Statistica Sinica Preprint No: SS-2022-0315	
Title	Power Enhancement for Dimension Detection of Gaussian Signals
Manuscript ID	SS-2022-0315
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0315
Complete List of Authors	Gaspard Bernard and Thomas Verdebout
Corresponding Authors	Thomas Verdebout
E-mails	tverdebout@gmail.com

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

Statistica Sinica

Power enhancement for dimension detection of Gaussian signals

Gaspard Bernard and Thomas Verdebout

Université libre de Bruxelles (ULB)

Abstract: We consider the classical problem of testing $\mathcal{H}_{0q}^{(n)} : \lambda_q^{(n)} > \lambda_{q+1}^{(n)} = \ldots = \lambda_p^{(n)}$, where $\lambda_1^{(n)}, \ldots, \lambda_p^{(n)}$ are the ordered latent roots of covariance matrices $\mathbf{\Sigma}^{(n)}$. We show that the usual Gaussian procedure, $\phi^{(n)}$, for this problem essentially shows no power against alternatives of weaker signals of the form $\mathcal{H}_{1q}^{(n)} : \lambda_q^{(n)} = \lambda_{q+1}^{(n)} = \ldots = \lambda_p^{(n)}$, which is problematic if it is used to perform inference on the true dimension of the signal. We show that the same test $\phi^{(n)}$ enjoys some local and asymptotic optimality properties for detecting alternatives to the equality of the $p-q$ smallest roots of $\mathbf{\Sigma}^{(n)}$, provided that $\lambda_q^{(n)}$ and $\lambda_{q+1}^{(n)}$ are sufficiently separated. We obtain tests, $\phi_{\text{new}}^{(n)}$, for the problem that retain the local and asymptotic optimality properties of $\phi^{(n)}$ when $\lambda_q^{(n)}$ and $\lambda_{q+1}^{(n)}$ are sufficiently separated and properly detect alternatives of the form $\mathcal{H}_{1q}^{(n)}$. We illustrate the performances of our tests using simulations and on a gene expression data set, where we also discuss the problem of estimating the dimension of the signal.

Key words and phrases: signal dimension; hypothesis testing; latent roots.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

1. Introduction

Principal component analysis (PCA) is a popular technique for performing unsupervised dimension reduction. The main objective of a PCA is to extract a low-dimensional signal from the data. This can be achieved by first identifying a spiked structure in the underlying $p \times p$ positive-definite covariance matrix $\mathbf{\Sigma}$ using the data at hand. In the very popular spiked covariance models, the underlying covariance matrix $\mathbf{\Sigma}$ has eigenvalues $\lambda_1 \geq \ldots \geq \lambda_q > \sigma^2 = \ldots = \sigma^2 > 0$; see, for instance, Johnstone (2001). In the spiked covariance model, the q largest eigenvalues of $\mathbf{\Sigma}$ are well separated from the rest, and the data at hand can therefore be seen as q-dimensional data contaminated with noise. Inference within spiked covariance models has been considered by Li et al. (2020), Paindaveine et al. (2020a,b), and Bao et al. (2022), among others. In the context of spiked models, and in a PCA in general, an important problem is testing the equality of the $p-q$ smallest eigenvalues $\mathcal{H}_{0q} : \lambda_q > \lambda_{q+1} = \ldots = \lambda_p$ of $\mathbf{\Sigma}$. Under $\mathcal{H}_{0q}$, the smallest $p-q$ eigenvalues are equal so that they correspond to some noise. As a result, selecting more than q principal components is useless. Tests for $\mathcal{H}_{0q}$ are typically used before selecting the number of components to keep. The problem is not new. Bartlett (1950) used tests for $\mathcal{H}_{0q}$ to determine the number of significant factors in a data set of

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

measurements of the reading speed, reading power, arithmetic speed and arithmetic power for 140 children. Tests for $\mathcal{H}_{0q}$ can also be used to check the suitability of a data set for factor analysis, as Şahan et al. (2019), who assessed wether a psychological questionnaire is consistent. In the same spirit, Chakraborty et al. (2020) used tests for $\mathcal{H}_{0q}$ to ensure that every PCA-based sub-indicator is relevant when constructing a socioeconomic index. Finally, as mentioned in Kritchman and Nadler (2009), getting rid of the noise is a critical preliminary step when treating the output of a collection of sensors.

The (full) sphericity problem ($q = 0$ with λ_0 arbitrarily large) has been studied by Ledoit and Wolf (2002), Onatski et al. (2014), Tian et al. (2015), Li and Yao (2016), and Paindaveine and Verdebout (2016) in the high-dimensional case, while Hallin and Paindaveine (2006) proposed locally and asymptotically optimal tests based on signed ranks. Cuesta-Albertos et al. (2009) proposed tests based on random projections, Henze et al. (2014) provided tests based on the characteristic function, and Francq et al. (2017) considered the problem in a time series context. Fixing $q < p - 1$, the problem of testing the equality of the smallest $p - q$ eigenvalues $\mathcal{H}_{0q} : \lambda_q > \lambda_{q+1} = \ldots = \lambda_p$ has also been investigated thoroughly in the multivariate statistics literature. Methods for determining the dimension of a signal

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

can be traced back to the work of Lawley (1956), who developed Gaussian likelihood ratio tests to check the equality of the smallest eigenvalues. A pseudo-Gaussian test that is valid under elliptical assumptions has been proposed in Waternaux (1984). The local asymptotic powers of robust tests have been obtained in Tyler (1983), and other procedures have been investigated by Nadler (2010), Luo and Li (2016), and Nordhausen et al. (2022) among others. High-dimensional tests have been studied in Schott (2006) and, more recently, in Virta (2021).

In the present study, our objective is to provide tests for $\mathcal{H}_{0q} : \lambda_q > \lambda_{q+1} = \ldots = \lambda_p$ that can detect alternatives of *stronger* signals, under which $\lambda_{q+1}, \ldots, \lambda_p$ are not equal, and alternatives of *weaker* signals, under which λ_q and λ_{q+1} are "too close to each other." Note that our tests for $\mathcal{H}_{0q}$ can be adapted easily to tests for other restrictions, such as $\lambda_{q_1} > \lambda_{q_1+1} = \ldots = \lambda_{q_2} > \lambda_{q_2+1}$, for some q_1 and q_2. To properly formalize the problem, we consider a triangular array context, in which the nth line of the array consists of independent and identically distributed (i.i.d.) p-variate Gaussian vectors $\mathbf{X}_{1n}, \ldots, \mathbf{X}_{nn}$ with common covariance matrix $\mathbf{\Sigma}^{(n)} = \boldsymbol{\beta}\mathbf{\Lambda}^{(n)}\boldsymbol{\beta}'$, where $\boldsymbol{\beta}$ is orthogonal and $\mathbf{\Lambda}^{(n)} := \text{diag}(\lambda_1^{(n)}, \ldots, \lambda_p^{(n)})$ is a diagonal matrix of positive ordered eigenvalues that may change with n. Within such sequences of experiments, we consider the (sequence of)

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

hypotheses testing problems characterized by null hypothesis of the form

$$\mathcal{H}_{0q}^{(n)} : \{\lambda_{q+1}^{(n)} = \ldots = \lambda_p^{(n)}\} \cap \{n^{1/2}(\lambda_q^{(n)} - \lambda_{q+1}^{(n)}) \to \infty \text{ as } n \to \infty\}. \quad (1.1)$$

Under $\mathcal{H}_{0q}^{(n)}$, the smallest $p-q$ underlying latent roots are equal, and $\lambda_q^{(n)}$ and $\lambda_{q+1}^{(n)}$ are sufficiently separated in the sense that $n^{1/2}(\lambda_q^{(n)} - \lambda_{q+1}^{(n)}) \to \infty$ as $n \to \infty$. Here, we adapt the aforementioned sequence of hypotheses testing problems to detect the signal dimension. Indeed, a rejection of $\mathcal{H}_{0q}^{(n)}$ indicates that the smallest roots are not equal, in which case the signal is *stronger*, or that $\lambda_q^{(n)}$ and $\lambda_{q+1}^{(n)}$ are too close to each other, in which case the signal is *weaker*. Note that the consistency of an empirical projection on the first q principal axes holds only if $n^{1/2}(\lambda_q^{(n)} - \lambda_{q+1}^{(n)})$ diverges to ∞ as $n \to \infty$; this makes the testing problem associated with $\mathcal{H}_{0q}^{(n)}$ in (1.1) a natural problem to tackle in the context. Alternatives to $\mathcal{H}_{0q}^{(n)}$ (for $q \geq 1$) are of two different types:

(i) type-I alternatives, under which the smallest $p-q$ eigenvalues are not equal and $n^{1/2}(\lambda_q^{(n)} - \lambda_{q+1}^{(n)}) \to \infty$ as $n \to \infty$;

(ii) type-II alternatives, under which $\lambda_q^{(n)}$ and $\lambda_{q+1}^{(n)}$ are too close to each other in the sense that

$$(\lambda_q^{(n)} - \lambda_{q+1}^{(n)}) = O(n^{-1/2})$$

as $n \to \infty$ and $\lambda_{q+1}^{(n)} = \ldots = \lambda_p^{(n)}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

We begin by examining the asymptotic behavior of the classical test $\phi^{(n)}$ for the problem studied in Schott (2006) and Virta (2021). We show that the test $\phi^{(n)}$, which is asymptotically equivalent to the Gaussian likelihood ratio test (LRT) for the equality of the smallest eigenvalues, behaves quite well against type-I alternatives but behaves poorly against alternatives of type II. Indeed, if $n^{1/2}(\lambda_q^{(n)} - \lambda_{q+1}^{(n)}) = O(1)$ as $n \to \infty$, the limiting power $\lim_{n\to\infty} \mathrm{E}[\phi^{(n)}]$ of the test is far below the asymptotic nominal level α. It follows directly that $\phi^{(n)}$ is unable to detect alternatives of weaker signals (alternatives of type II). The two main contributions of this study are as follows. First, we show that the test $\phi^{(n)}$ enjoys some local and asymptotic optimality properties when detecting type-I alternatives within a triangular array context. Second, we obtain tests for the problem that retain the aforementioned optimality properties, but can also detect alternatives of type II. The idea underpinning our new tests lies in the concept of *preliminary test estimators* studied by Saleh (2006) and Paindaveine et al. (2021). Our tests can be viewed as *preliminary test tests*, guided by the power enhancement principle studied recently in a high-dimensional setup by Fan et al. (2015) and Kock and Preinerstorfer (2019). We show using simulations that the estimator of the signal dimension based on $\phi^{(n)}$ studied in Nordhausen et al. (2022) can be improved using an estimator based on our new test.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

The rest of the paper is organized as follows. In Section 2, we present notation used in the rest of the paper and discuss the asymptotic equivalence between $\phi^{(n)}$ and the LRT for the equality of eigenvalues. In Sections 3 and 4, we study the asymptotic properties of $\phi^{(n)}$ against alternatives of type II and type I, respectively. In Section 5, we propose new tests for the problem, and show that the latter procedures enjoy many attractive properties. In Section 6, we demonstrate our method using a gene expression data set and discuss the problem of estimating the signal dimension. Additional Monte Carlo simulation results and technical details are contained in the Supplementary Material.

2. Testing the equality of eigenvalues

We consider triangular arrays of observations where the nth line of the array consists of i.i.d. observations $\mathbf{X}_{n1}, \ldots, \mathbf{X}_{nn}$ that follow a common Gaussian distribution with mean zero (without loss of generality, because in the Gaussian case, location and scatter parameters are "orthogonal"; e.g., see Hallin et al. (2010)) and covariance matrix $\mathbf{\Sigma}^{(n)}$ that admits the spectral decomposition

$$\mathbf{\Sigma}^{(n)} = \boldsymbol{\beta}\mathbf{\Lambda}^{(n)}\boldsymbol{\beta}' = \sum_{j=1}^{p} \lambda_j^{(n)} \boldsymbol{\beta}_j \boldsymbol{\beta}_j', \tag{2.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)$ is an orthogonal matrix and $\boldsymbol{\Lambda}^{(n)} = \mathrm{diag}(\lambda_1^{(n)}, \ldots, \lambda_p^{(n)})$ is a diagonal matrix of finite positive (well-ordered) eigenvalues. Throughout, $\mathrm{diag}(\mathbf{A}_1, \ldots, \mathbf{A}_m)$ denotes a block-diagonal matrix with blocks $\mathbf{A}_1, \ldots, \mathbf{A}_m$. We write $\mathrm{P}^{(n)}_{\boldsymbol{\beta}, \boldsymbol{\lambda}^{(n)}}$ for this Gaussian triangular array hypothesis, parametrized by $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}^{(n)} := (\lambda_1^{(n)}, \ldots, \lambda_p^{(n)})'$.

Fixing $0 \leq q < p-1$, we consider the testing problem characterized by sequences of null hypotheses of the form $\mathcal{H}_{0q}^{(n)}$ in (1.1), where for $q = 0$, $\lambda_0^{(n)}$ can be defined arbitrarily in such a way that $n^{1/2}(\lambda_0^{(n)} - \lambda_1^{(n)}) \to \infty$ as $n \to \infty$ so that, still for $q = 0$, the sequence of problems coincides with the full sphericity problem. We therefore tacitly assume that $\lambda_0^{(n)} = \lambda_1^{(n)} + 1$ throughout. When testing the equality of the smallest roots of a covariance matrix, the classical Gaussian LRT $\phi_{\mathrm{LRT}}^{(n)}$ rejects the null hypothesis at the asymptotic level α when $(d(p,q) := (p-q+2)(p-q-1)/2)$

$$L_q^{(n)} := -n \log\{ \prod_{j=q+1}^{p} \hat{\lambda}_j / ((p-q)^{-1} \sum_{j=q+1}^{p} \hat{\lambda}_j)^{p-q} \} > \chi^2_{d(p,q);1-\alpha}, \qquad (2.2)$$

where $\chi^2_{\nu;\delta}$ is the quantile of order δ of a chi-squared distribution with ν degrees of freedom, and $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ are the ordered eigenvalues of the empirical covariance matrix $\mathbf{S}^{(n)} := n^{-1} \sum_{i=1}^{n} \mathbf{X}_{ni} \mathbf{X}_{ni}'$; see, for instance, Muirhead (1982). Another classical test $\phi^{(n)}$ for the same problem rejects

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

the null hypothesis at the asymptotic level α when

$$T_q^{(n)} = \frac{n(\sum_{j=q+1}^{p} \hat{\lambda}_j^2 - (p-q)^{-1}(\sum_{j=q+1}^{p} \hat{\lambda}_j)^2)}{2((p-q)^{-1}\sum_{j=q+1}^{p} \hat{\lambda}_j)^2} > \chi^2_{d(p,q);1-\alpha}. \tag{2.3}$$

The test statistic $T_q^{(n)}$ is well known; Schott (2006) and Virta (2021) recently studied its high-dimensional properties. We have the following result (the proof follows directly from the proof of Theorem 5.1 in Tyler (1983)).

Lemma 1. *Let* $\mathbf{1}_p := (1, \ldots, 1)' \in \mathbb{R}^p$ *and*

$$\boldsymbol{\lambda}^{(n)} = (\lambda_1^{(n)}, \ldots, \lambda_q^{(n)}, \underline{\lambda}_{p-q}^{(n)} \mathbf{1}'_{p-q})', \tag{2.4}$$

where $\lambda_1^{(n)} \geq \ldots \geq \lambda_q^{(n)} \geq \underline{\lambda}_{p-q}^{(n)}$. *Then* $L_q^{(n)} - T_q^{(n)} = o_{\mathrm{P}}(1)$ *as* $n \to \infty$ *under* $\mathrm{P}^{(n)}_{\boldsymbol{\beta},\boldsymbol{\lambda}^{(n)}}$ *as* $n \to \infty$.

Lemma 1 shows that the Gaussian LRT $\phi^{(n)}_{\mathrm{LRT}}$ and the test $\phi^{(n)}$ enjoy a similar asymptotic behavior under $\mathrm{P}^{(n)}_{\boldsymbol{\beta},\boldsymbol{\lambda}^{(n)}}$, with $\boldsymbol{\lambda}^{(n)}$ as in (2.4). It follows directly from the definition of contiguity that their asymptotic behaviors also coincide under contiguous sequences. In particular, their local and asymptotic power coincide under contiguous alternatives of type I. Moreover, because the result obtained in Lemma 1 does not depend on the asymptotic behavior of $(\lambda_q^{(n)} - \underline{\lambda}_{p-q}^{(n)})$, the asymptotic behaviors of $\phi^{(n)}_{\mathrm{LRT}}$ and $\phi^{(n)}$ also coincide under alternatives of type II. In the rest of the paper, all asymptotic results for $\phi^{(n)}$ therefore also hold for $\phi^{(n)}_{\mathrm{LRT}}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

Our objective in the next two sections is to derive the asymptotic behavior of $\phi^{(n)}$ against both types of alternatives. We need the following notation: as usual, $\mathrm{vec}(\mathbf{A})$ stands for the vector obtained by stacking the columns of a matrix $\mathbf{A}$. Letting $\mathbf{A} \otimes \mathbf{B}$ stand for the Kronecker product between two matrices $\mathbf{A}$ and $\mathbf{B}$ ($\mathbf{A}^{\otimes 2} := \mathbf{A} \otimes \mathbf{A}$), the *commutation matrix* $\mathbf{K}_{k,\ell}$, such that $\mathbf{K}_{k,\ell}(\mathrm{vec}\,\mathbf{A}) = \mathrm{vec}(\mathbf{A}')$ for any $k \times \ell$ matrix $\mathbf{A}$, satisfies $\mathbf{K}_{p,k}(\mathbf{A} \otimes \mathbf{B}) = (\mathbf{B} \otimes \mathbf{A})\mathbf{K}_{q,\ell}$, for any $k \times \ell$ matrix $\mathbf{A}$ and $p \times q$ matrix $\mathbf{B}$; see, for example Magnus and Neudecker (2007). In the sequel we write $\mathbf{K}_k := \mathbf{K}_{k,k}$.

3. Asymptotic behavior against type-II alternatives

We now discuss the limiting behavior of $T_q^{(n)}$ (and therefore of $L_q^{(n)}$) under alternatives of type II. To do so, we consider sequences of models $\mathrm{P}^{(n)}_{\boldsymbol{\beta},\boldsymbol{\lambda}^{(n)}}$ such that the sequence $\boldsymbol{\lambda}^{(n)}$ provides alternatives of type II. Accordingly, the covariance matrix $\boldsymbol{\Sigma}^{(n)}$ in (2.1) has eigenvalues $\boldsymbol{\lambda}^{(n)} = (\lambda_1^{(n)}, \ldots, \lambda_p^{(n)})'$ of the form

$$\lambda_1^{(n)} := 1+r_1^{(n)}v_1 \geq \lambda_2^{(n)} := 1+r_2^{(n)}v_2 \geq \ldots \geq \lambda_q^{(n)} := 1+r_q^{(n)}v_q > \lambda_{q+1}^{(n)} = \ldots = \lambda_p^{(n)} = 1, \tag{3.1}$$

for some *rates* vector $\mathbf{r}^{(n)} := (r_1^{(n)}, \ldots, r_q^{(n)})'$ and some positive *localization* parameters $\mathbf{v} := (v_1, \ldots, v_q)'$, such that (3.1) holds for all n. More precisely,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

$r_j^{(n)}$ $(j = 1, \ldots, q)$ can be such that $r_j^{(n)} \equiv 1$ for all n, or such that $r_j^{(n)} \to 0$ as $n \to \infty$. Alternatives to $\mathcal{H}_{0q}^{(n)}$ of type II are such that $n^{1/2} r_q^{(n)}$ is $O(1)$ (and potentially $o(1)$) as $n \to \infty$. Note that the various tests compared here are clearly invariant with respect to scale transformations of the form $(\mathbf{X}_{n1}, \ldots, \mathbf{X}_{nn}) \to (s\mathbf{X}_{n1}, \ldots, s\mathbf{X}_{nn})$, for $s \in \mathbb{R}$. Thus, when we study the asymptotic behavior of $T_q^{(n)}$, or any other invariant test statistic, we can safely assume in our asymptotic analysis that the eigenvalues $\lambda_{q+1}^{(n)} = \ldots = \lambda_p^{(n)}$ in (3.1) are equal to one without loss of generality. As shown below, the asymptotic behavior of $T_q^{(n)}$ under $\mathrm{P}_{\boldsymbol{\beta}, \boldsymbol{\lambda}^{(n)}}^{(n)}$ with $\boldsymbol{\lambda}^{(n)}$ as in (3.1) depends on the rates in $\mathbf{r}^{(n)}$. We assume that the rates vector

$$\mathbf{r}^{(n)} = (\underbrace{r_1^{(n)}, \ldots, r_{s_1}^{(n)}}_{\text{block 1}}, \underbrace{r_{s_1+1}^{(n)}, \ldots, r_{s_2}^{(n)}}_{\text{block 2}}, \underbrace{r_{s_2+1}^{(n)}, \ldots, r_{s_3}^{(n)}}_{\text{block 3}}, \underbrace{r_{s_3+1}^{(n)}, \ldots, r_q^{(n)}}_{\text{block 4}})' \quad (3.2)$$

contains four blocks: in block 1, $r_j^{(n)}$ are all equal to one; in block 2, $r_j^{(n)}$ are $o(1)$ and $n^{1/2} r_j^{(n)} \to \infty$; in block 3, $r_j^{(n)} \equiv n^{-1/2}$; and in block 4, $r_j^{(n)}$ are $o(n^{-1/2})$. Of course, the blocks can be empty; for instance, $s_1 = 0$ indicates that the first block is empty, and block 2 is empty if $s_2 - s_1 = 0$, and so on. Under $\mathcal{H}_{0q}^{(n)}$, blocks 3 and 4 are empty. The setup is illustrated in Figure 1 below.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

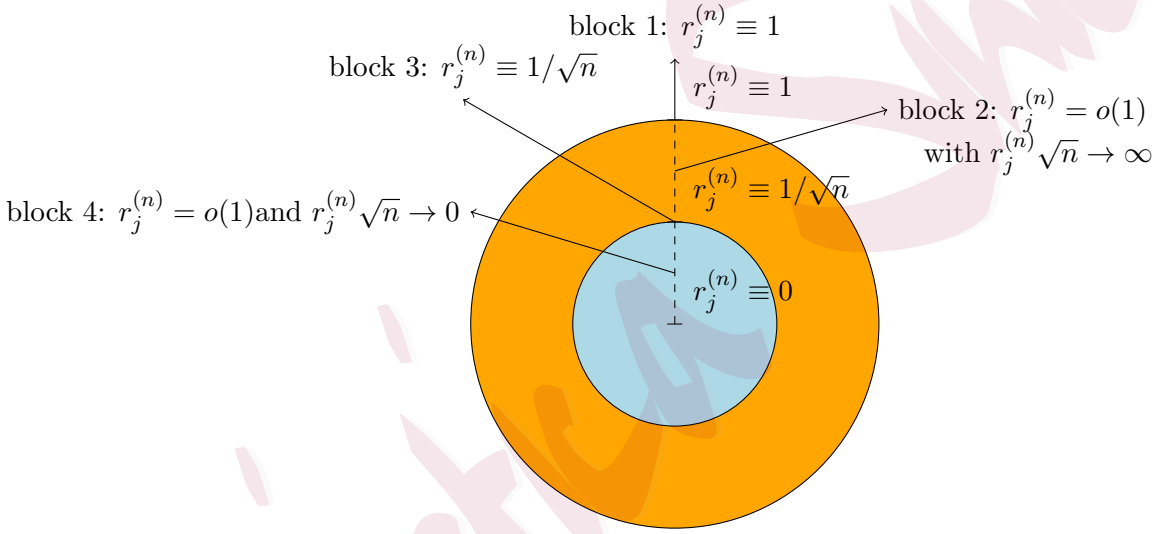


Figure 1: Illustration of how the eigenvalues are separated in blocks in the data-generating process.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

We have the following result.

Proposition 1. *Let* $\mathbf{r}^{(n)}$ *and* $\mathbf{v}$ *be such that* (3.1) *holds and such that for* $0 \le s_1 \le s_2 \le s_3 \le q$, *(i)* $r_j^{(n)} \equiv 1$ *for each* $1 \le j \le s_1$, *(ii)* $r_j^{(n)} = o(1)$ *with* $n^{1/2} r_j^{(n)} \to \infty$, *for each* $s_1 < j \le s_2$, *(iii)* $r_j^{(n)} = n^{-1/2}$, *for each* $s_2 < j \le s_3$ *and (iv)* $r_j^{(n)} = o(n^{-1/2})$, *for each* $s_3 < j \le q$. *Furthermore, let*

$$\mathbf{Z}(v_1, \ldots, v_{s_1}) = \begin{pmatrix} \mathbf{Z}_{11} & \mathbf{Z}_{21}' \\ \mathbf{Z}_{21} & \mathbf{Z}_{22} \end{pmatrix}$$

be a $p \times p$ *matrix, where* $\mathbf{Z}_{11}$ *is the* $s_2 \times s_2$ *upper-left block of* $\mathbf{Z}(v_1, \ldots, v_{s_1})$, $\mathbf{Z}_{22}$ *is the* $(p - s_2) \times (p - s_2)$ *lower-right block of* $\mathbf{Z}(v_1, \ldots, v_{s_1})$, *etc, such that*

$$\mathrm{vec}(\mathbf{Z}(v_1, \ldots, v_{s_1})) \sim \mathcal{N}_{p^2}(\mathbf{0}, (\mathbf{I}_{p^2} + \mathbf{K}_p)(\mathrm{diag}(1 + v_1, \ldots, 1 + v_{s_1}, \mathbf{1}_{p-s_1}'))^{\otimes 2}).$$

Then, as $n \to \infty$ *under* $\mathrm{P}^{(n)}_{\boldsymbol{\beta}, \boldsymbol{\lambda}^{(n)}}$ *with* $\boldsymbol{\lambda}^{(n)}$ *as in* (3.1), $T_q^{(n)}$ *converges weakly to*

$$\frac{1}{2}\Big(\sum_{j=q+1}^{p} \ell_j^2 - (p-q)^{-1}\Big(\sum_{j=q+1}^{p} \ell_j\Big)^2\Big), \tag{3.3}$$

where $(\ell_{q+1}, \ldots, \ell_p)$ *are the* $p - q$ *smallest roots of*

$$\mathbf{Z}_{22} + \mathrm{diag}(v_{s_2+1}, \ldots, v_{s_3}, \mathbf{0}_{q-s_3}', \mathbf{0}_{p-q}').$$

See the Supplementary Material for a proof. Proposition 1 states that the asymptotic behavior of $T_q^{(n)}$ depends crucially on the content of the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

various blocks in (3.2). In particular, under $\mathcal{H}_{0q}^{(n)}$, that is, if $s_1 \leq s_2 = q$ ($s_3 - s_2 = 0$), and thus the blocks 3 and 4 in (3.2) are empty (and therefore $n^{1/2} r_q^{(n)} \to \infty$ as $n \to \infty$), $\boldsymbol{\ell}_{p-q} := (\ell_{q+1}, \ldots, \ell_p)$ are the $p - q$ eigenvalues of the $(p-q) \times (p-q)$ matrix $\mathbf{Z}_{22}$ in Proposition 1. It is then easy to see that the resulting weak limit of $T_q^{(n)}$ is chi-squared with $d(p, q)$ degrees of freedom. It follows that the test $\phi^{(n)}$ is asymptotically valid for sequences of testing problems with null hypotheses $\mathcal{H}_{0q}^{(n)}$. If $n^{1/2} r_q^{(n)}$ does not diverge to ∞, that is, under alternatives of type II, the test statistic $T_q^{(n)}$ does not converge weakly to a chi-squared random variable with $d(p, q)$ degrees of freedom. Its asymptotic behavior is nevertheless completely characterized by Proposition 1. In Figure 2, we provide approximations of

$$\lim_{n\to\infty} \mathrm{E}[\phi^{(n)}] = \lim_{n\to\infty} \mathrm{P}[T_q^{(n)} > \chi^2_{d(p,q);1-\alpha}],$$

for $\alpha = .05$, $p = 8$ and various values of q under triangular arrays of observations with covariance $\boldsymbol{\Sigma}^{(n)}(b) = \mathrm{diag}((1 + n^{-b})\mathbf{1}_q, \mathbf{1}_{p-q})$, for $b = 0, 1/4, 1/2, 1$. For $b < 1/2$, the corresponding sequences of models belong to $\mathcal{H}_{0q}^{(n)}$, whereas for $b \geq 1/2$, the sequences of models are alternatives of type II. The approximations of $\lim_{n\to\infty} \mathrm{E}[\phi^{(n)}]$ are based on $100,000$ replications of the random variable in (3.3). Figure 2 clearly shows that the test $\phi^{(n)}$ is asymptotically valid for the problem at hand, but is blind to alternatives of type II. For $b \geq 1/2$, $\lim_{n\to\infty} \mathrm{E}[\phi^{(n)}]$ is far below the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

nominal level $\alpha = .05$. The results of Monte Carlo simulations, provided in the "Further simulations" section of the Supplementary Material clearly confirm the asymptotic behavior of $T_q^{(n)}$ obtained in Proposition 1. Two natural questions then arise. First, does the test $\phi^{(n)}$ enjoy some asymptotic optimality properties against local alternatives of type I (for any $\mathbf{r}_n$, such that $n^{1/2} r_q^{(n)} \to \infty$, not only in the classical $r_q^{(n)} \equiv 1$ case)? Second, the test $\phi^{(n)}$ clearly does not properly detect alternatives of type II; Figure 2 shows that the limiting power of $\phi^{(n)}$ against such alternatives can be almost zero. Thus, can we obtain tests that detect alternatives of type II, without losing too much power with respect to $\phi^{(n)}$ against local alternatives of type I?

4. Asymptotic behavior against type-I alternatives

In this section, we address the first of the two aforementioned questions by determining wether the test $\phi^{(n)}$ (and therefore $\phi_{\mathrm{LRT}}^{(n)}$) enjoys some optimality properties against alternatives of type I. Consider the $(p-q)$-dimensional observations

$$\mathbf{Y}_{ni} := (\boldsymbol{\beta}_{q+1}, \ldots, \boldsymbol{\beta}_p)' \mathbf{X}_{ni}, \quad i = 1, \ldots, n,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

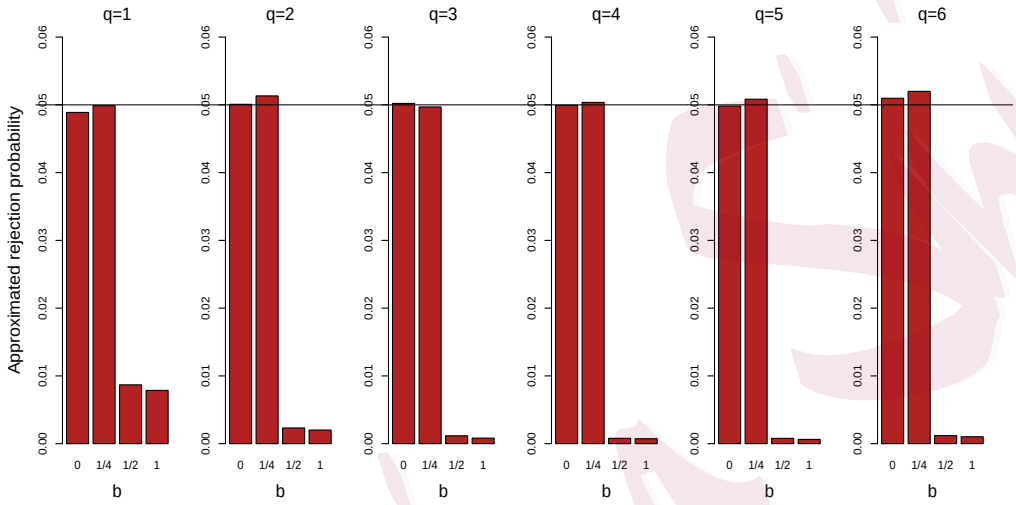


Figure 2: Approximations of $\lim_{n\to\infty} \mathrm{E}[\phi^{(n)}]$ for $p = 8$ and various values of q under triangular arrays of observations with covariance $\mathbf{\Sigma}^{(n)}(b) = \mathrm{diag}((1 + n^{-b})\mathbf{1}_q, \mathbf{1}_{p-q})$. The test $\phi^{(n)}$ is performed at the nominal level $\alpha = .05$. The approximation is based on $100,000$ replications of the random variable in (3.3).

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

obtained by selecting the last $p-q$ components of the rotated sample $\boldsymbol{\beta}'\mathbf{X}_{n1},\ldots,\boldsymbol{\beta}'\mathbf{X}_{nn}$ and define

$$\mathbf{S}_{\mathbf{Y}}^{(n)} := n^{-1}\sum_{i=1}^{n}\mathbf{Y}_{ni}\mathbf{Y}_{ni}' = (\boldsymbol{\beta}_{q+1},\ldots,\boldsymbol{\beta}_p)'\mathbf{S}^{(n)}(\boldsymbol{\beta}_{q+1},\ldots,\boldsymbol{\beta}_p),$$

where $\mathbf{S}^{(n)}$ (defined above (2.3)) is the empirical covariance matrix associated with the original sample. The $\mathbf{Y}_{ni}$ are i.i.d. with covariance matrix

$$\boldsymbol{\Sigma}_{\mathbf{Y}}^{(n)} := (\boldsymbol{\beta}_{q+1},\ldots,\boldsymbol{\beta}_p)'\boldsymbol{\Sigma}^{(n)}(\boldsymbol{\beta}_{q+1},\ldots,\boldsymbol{\beta}_p). \tag{4.1}$$

An asymptotically maximin test for the null hypothesis of sphericity $\mathcal{H}_0 : \boldsymbol{\Sigma}_{\mathbf{Y}}^{(n)} = \delta\mathbf{I}_{p-q}$, with $\delta > 0$, against contiguous local alternatives of type I has been proposed by Hallin and Paindaveine (2006). A test ϕ^* is called maximin in the class $\mathcal{C}_\alpha$ of level-α tests for a problem of testing some null hypothesis $\mathcal{H}_0$ against $\mathcal{H}_1$ if (i) ϕ^* has level α, and (ii) the power of ϕ^* is such that

$$\inf_{\mathrm{P}\in\mathcal{H}_1}\mathbb{E}_{\mathrm{P}}[\phi^*] \geq \sup_{\phi\in\mathcal{C}_\alpha}\inf_{\mathrm{P}\in\mathcal{H}_1}\mathbb{E}_{\mathrm{P}}[\phi].$$

Note that if $\boldsymbol{\lambda}^{(n)}$ belongs to $\mathcal{H}_{0q}^{(n)}$, $\boldsymbol{\lambda}^{(n)}+n^{-1/2}\boldsymbol{\ell}$ can only be an alternative of type I (and not of type II). The asymptotically maximin test against local alternatives of type I in Hallin and Paindaveine (2006), denoted here by

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

$\phi^{(n)}_{\boldsymbol{\beta}}$, rejects the null hypothesis at the asymptotic level α when

$$T^{(n)}_q(\boldsymbol{\beta}) \;=\; \frac{n}{2}\Big(\frac{p-q}{\mathrm{tr}(\mathbf{S}^{(n)}_{\mathbf{Y}})}\Big)^2\big(\mathrm{tr}((\mathbf{S}^{(n)}_{\mathbf{Y}})^2)-(p-q)^{-1}(\mathrm{tr}^2(\mathbf{S}^{(n)}_{\mathbf{Y}}))>\chi^2_{d(p,q);1-\alpha}. \tag{4.2}$$

Of course, in practice, the eigenvectors $\boldsymbol{\beta}_{q+1},\ldots,\boldsymbol{\beta}_p$ are rarely specified and, in general, need to be estimated. The most natural estimators of $\boldsymbol{\beta}_{q+1},\ldots,\boldsymbol{\beta}_p$ in the present Gaussian context are the eigenvectors $\hat{\boldsymbol{\beta}}_{q+1},\ldots,\hat{\boldsymbol{\beta}}_p$ associated with the $p-q$ smallest eigenvalues of

$$\mathbf{S}^{(n)} =: \sum_{j=1}^{p}\hat{\lambda}_j\hat{\boldsymbol{\beta}}_j\hat{\boldsymbol{\beta}}_j'.$$

Below, $\hat{\boldsymbol{\beta}} := (\hat{\boldsymbol{\beta}}_1,\ldots,\hat{\boldsymbol{\beta}}_p)$ stands for the $p\times p$ orthogonal matrix collecting the eigenvectors of $\mathbf{S}^{(n)}$. Plugging these estimators into $T^{(n)}_q(\boldsymbol{\beta})$ yields the test statistic $T^{(n)}_q$ in (2.3). Thus, to study the potential asymptotic equivalence between $T^{(n)}_q$ and $T^{(n)}_q(\boldsymbol{\beta})$, we need to control the asymptotic cost of the substitution of $\boldsymbol{\beta}_{q+1},\ldots,\boldsymbol{\beta}_p$ with $\hat{\boldsymbol{\beta}}_{q+1},\ldots,\hat{\boldsymbol{\beta}}_p$. Still in the same model, letting

$$\mathbf{E}^{(n)} = \begin{pmatrix}\mathbf{E}^{(n)}_{11} & \mathbf{E}^{(n)}_{12}\\ \mathbf{E}^{(n)}_{21} & \mathbf{E}^{(n)}_{22}\end{pmatrix} := \hat{\boldsymbol{\beta}}'\boldsymbol{\beta}, \tag{4.3}$$

where $\mathbf{E}^{(n)}_{11}$ and $\mathbf{E}^{(n)}_{22}$ are the $q\times q$ upper-diagonal and $(p-q)\times(p-q)$ lower-diagonal blocks, respectively, of $\mathbf{E}^{(n)}$, we have the following result.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

Proposition 2. *As $n \to \infty$ under $\mathrm{P}^{(n)}_{\boldsymbol{\beta},\boldsymbol{\lambda}^{(n)}}$ with $\boldsymbol{\lambda}^{(n)}$ as in* (3.1)*,*

(i) if $n^{1/2} r_q^{(n)} \to \infty$ as $n \to \infty$, $n^{1/2}\mathrm{diag}((\mathbf{r}^{(n)})')\mathbf{E}_{12}^{(n)} = O_{\mathrm{P}}(1)$ as $n \to \infty$ and $\mathbf{E}_{22}^{(n)}(\mathbf{E}_{22}^{(n)})' = \mathbf{I}_{p-q} + o_{\mathrm{P}}(1)$ as $n \to \infty$;

(ii) if $n^{1/2} r_q^{(n)} \to c < \infty$ as $n \to \infty$, we have that $\mathbf{E}_{12}^{(n)}$ is not $o_{\mathrm{P}}(1)$ as $n \to \infty$.

See the Supplementary Material for a proof. Proposition 2 shows that the consistency of the underlying eigenvectors can only hold when $n^{1/2}(\lambda_q^{(n)} - \lambda_{q+1}^{(n)})$ diverges to infinity as $n \to \infty$ with rates depending on $r_1^{(n)}, \ldots, r_q^{(n)}$. This fact naturally yields to the following question: are the tests $\phi_{\boldsymbol{\beta}}^{(n)}$ and $\phi^{(n)}$ asymptotically equivalent under $\mathcal{H}_{0q}^{(n)}$ (and therefore also under contiguous alternatives of type I)? The following result provides a positive answer.

Proposition 3. *Assume that $\boldsymbol{\lambda}^{(n)}$ as in* (3.1) *is such that $n^{1/2} r_q^{(n)} \to \infty$ as $n \to \infty$. Then, $T_q^{(n)} - T_q^{(n)}(\boldsymbol{\beta})$ is $o_{\mathrm{P}}(1)$ under $\mathrm{P}^{(n)}_{\boldsymbol{\beta},\boldsymbol{\lambda}^{(n)}}$ as $n \to \infty$.*

See the Supplementary Material for a proof. Proposition 3 states that the test $\phi^{(n)}$ and the test $\phi_{\boldsymbol{\beta}}^{(n)}$ are asymptotically equivalent under the null hypothesis $\mathcal{H}_{0q}^{(n)}$, and therefore also under contiguous alternatives. This shows directly that the three tests $\phi^{(n)}$, $\phi_{\mathrm{LRT}}^{(n)}$, and $\phi_{\boldsymbol{\beta}}^{(n)}$ enjoy the same asymptotic local power properties against the same contiguous alternatives

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

of type I. In particular they are locally and asymptotically maximin for the sphericity of $\mathbf{\Sigma}_{\mathbf{Y}}^{(n)}$ in (4.1), and therefore enjoy some local and asymptotic optimality property for detecting alternatives of type I. The results of Monte our Carlo simulations, provided in the "Further simulations" section of the Supplementary Material confirm the results presented in this section.

5. New tests

As shown in the previous sections, the test $\phi^{(n)}$ (and therefore $\phi_{\text{LRT}}^{(n)}$) enjoys some local and asymptotic optimality properties against alternatives of type I, but is blind to alternatives of type II. This is often problematic, because, in general, the purpose of this test is to provide information on the dimension of the underlying signal. Here, we propose tests that combine the properties of (i) being asymptotically equivalent to $\phi^{(n)}$ under $\mathcal{H}_{0q}^{(n)}$ (and therefore also under contiguous alternatives of type I) and (ii) being able to detect alternatives of type II. More precisely, we consider tests of the form

$$\phi_{\text{new}}^{(n)} := \mathbb{I}[T_q^{(n)} > \chi^2_{d(p,q);1-\alpha}]\mathbb{I}[T_{q,q+1}^{(n)} > \chi^2_{2;1-\gamma}] + \mathbb{I}[T_{q,q+1}^{(n)} \leq \chi^2_{2;1-\gamma}], \tag{5.1}$$

for $\alpha \in (0,1)$ and $\gamma \in (0,1)$, where

$$T_{q,q+1}^{(n)} := \frac{n(\sum_{j=q}^{q+1} \hat{\lambda}_j^2 - \frac{1}{2}(\sum_{j=q}^{q+1} \hat{\lambda}_j)^2)}{\frac{1}{2}(\sum_{j=q}^{q+1} \hat{\lambda}_j)^2} \tag{5.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

is a natural test statistic to test the equality of $\lambda_q^{(n)}$ and $\lambda_{q+1}^{(n)}$. Note that in (5.1), we take the convention $T_{0,1}^{(n)} \equiv +\infty$ so that, for testing $\mathcal{H}_{00}^{(n)}$, the tests $\phi_{\text{new}}^{(n)}$ and $\phi^{(n)}$ do coincide. The test $\phi_{\text{new}}^{(n)}$ can be viewed as a "preliminary test" test that rejects $\mathcal{H}_{0q}^{(n)}$ for large values of $T_q^{(n)}$, provided that $T_{q,q+1}^{(n)}$ is large enough, and also rejects when $T_{q,q+1}^{(n)}$ is too small. The idea underpinning this test lies in the concept of "preliminary test estimators" studied in Saleh (2006) and Paindaveine et al. (2021). We have the following result, obtained, without loss of generality, under sequences of models $\mathrm{P}_{\boldsymbol{\beta},\boldsymbol{\lambda}^{(n)}}^{(n)}$, with $\boldsymbol{\lambda}^{(n)}$ as in (3.1).

Proposition 4. *Assume that* $\boldsymbol{\lambda}^{(n)}$ *as in* (3.1) *is such that* $n^{1/2}r_q^{(n)} \to \infty$ *as* $n \to \infty$. *Then, under* $\mathrm{P}_{\boldsymbol{\beta},\boldsymbol{\lambda}^{(n)}}^{(n)}$, $\phi_{\text{new}}^{(n)} - \phi^{(n)}$ *is* $o_{\mathrm{P}}(1)$ *as* $n \to \infty$.

See the Supplementary Material for a proof. It follows directly from Proposition 4 that $\phi_{\text{new}}^{(n)}$ is asymptotically valid, because under $\mathcal{H}_{0q}^{(n)}$, $\lim_{n\to\infty} \mathrm{E}[\phi_{\text{new}}^{(n)}] = \alpha$. Moreover, $\phi_{\text{new}}^{(n)}$ inherits the local and asymptotic properties of $\phi^{(n)}$ under contiguous alternatives of type I. As shown below using simulations and as expected, the test $\phi_{\text{new}}^{(n)}$ shows far better power properties than $\phi^{(n)}$ against alternatives of type II. Indeed, assume that $\boldsymbol{\lambda}^{(n)}$ in (3.1) is such that it belongs to alternatives of type II with $n^{1/2}r_q^{(n)} \to 0$ as $n \to \infty$. Following the same rationale as in Section 3,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

because $\lim_{n\to\infty} \mathrm{P}^{(n)}_{\boldsymbol{\beta},\boldsymbol{\lambda}^{(n)}}\left[T^{(n)}_{q,q+1} > \chi^2_{2;1-\gamma}\right] \leq \gamma$, for $\gamma \in (0,1)$,

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{P}^{(n)}_{\boldsymbol{\beta},\boldsymbol{\lambda}^{(n)}}\left[\phi^{(n)}_{\text{new}} = 1\right] &\geq \lim_{n\to\infty} \mathrm{P}^{(n)}_{\boldsymbol{\beta},\boldsymbol{\lambda}^{(n)}}\left[T^{(n)}_{q,q+1} \leq \chi^2_{2;1-\gamma}\right] \\
&\geq 1 - \lim_{n\to\infty} \mathrm{P}^{(n)}_{\boldsymbol{\beta},\boldsymbol{\lambda}^{(n)}}\left[T^{(n)}_{q,q+1} > \chi^2_{2;1-\gamma}\right] \\
&\geq 1 - \gamma,
\end{aligned}$$

so that small values of γ necessarily result in large asymptotic power of $\phi^{(n)}_{\text{new}}$ against type II alternatives.

To illustrate the properties of the new tests, we perform Monte Carlo simulations. We generate $M = 2,000$ independent samples of i.i.d. observations

$$\mathbf{X}^{(b,\tau)}_1, \ldots, \mathbf{X}^{(b,\tau)}_n,$$

for $\tau = 0, 1, 2, 4, 6, 8$ and $b = 0, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2$. The $\mathbf{X}^{(b,\tau)}_i$ are i.i.d. with a common ($p =$)5-dimensional Gaussian distribution with mean zero and covariance matrix

$$\boldsymbol{\Sigma}(b,\tau) = \mathrm{diag}(3, 1 + n^{-b}, 1 + n^{-b}, 1, 1 - \frac{\tau}{n^{1/2}}).$$

We compare the classical test $\phi^{(n)}$ performed at the asymptotic nominal level $\alpha = .05$ with three versions of the $\phi^{(n)}_{\text{new}}$ test (all with $\alpha = .05$ in (5.1)) based on $\gamma = .9$, $\gamma = .5$, and $\gamma = .05$. All tests are performed for $\mathcal{H}^{(n)}_{03}$ ($q = 3$). The couples $(\tau, b) = (0, 0), (\tau, b) = (0, \frac{1}{8})$, and $(\tau, b) = (0, \frac{1}{4})$ provide data-generating processes under $\mathcal{H}^{(n)}_{03}$, while all other couples provide

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

data-generating processes under the alternative. In particular, the values $(\tau, b) = (0, \frac{1}{2}), (\tau, b) = (0, 1)$, and $(\tau, b) = (0, 2)$ provide alternatives that are purely of type II, and the couples (τ, b) with $\tau > 0$ and $b < 1/2$ provide alternatives that are purely of type I. In Figures 3 and 4, we provide the empirical rejection frequencies (out of the $2,000$ replications) of the four tests as functions of τ for sample sizes $n = 500$ and $n = 10,000$, respectively. The two figures show that the new tests $\phi^{(n)}_{\text{new}}$ behave as predicted by the asymptotic theory. They enjoy the same empirical power curves as $\phi^{(n)}$ when $\lambda^{(n)}_q$ is not too close to $\lambda^{(n)}_{q+1}$. Of course, there is some "continuity phenomenon" that implies that for finite samples, the nominal level constraint holds essentially for $(\tau, b) = (0, 0)$ and $(\tau, b) = (0, \frac{1}{8})$ only. The situation improves as n becomes larger, as shown in Figure 4. This is a finite-sample effect since, because, as explained below Proposition 4, $\phi^{(n)}_{\text{new}}$ is asymptotically valid. For large values of γ, the same "continuity phenomenon" is more pronounced, with a larger power enhancement. The new tests $\phi^{(n)}_{\text{new}}$ outperform $\phi^{(n)}$ in terms of detecting alternatives of type II, as expected.

6. Estimation of the signal dimension and a real-data application

In this Section, we demonstrate the usefulness of our method by applying it to the data used in Cho et al. (1998) on gene expressions. The data

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

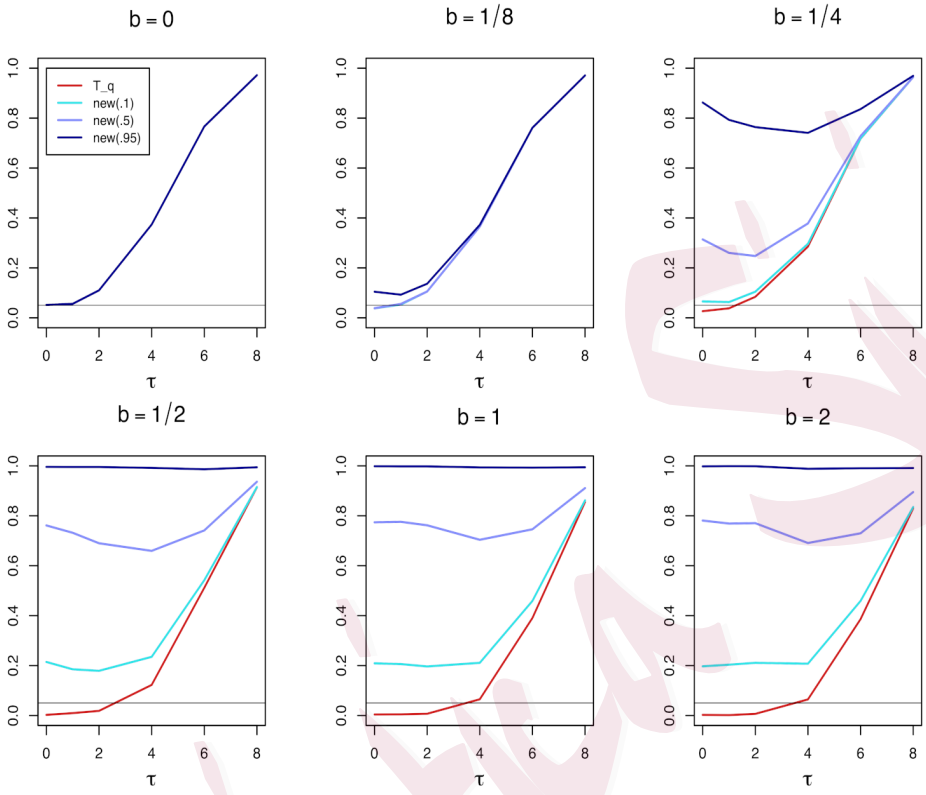


Figure 3: Empirical rejection frequencies of the classical $\phi^{(n)}$ performed at the asymptotic nominal level .05 and three versions of the $\phi^{(n)}_{\text{new}}$ test (all with $\alpha = .05$ in (5.1)), based on three choices of γ: $\gamma = .9$ (denoted as new(.1)), $\gamma = .5$ (denoted as new(.5)), and $\gamma = .05$ (denoted as new(.95)). The sample size is $n = 500$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

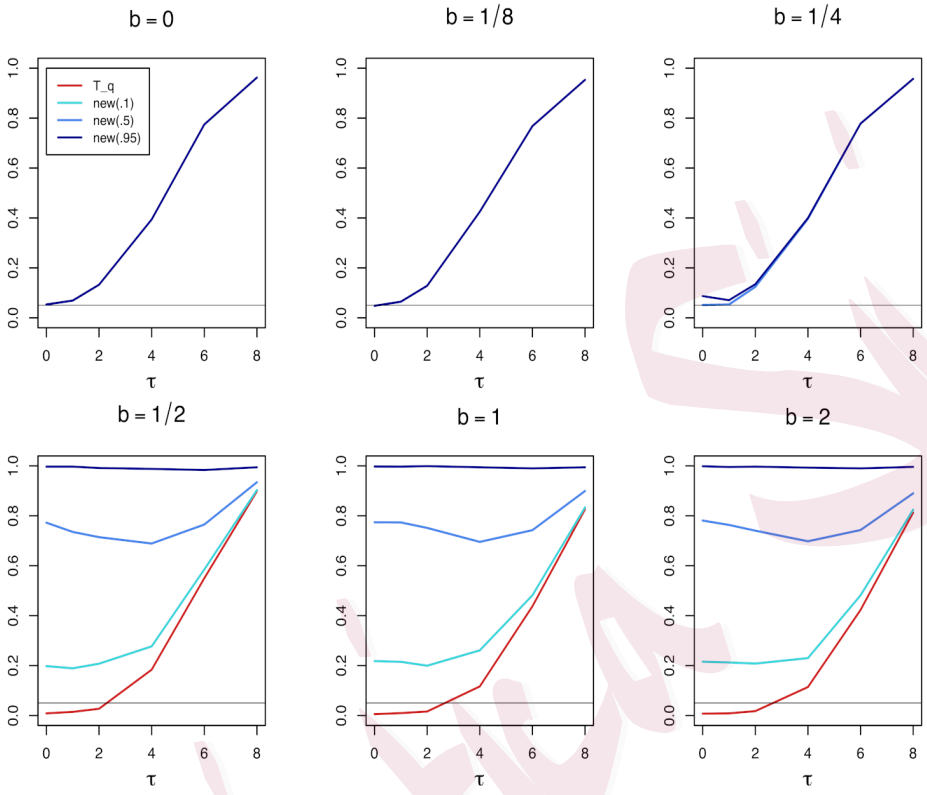


Figure 4: Empirical rejection frequencies of the classical $\phi^{(n)}$ performed at the asymptotic nominal level .05 and three versions of the $\phi^{(n)}_{\text{new}}$ test (all with $\alpha = .05$ in (5.1)), based on three different choices of γ: $\gamma = .9$ (denoted as new(.1)), $\gamma = .5$ (denoted as new(.5)), and $\gamma = .05$ (denoted as new(.95)). The sample size is $n = 10,000$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

set contains data on $n = 384$ gene expressions, measured at $p = 17$ time points, and is available online at http://faculty.washington.edu/kayee/pca/ . As explained in Cho et al. (1998), expression levels peak at different time points, corresponding to the five phases of cell cycles. The gene expressions are partitioned into five classes, corresponding to each phase of the cycle. Following Yeung and Ruzzo (2001), it is important to provide methods that cluster such data sets in order to recover the cell cycles. Following Yeung and Ruzzo (2001), a PCA is performed before clustering to reduce the noise level in the data. Then, the clustering is based on the noise-free data set. Deleting the noise is crucial in the Yeung and Ruzzo (2001) procedure.

We show how our tests can be used to construct an estimator of the signal dimension. The signal dimension k is the value $q \in \{0, \ldots, p-2\}$ for which $\mathcal{H}_{0q}^{(n)}$ holds. Note that if $\mathcal{H}_{0q}^{(n)}$ does not hold for any $q \in \{0, \ldots, p-2\}$, we then put $k = p - 1$; in such a case, the signal does not contain noise. As shown in Nordhausen et al. (2022), a consistent estimator $\hat{k}$ of k can be obtained as follows: letting $b_q^{(n)}$, $q = 0, \ldots, p-2$ be positive sequences such that $b_q^{(n)} \to \infty$ and $b_q^{(n)} = o(n)$ as $n \to \infty$ for all q, the estimator $\hat{k}$ is defined as

$$\hat{k} := \min\{q \in \{0, \ldots, p-2\},\ T_q^{(n)} < b_q^{(n)}\}, \tag{6.1}$$

with $\hat{k} := p - 1$ if the minimum above is not achieved. Using the test $\phi_{\text{new}}^{(n)}$,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

we define a new estimator of k as

$$\hat{k}_{\text{new}} = \min\{q \in \{0, \dots, p-2\}, \mathbb{I}[T_q^{(n)} > b_q^{(n)}]\mathbb{I}[T_{q,q+1}^{(n)} > c^{(n)}] + \mathbb{I}[T_{q,q+1}^{(n)} \leq c^{(n)}] = 0\}, \tag{6.2}$$

for some positive sequence $c^{(n)} \to \infty$ such that $c^{(n)} = o(n)$ as $n \to \infty$, and as $\hat{k}_{\text{new}} := p - 1$ if the minimum is not achieved. A consistency result for $\hat{k}_{\text{new}}$ is provided in the Supplementary Material. Here, we compare the small-sample properties of the estimators $\hat{k}$ and $\hat{k}_{\text{new}}$ using Monte Carlo simulations, before using them on the real data. We generate $M = 2,000$ independent samples of i.i.d. observations $\mathbf{X}_1^{(b,\tau^{(n)})}, \dots, \mathbf{X}_n^{(b,\tau^{(n)})}$ from a common ($p =$)3-dimensional Gaussian distribution with mean zero and covariance matrix $\mathbf{\Sigma}(b, \tau^{(n)}) = \text{diag}(1 + n^{-b}, 1, 1 - \tau^{(n)})$. We simulate observations with $\tau^{(n)} = 0, n^{-1/2}, .99$ and $b = 0, \frac{1}{2}, 1$. At each replication, we compute three versions of the estimator $\hat{k}$ in (6.1): one for each $b_q^{(n)} \in \{\log(n), \chi^2_{d(p,q),.95}, n^{1/2}\}$, $q = 0, \dots, p-2$. We also compute 12 versions of the estimator $\hat{k}_{\text{new}}$ in (6.2): one for each couple $(b_q^{(n)}, c^{(n)})$, with $b_q^{(n)} \in \{\log(n), \chi^2_{d(p,q),.95}, n^{1/2}\}$ and $c^{(n)} \in \{\chi^2_{2;.05}, \chi^2_{2;.1}, \chi^2_{2;.95}, n^{1/2}\}$. We compare the estimators with the true value of k, given by

$$k = (p-1)\mathbb{I}[\tau^{(n)} > 0] + (\mathbb{I}[b < \frac{1}{2}] + (p-1)\mathbb{I}[b \geq \frac{1}{2}])\mathbb{I}[\tau^{(n)} = 0].$$

In Figures 5, 6, and 7, we provide the frequencies (among the 2,000 replica-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

tions) of good selection of k for the various estimators. More explicitly, we compute the proportion of replications for which $\hat{k} = k$ and $\hat{k}_{\text{new}} = k$. The figures show that the new selectors perform equivalently to $\hat{k}$ when $b = 0$ (the two largest eigenvalues are sufficiently separated), and outperform $\hat{k}$ when $b > 0$. This is in line with the result that the tests associated with the selectors $\hat{k}_{\text{new}}$ perform better in terms of detecting alternatives of type II. When $\tau = .99$, the two smallest eigenvalues are strongly separated, and all estimators select the signal dimension perfectly.

In practice, the selection of $c^{(n)}$ and $b_q^{(n)}$ remains problematic. Virta and Nordhausen (2019) encountered a similar problem when selecting the $b_q^{(n)}$ used to compute the classical estimator $\hat{k}$. Our recommendation is similar to theirs: use $b_q^{(n)} = \chi^2_{d(p,q);1-\alpha}$ and $c^{(n)} = \chi^2_{2;1-\alpha}$ as default choices, for some reasonable α. This choice is in line with the discussion in Section 5 about the asymptotic power of $\phi^{(n)}_{\text{new}}$ under type-II alternatives.

The simulation results show that our estimator $\hat{k}_{\text{new}}$ performs quite well. We therefore use it to estimate the signal dimension of the log-transformed data set described earlier , comprising $n = 384$ gene expressions measured at $p = 17$ time points. Because we can question the Gaussianity in this practical example, we use estimators based on robustified versions of our test statistics, namely, the pseudo-Gaussian test statistics, in the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

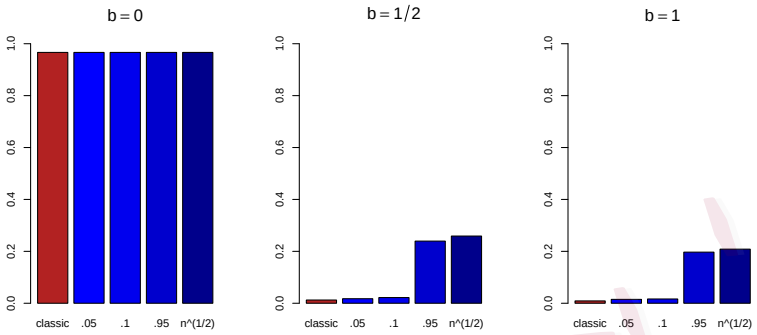
sense of Waternaux (1984) (see also Hallin et al. (2010)). The pseudo-Gaussian test statistics use estimated kurtosis coefficients to extend the asymptotic validity of parametric Gaussian procedures to the class of elliptical distributions (with finite moments of order four). They furthermore keep the local and asymptotic power properties of the same parametric Gaussian procedures in the Gaussian case. Letting $\hat{\kappa}^{(n)}$ be a consistent estimator of the underlying kurtosis parameter (see Waternaux (1984) for details), the pseudo-Gaussian test statistics are $\tilde{T}_q^{(n)} := (1+\hat{\kappa}^{(n)})^{-1} T_q^{(n)}$ and $\tilde{T}_{q,q+1}^{(n)} := (1+\hat{\kappa}^{(n)})^{-1} T_{q,q+1}^{(n)}$. We compute (pseudo-Gaussian versions of) $\hat{k}$ with $b_q^{(n)} \in \{\log(n), \chi^2_{d(p,q),.95}, n^{1/2}\}$, for $q = 0, \ldots, p-2$, and 12 (pseudo-Gaussian versions of) estimators $\hat{k}_{\text{new}}$, one for each couple $(b_q^{(n)}, c^{(n)})$, with $b_q^{(n)} \in \{\log(n), \chi^2_{d(p,q),.95}, n^{1/2}\}$ and $c^{(n)} \in \{\chi^2_{2;.05}, \chi^2_{2;.1}, \chi^2_{2;.95}, n^{1/2}\}$. In Figure 8, we provide the values taken by the various estimators. Figure 8 reveals that, although the small eigenvalues look close to each other, the data set does not contain much noise; the classical estimator $\hat{k}$ estimates the dimension of the signal at 13 or 14. Our new estimators with $c^{(n)} \in \{\chi^2_{2;.95}, n^{1/2}\}$ indicate that the data contain no noise. Given the performance of the various estimators on simulated examples, we suggest that every principal component should be considered significant and kept in the data set if the goal is to explain the maximal amount of variance possible. Because the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

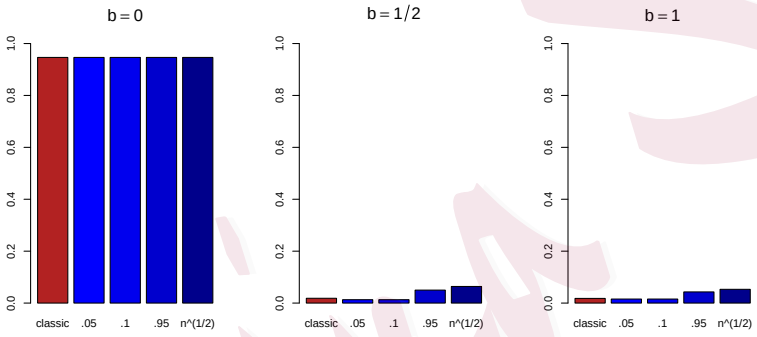
data set is noiseless, any dimension reduction based on a PCA will come at a cost of relevant information. There is no denoising step to conduct here, and any further dimension reduction technique should be performed on the full data set.

Statistica Sinica

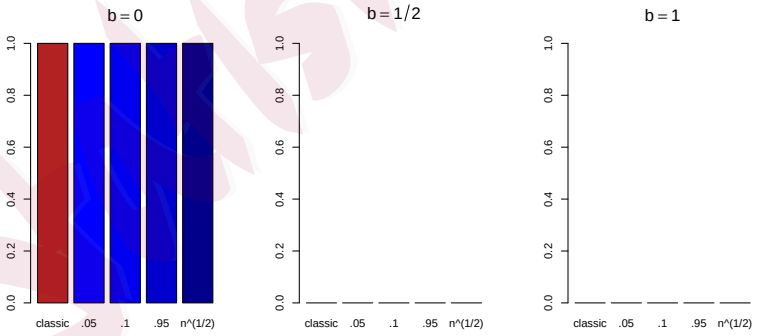
Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315



(a) $b_q^{(n)} \equiv \log(n)$



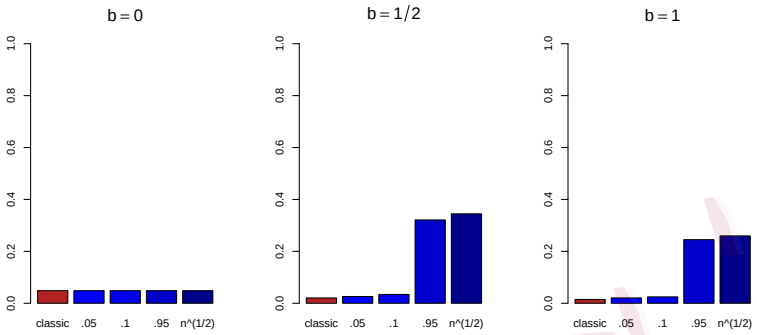
(b) $b_q^{(n)} \equiv \chi^2_{d(p,q),.95}$



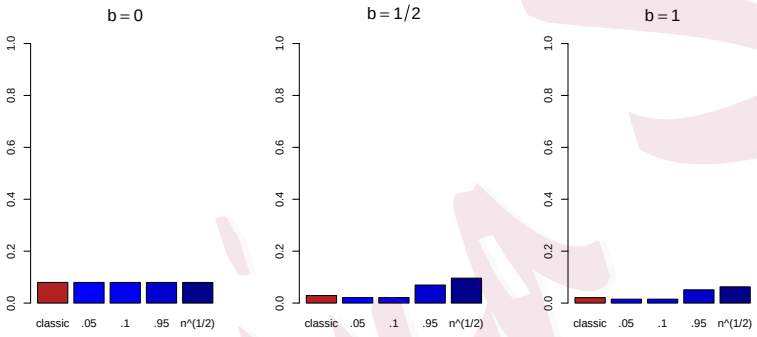
(c) $b_q^{(n)} \equiv n^{1/2}$

Figure 5: Proportion of good selection of k for three estimators $\hat{k}$ (in red, denoted as "classic") and for estimators $\hat{k}_{\text{new}}$ (in blue) with $b_q^{(n)} \in \{\log(n), \chi^2_{d(p,q),.95}, n^{1/2}\}$, and $c^{(n)} \in \{\chi^2_{2;.05}, \chi^2_{2;.1}, \chi^2_{2;.95}, n^{1/2}\}$. The sample size is $n = 1000$ and $\tau^{(n)} = 0$.

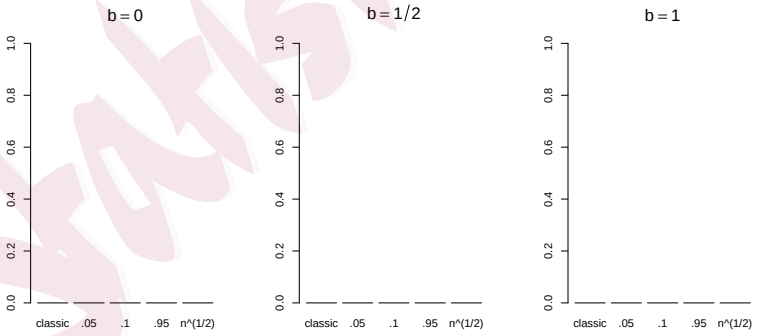
Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315



(a) $b_q^{(n)} \equiv \log(n)$



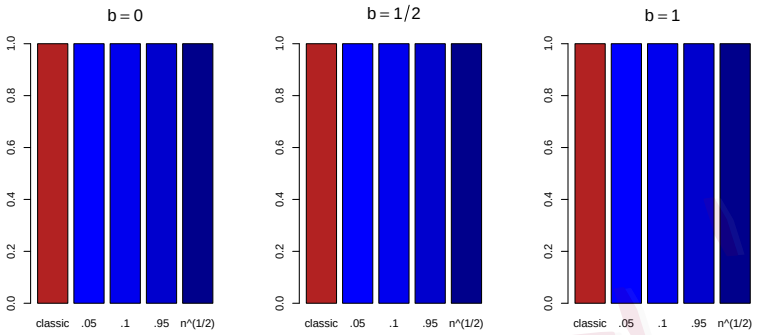
(b) $b_q^{(n)} \equiv \chi^2_{d(p,q),.95}$



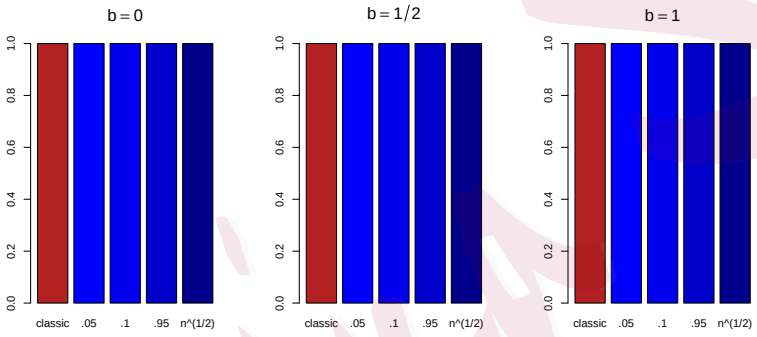
(c) $b_q^{(n)} \equiv n^{1/2}$

Figure 6: Proportion of good selection of k for three estimators $\hat{k}$ (in red, denoted as "classic") and for estimators $\hat{k}_{\text{new}}$ (in blue) with $b_q^{(n)} \in \{\log(n), \chi^2_{d(p,q),.95}, n^{1/2}\}$, and $c^{(n)} \in \{\chi^2_{2;.05}, \chi^2_{2;.1}, \chi^2_{2;.95}, n^{1/2}\}$. The sample size is $n = 1000$ and $\tau^{(n)} = n^{-1/2}$.

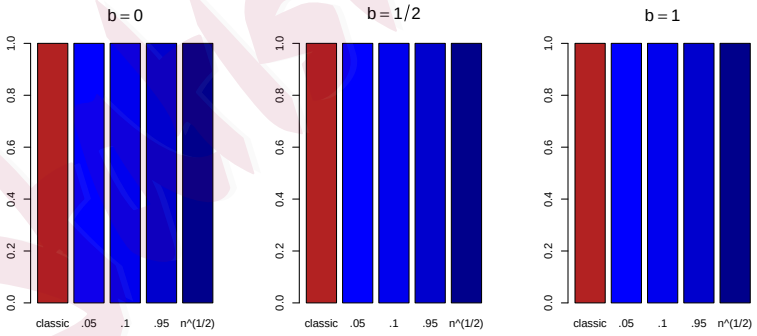
Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315



(a) $b_q^{(n)} \equiv \log(n)$



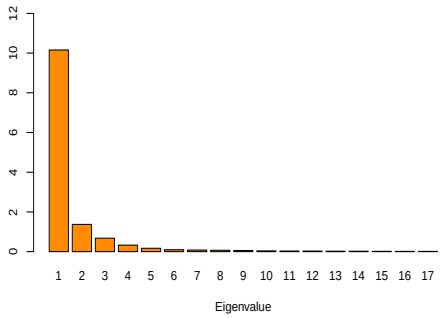
(b) $b_q^{(n)} \equiv \chi^2_{d(p,q),.95}$



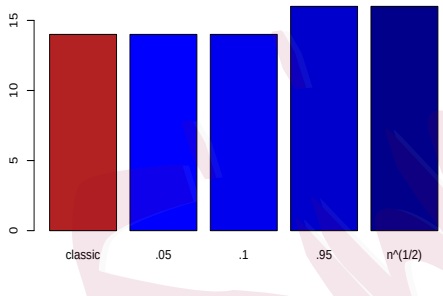
(c) $b_q^{(n)} \equiv n^{1/2}$

Figure 7: Proportion of good selection of k for three estimators $\hat{k}$ (in red, denoted as "classic") and for estimators $\hat{k}_{\text{new}}$ (in blue) with $b_q^{(n)} \in \{\log(n), \chi^2_{d(p,q),.95}, n^{1/2}\}$, and $c^{(n)} \in \{\chi^2_{2;.05}, \chi^2_{2;.1}, \chi^2_{2;.95}, n^{1/2}\}$. The sample size is $n = 1000$ and $\tau^{(n)} = .99$.

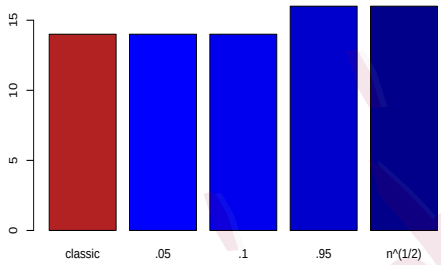
Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315



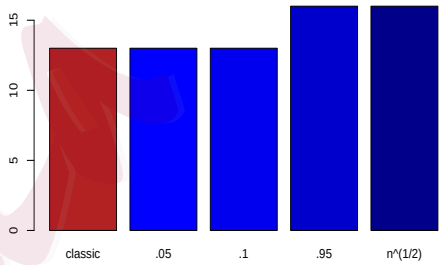
(a) Eigenvalues



(b) $b_q^{(n)} \equiv \log(n)$



(c) $b_q^{(n)} \equiv \chi^2_{d(p,q),.95}$



(d) $b_q^{(n)} \equiv n^{1/2}$

Figure 8: (a) the eigenvalues of the log-transformed gene expression data set; (b)-(d) the values (between zero and 16) taken by the estimators $\hat{k}$ (in red, denoted as "classic") for $b_q^{(n)} \in \{\log(n), \chi^2_{d(p,q),.95}, n^{1/2}\}$ and $\hat{k}_{\text{new}}$ (in blue) for $b_q^{(n)} \in \{\log(n), \chi^2_{d(p,q),.95}, n^{1/2}\}$ and $c^{(n)} \in \{\chi^2_{2;.05}, \chi^2_{2;.1}, \chi^2_{2;.95}, n^{1/2}\}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

7. Conclusion

We have studied procedures for testing problems characterized by null hypotheses of the form

$$\mathcal{H}_{0q}^{(n)} : (\lambda_{q+1}^{(n)} = \ldots = \lambda_p^{(n)}) \cap (n^{1/2}(\lambda_q^{(n)} - \lambda_{q+1}^{(n)}) \to \infty \text{ as } n \to \infty).$$

We have shown that $\phi^{(n)}$ (or, equivalently, $\phi_{\text{LRT}}^{(n)}$) enjoys some local and asymptotic optimality properties against alternatives of type I, but is blind to alternatives of type II. Our proposed tests for the problem that retain the local and asymptotic optimality properties of $\phi^{(n)}$ against alternatives of type I, and are able to detect alternatives of type II. In Proposition 2, we show that the consistency of an empirical projection on the first q principal axes can hold only if $n^{1/2}(\lambda_q^{(n)} - \lambda_{q+1}^{(n)})$ diverges to ∞ as $n \to \infty$. This makes $\mathcal{H}_{0q}^{(n)}$ a natural sequence of null hypotheses to test in order to perform an inference on the signal dimension. We then used our tests to build a new estimator of the signal dimension, which performs quite well, as shown in a simulation study.

Note that our asymptotic analysis concerns classical Gaussian estimators and tests. However, the same type of analysis will hold for robust tests built, for instance, on the eigenvalues of empirical robust scatter matrices $\mathbf{R}^{(n)}$ in setups in which the distribution of $\mathbf{U}^{(n)} := (\mathbf{\Sigma}^{(n)})^{-1/2} n^{1/2} (\mathbf{R}^{(n)} -$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

$\mathbf{\Sigma}^{(n)})(\mathbf{\Sigma}^{(n)})^{-1/2}$ is spherically symmetric, and the weak limit of $\text{vec}(\mathbf{U}^{(n)})$ is Gaussian. As explained in the real-data illustration, we can correct Gaussian LRTs using empirical kurtosis coefficients to obtain tests that are asymptotically valid under elliptical distributions with finite fourth-order moments; see, for instance, the pseudo-Gaussian tests in Waternaux (1984) and Hallin et al. (2010). Simulations illustrating the properties of these pseudo-Gaussian procedures in non-Gaussian settings are provided in the Supplementary Material.

Finally, in our study, the dimension p is fixed. It would be natural to extend our results to the high-dimensional case considered, for instance, in Forzani et al. (2017) and Virta (2021). This is left to future research.

Supplementary Material

The supplement contains various simulation studies to illustrate our results, all the technical proofs and a consistency result for the new estimator of the dimension of the signal.

Acknowledgments

Thomas Verdebout's research was supported by an ARC grant of the Communauté Française de Belgique and a Projet de Recherche (PDR) from the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

Fonds National de la Recherche Scientifique (FNRS).

References

Bao, Z., Ding, X., Wang, J. and Wang, K. (2022). Statistical inference for principal components of spiked covariance matrices. *Ann. Statist. 50*(2), 1144–1169.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *Br. J. Stat. Psychol. 3*(2), 77–85.

Chakraborty, L., Rus, H., Thistlethwaite, J. and Scott, D. (2020). A place-based socioeconomic status index: Measuring social vulnerability to flood hazards in the context of environmental justice. *Int. J. Disaster Risk Reduct. 43*.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L. et al. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. cell 2*(1), 65–73.

Cuesta-Albertos, J. A., Cuevas, A. and Fraiman, R. (2009). On projection-based tests for directional and compositional data. *Stat. Comput. 19*(4), 367–380.

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

Forzani, L., Gieco, A. and Carlos, T. (2017). Likelihood ratio test for partial sphericity in high and ultra-high dimensions. *J. Multivar. Anal. 159*, 18–38.

Francq, C., Jiménez-Gamero, M. and Meintanis, S. G. (2017). Tests for conditional ellipticity in multivariate garch models. *J. Econometrics 196*(2), 305–319.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

Hallin, M. and Paindaveine D. (2006). Semiparametrically efficient rank-based inference for shape. I. Optimal rank-based tests for sphericity. *Ann. Statist. 34*(6), 2707–2756.

Hallin, M., Paindaveine, D. and Verdebout, T. (2010). Optimal rank-based testing for principal components. *Ann. Statist. 38*(6), 3245–3299.

Henze, N., Hlávka, Z. and Meintanis, S. G. (2014). Testing for spherical symmetry via the empirical characteristic function. *Statistics 48*(6), 1282–1296.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist. 29*(2), 295–327.

Kock, A. B. and Preinerstorfer, D. (2019). Power in high-dimensional testing problems. *Econometrica 87*(3), 1055–1069.

Kritchman, S. and Nadler, B. (2009). Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Trans. Signal Process. 57*(10), 3930–3941.

Lawley, D. N. (1956). Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika 43*(1), 128–136.

Ledoit, O. and Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Ann. Statist. 30*(4), 1081–1102.

Li, Z., Han, F. and Yao, J. (2020). Asymptotic joint distribution of extreme eigenvalues and trace of large sample covariance matrix in a generalized spiked population model. *Ann.*

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

Statist. 48(6), 3138–3160.

Li, Z. and Yao, J. (2016). Testing the sphericity of a covariance matrix when the dimension is much larger than the sample size. *Electron. J. Statist. 10*(2), 2973–3010.

Luo, W. and Li, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika 103*(4), 875–887.

Magnus, J. R. and Neudecker, H. (2007). *Matrix Differential Calculus with Applications in Statistics and Econometrics* (3rd edition ed.). Chichester: John Wiley & Sons.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. New Jersey: John Wiley & Sons.

Nadler, B. (2010). Nonparametric detection of signals by information theoretic criteria: performance analysis and an improved estimator. *IEEE Trans. Signal Process. 58*(5), 2746–2756.

Nordhausen, K., Oja, H. and Tyler, D. E. (2022). Asymptotic and bootstrap tests for subspace dimension. *J. Multivar. Anal. 188*, 104830.

Onatski, A., Moreira, M. J. and Hallin, M. (2014). Signal detection in high dimension: the multispiked case. *Ann. Statist. 42*(1), 225–254.

Paindaveine, D., Rasoafaraniaina, J. and Verdebout, T. (2021). Preliminary multiple-test estimation, with applications to k-sample covariance estimation. *J. Am. Stat. Assoc. 117*(540), 1904–1915.

Paindaveine, D., Remy, J. and Verdebout, T. (2020a). Sign tests for weak principal directions.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

Bernoulli 26(4), 2987–3016.

Paindaveine, D., Remy, J. and Verdebout, T. (2020b). Testing for principal component directions under weak identifiability. *Ann. Statist. 48*(1), 324–345.

Paindaveine, D. and Verdebout, T. (2016). On high-dimensional sign tests. *Bernoulli 22*(3), 1745–1769.

Saleh, A. M. E. (2006). *Theory of preliminary test and Stein-type estimation with applications.* New Jersey: John Wiley & Sons.

Schott, J. R. (2006). A high-dimensional test for the equality of the smallest eigenvalues of a covariance matrix. *J. Multivar. Anal. 97*(4), 827–843.

Tian, X., Lu, Y. and Li, W. (2015). A robust test for sphericity of high-dimensional covariance matrices. *J. Multivar. Anal. 141*, 217–227.

Tyler, D. E. (1983). The asymptotic distribution of principal component roots under local alternatives to multiple roots. *Ann. Statist. 11*(4), 1232–1242.

Virta, J. (2021). Testing for subsphericity when n and p are of different asymptotic order. *Stat. Probab. Lett. 179*, 109209.

Virta, J. and Nordhausen, K. (2019). Estimating the number of signals using principal component analysis. *Stat 8*(1), e231.

Waternaux, C. M. (1984). Principal components in the nonnormal case: the test of equality of q roots. *J. Multivar. Anal. 14*(3), 323–335.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0315

REFERENCES

Yeung, K. Y. and Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics 17*(9), 763–774.

Şahan, C., Baydur, H. and Demiral, Y. (2019). A novel version of copenhagen psychosocial questionnaire-3: Turkish validation study. *Arch. Environ. Occup. Health 74*(6), 297–309.

Gaspard Bernard

Département de Mathématique and ECARES

Université libre de Bruxelles (ULB)

Campus Plaine, Boulevard du Triomphe, CP210

B-1050 Brussels, Belgium

E-mail: bernard.gaspard@ulb.be

Thomas Verdebout

Département de Mathématique and ECARES

Université libre de Bruxelles (ULB)

Campus Plaine, Boulevard du Triomphe, CP210

B-1050 Brussels, Belgium

E-mail: thomas.verdebout@ulb.be