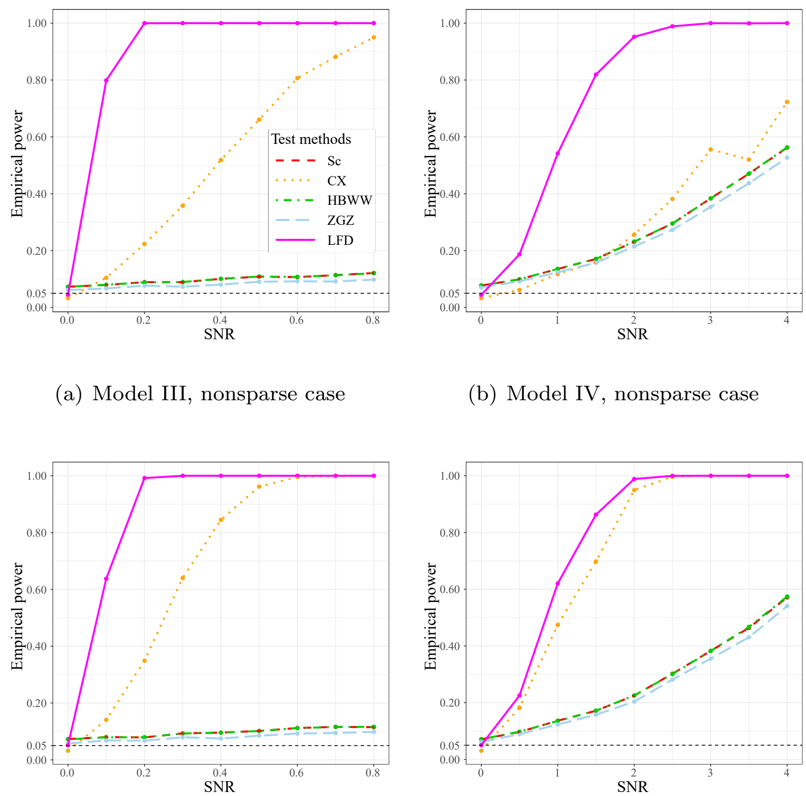


(c) Model I, sparse case

(d) Model II, sparse case

Figure 2. Empirical size and power of tests under model I and model II; $n_1 = n_2 = n_3 = 25$, $p = 800$.

Under the null hypothesis, we always take $\theta_1 = \cdots = \theta_k = \mathbf{0}_p$. We consider two structures for the alternative hypotheses: the nonsparse alternative, and the sparse alternative. In the nonsparse case, we take $\theta_1 = \kappa \mathbf{1}_p$, $\theta_2 = -\kappa \mathbf{1}_p$, and $\theta_3 = \mathbf{0}_p$, where κ is selected to make the SNR equal to specific values. In the sparse case, we take $\theta_1 = \kappa(\mathbf{1}_{p/5}^\top, \mathbf{0}_{4p/5}^\top)^\top$, $\theta_2 = \kappa(\mathbf{0}_{p/5}^\top, \mathbf{1}_{p/5}^\top, \mathbf{0}_{3p/5}^\top)^\top$, and $\theta_3 = \mathbf{0}_p$. Again, κ is selected to make the SNR equal to specific values. The simulation results are summarized in Figures 1–4, and show that in all scenarios, the empirical sizes of the LFD test are reasonably close to the nominal level 0.05. Under model I and model II, where the covariance matrices are nonspiked, the empirical power of the LFD test is slightly lower than that of the sum-of-squares tests, but is higher than that of the CX test. Under model III and model IV,



(a) Model III, nonsparse case

(b) Model IV, nonsparse case

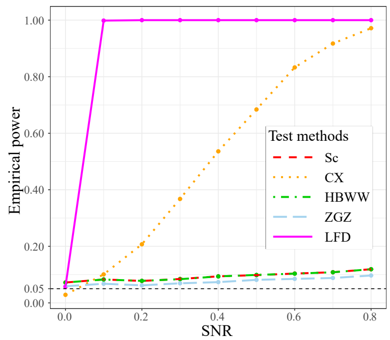
(c) Model III, sparse case

(d) Model IV, sparse case

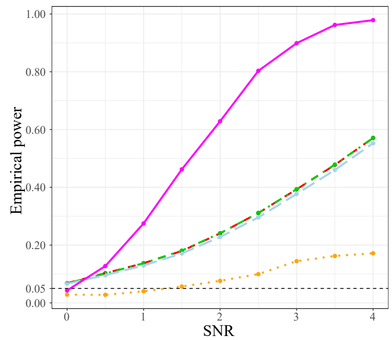
Figure 3. Empirical size and power of tests under model III and model IV; $n_1 = n_2 = n_3 = 20$, $p = 300$.

where the covariance matrices are spiked, the empirical power of the LFD test is significantly higher than that of the sum-of-squares tests. In addition, the LFD test exhibits higher empirical power than that of the CX test in most cases, except for model IV with sparse means. These simulation results verify our theoretical results that the LFD test is particularly powerful under the spiked covariance.

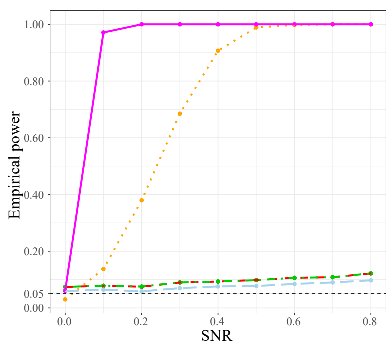
In our second simulation study, we investigate the effect of correlations between the variables. We consider the compound symmetry structure; that is, the diagonal elements of $\mathbf{\Sigma}$ are one, and the off-diagonal elements are ρ, with $0 \leq \rho < 1$. The parameter ρ characterizes the correlations between the variables. We take $\theta_1 = \kappa(\mathbf{1}_{p/5}^\top, \mathbf{0}_{4p/5}^\top)^\top$, $\theta_2 = \kappa(\mathbf{0}_{p/5}^\top, \mathbf{1}_{p/5}^\top, \mathbf{0}_{3p/5}^\top)^\top$, and $\theta_3 = \mathbf{0}_p$, where κ is selected such that $\mathbf{C}^\top \mathbf{\Theta}^\top \mathbf{\Theta} \mathbf{C}/(\sum_{i=2}^p \boldsymbol{\lambda}_i^2)^{1/2} = 5$. Figure 5 plots the empirical



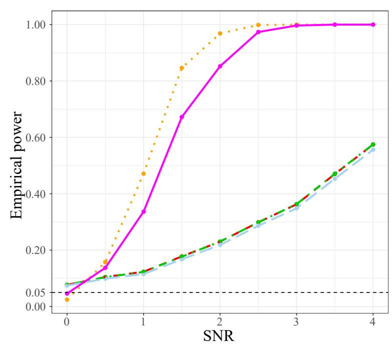
(a) Model III, nonsparse case



(b) Model IV, nonsparse case



(c) Model III, sparse case



(d) Model IV, sparse case

Figure 4. Empirical size and power of tests under model III and model IV; $n_1 = n_2 = n_3 = 25$, $p = 800$.

power for various tests versus ρ. We can see that the empirical power of the LFD test remains nearly constant as ρ varies, whereas the empirical power of the competing sum-of-squares tests decreases rapidly as ρ increases. When ρ is nonzero, the LFD test outperforms the competing tests significantly.

5. Concluding Remarks

Using the idea of the least favorable direction, we have proposed an LFD test for MANOVA in the high-dimensional setting. We have derived the asymptotic distribution of the LFD test statistic under both nonspiked and spiked covariances. The asymptotic local power functions are also given. Our theoretical results and simulation studies show that the LFD test exhibits power behavior

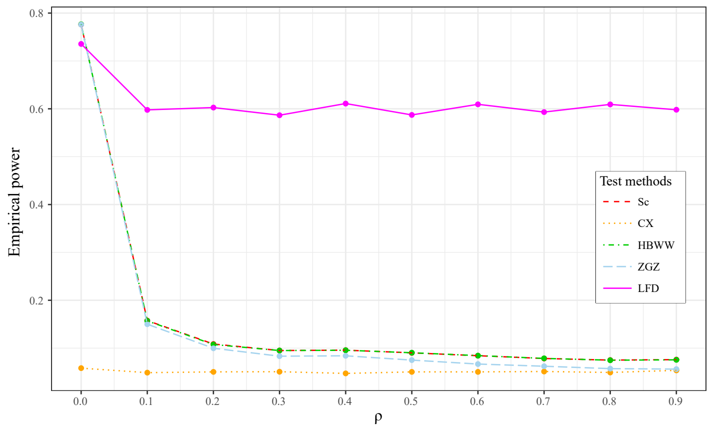


Figure 5. Empirical power of tests; $n_1 = n_2 = n_3 = 35$, $p = 1{,}000$.

comparable with that of existing tests when the covariance matrix is nonspiked, and tends to be much more powerful than existing tests when the covariance matrix is spiked.

Several interesting, but challenging problems remain. First, for the case of an unknown covariance structure, we proposed an adaptive LFD test procedure by consistently detecting the unknown covariance structure and estimating the unknown r. However, this procedure relies on a hyperparameter τ. Determining an optimal τ remains an interesting problem. Second, our theoretical results rely on the normality of the observations. In fact, our proofs use the independence of $\mathbf{XJC}$ and $\mathbf{Y}$. Note that $\mathbf{XJC}$ and $\mathbf{Y} = \mathbf{X}\tilde{\mathbf{J}}$ are both linear combinations of independent random vectors X_{ij}. It is known that the independence of linear combinations of independent random variables essentially characterizes the normality of the variables; see, for example, Kagan, Linnik and Rao (1973), Section 3.1. Hence our strategy is not feasible without the normality assumption. It is unclear whether the conclusions of our theorems hold without this assumption. Third, our theoretical results require $p/n \to \infty$. In fact, the asymptotic behavior of $T(\mathbf{X})$ will be different in the regime where $p/n \to$ constant. Random matrix theory may be useful to investigate the asymptotic behavior of $T(\mathbf{X})$ in this regime. We leave these topics for future research.

Supplementary Material

The online Supplementary Material presents proofs of the propositions and theorems.

Acknowledgments

We thank the associate editor and two anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Natural Science Foundation of China under Grant Nos. 11471035 and 11471030.

References

Ahn, J. and Marron, J. S. (2010). The maximal data piling direction for discrimination. *Biometrika* **97**, 254–259.

Aoshima, M., Shen, D., Shen, H., Yata, K., Zhou, Y.-H. and Marron, J. S. (2018). A survey of high dimension low sample size asymptotics. *Australian & New Zealand Journal of Statistics* **60**, 4–19.

Aoshima, M. and Yata, K. (2018). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statistica Sinica* **28**, 43–62.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statistica Sinica* **6**, 311–329.

Cai, T., Han, X. and Pan, G. (2019). Limiting laws for divergent spiked eigenvalues and largest non-spiked eigenvalue of sample covariance matrices. *The Annals of Statistics*. In press.

Cai, T., Ma, Z. and Wu, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory & Related Fields* **161**, 781–815.

Cai, T. T., Liu, W. and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 349–372.

Cai, T. T. and Xia, Y. (2014). High-dimensional sparse MANOVA. *Journal of Multivariate Analysis* **131**, 174–196.

Cao, M.-X., Park, J. and He, D.-J. (2019). A test for the k sample behrens–fisher problem in high dimensional data. *Journal of Statistical Planning and Inference* **201**, 86–102.

Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics* **38**, 808–835.

Fan, J., Wang, D., Wang, K. and Zhu, Z. (2019). Distributed estimation of principal eigenspaces. *The Annals of Statistics* **47**, 3009–3031.

Fan, J., Yuan, L. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**, 603–680.

Feng, L., Zou, C., Wang, Z. and Zhu, L. (2015). Two-sample behrens-fisher problem for high-dimensional data. *Statistica Sinica* **25**, 1297–1312.

Hu, J., Bai, Z., Wang, C. and Wang, W. (2017). On testing the equality of high dimensional mean vectors with unequal covariance matrices. *Annals of the Institute of Statistical Mathematics* **69**, 365–387.

Kagan, A., Linnik, Y. and Rao, C. (1973). *Characterization Problems in Mathematical Statistics*. 1st Edition. Wiley, New York.

Katayama, S., Kano, Y. and Srivastava, M. S. (2013). Asymptotic distributions of some test criteria for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis* **116**, 410–421.

Koltchinskii, V. and Lounici, K. (2016). Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **52**, 1976–2013.

Ma, Y., Lan, W. and Wang, H. (2015). A high dimensional two-sample test under a low dimensional factor structure. *Journal of Multivariate Analysis* **140**, 162–170.

Nadler, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics* **36**, 2791–2817.

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics* **24**, 220–238.

Schott, J. R. (2007). Some high-dimensional tests for a one-way MANOVA. *Journal of Multivariate Analysis* **98**, 1825–1839.

Shen, D., Shen, H. and Marron, J. S. (2016). A general framework for consistency of principal component analysis. *Journal of Machine Learning Research* **17**, 1–34.

Srivastava, M. S. (2007). Multivariate theory for analyzing high dimensional data. *Journal of the Japan Statistical Society* **37**, 53–86.

Srivastava, M. S. and Kubokawa, T. (2013). Tests for multivariate analysis of variance in high dimension under non-normality. *Journal of Multivariate Analysis* **115**, 204–216.

Tsai, C.-A. and Chen, J. J. (2009). Multivariate analysis of variance test for gene set analysis. *Bioinformatics* **25**, 897–903.

Verstynen, T., Diedrichsen, J., Albert, N., Aparicio, P. and Ivry, R. (2005). Ipsilateral motor cortex activity during unimanual hand movements relates to task complexity. *Journal of Neurophysiology* **93**, 1209–1222.

Wang, R. and Xu, X. (2018). On two-sample mean tests under spiked covariances. *Journal of Multivariate Analysis* **167**, 225–249.

Wang, R. and Xu, X. (2019). A feasible high dimensional randomization test for the mean vector. *Journal of Statistical Planning and Inference* **199**, 160–178.

Wang, W. and Fan, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of Statistics* **45**, 1342–1374.

Yamada, T. and Himeno, T. (2015). Testing homogeneity of mean vectors under heteroscedasticity in high-dimension. *Journal of Multivariate Analysis* **139**, 7–27.

Yata, K. and Aoshima, M. (2012). Effective pca for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of Multivariate Analysis* **105**, 193–215.

Yata, K. and Aoshima, M. (2013). Pca consistency for the power spiked model in high-dimensional settings. *Journal of Multivariate Analysis* **122**, 334–354.

Zhang, J.-T., Guo, J. and Zhou, B. (2017). Linear hypothesis testing in high-dimensional one-way manova. *Journal of Multivariate Analysis* **155**, 200–216.

Zhao, J. and Xu, X. (2016). A generalized likelihood ratio test for normal mean when p is greater than n. *Computational Statistics & Data Analysis* **99**, 91–104.

Zhou, B., Guo, J. and Zhang, J.-T. (2017). High-dimensional general linear hypothesis testing under heteroscedasticity. *Journal of Statistical Planning and Inference* **188**, 36–54.

Rui Wang

School of Mathematics and Statistics, Beijing Institute of Technology, Beijing, 100081, China.

E-mail: wangruiphd@bit.edu.cn

Xingzhong Xu

School of Mathematics and Statistics, Beijing Institute of Technology, Beijing, 100081, China.
Beijing Key Laboratory on MCAACI, Beijing Institute of Technology, Beijing, 100081, China.

E-mail: xuxz@bit.edu.cn

(Received October 2017; accepted June 2019)