<table>
<thead>
<tr>
<th>Title</th>
<th>Power Enhancement for Dimension Detection of Gaussian Signals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>SS-2022-0315</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://www.stat.sinica.edu.tw/statistica/">http://www.stat.sinica.edu.tw/statistica/</a></td>
</tr>
<tr>
<td>DOI</td>
<td>10.5705/ss.202022.0315</td>
</tr>
<tr>
<td>Complete List of Authors</td>
<td>Gaspard Bernard and Thomas Verdebout</td>
</tr>
<tr>
<td>Corresponding Authors</td>
<td>Thomas Verdebout</td>
</tr>
<tr>
<td>E-mails</td>
<td><a href="mailto:tverdebout@gmail.com">tverdebout@gmail.com</a></td>
</tr>
</tbody>
</table>

Notice: Accepted version subject to English editing.
Power enhancement for dimension detection of Gaussian signals

Gaspard Bernard and Thomas Verdebout

Université libre de Bruxelles (ULB)

Abstract: We consider in the present paper the classical problem of testing
$H_{0q}^{(n)} : \lambda_1^{(n)} > \lambda_2^{(n)} = \ldots = \lambda_{q+1}^{(n)}$, where $\lambda_1^{(n)}, \ldots, \lambda_p^{(n)}$ are the ordered latent roots of covariance matrices $\Sigma^{(n)}$. We show that the usual Gaussian procedure $\phi^{(n)}$ for this problem essentially shows no power against alternatives of weaker signals of the form $H_{1q}^{(n)} : \lambda_q^{(n)} = \lambda_{q+1}^{(n)} = \ldots = \lambda_p^{(n)}$. This is very problematic if the latter procedure 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 to detect alternatives to the equality of the $p-q$ smallest roots of $\Sigma^{(n)}$ provided that $\lambda_q^{(n)}$ and $\lambda_{q+1}^{(n)}$ are sufficiently separated. We obtain tests $\phi_{new}^{(n)}$ for the problem that keep 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 $H_{1q}^{(n)}$. Our results are illustrated via simulations and on a gene expression dataset from which we also discuss the problem of estimating the dimension of the signal.

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

Nowadays, dimension reduction plays a very important role in many applications. One of the most popular technique to perform unsupervised dimension reduction is Principal Component Analysis (PCA). The main objective of PCA is to extract a low-dimensional signal from the data. This can be achieved by first exhibiting a spiked structure in the underlying $p \times p$ positive definite covariance matrix $\Sigma$ using the data at hand. In the very popular spiked covariance models, the underlying covariance matrix $\Sigma$ 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 $\Sigma$ 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 very recently in Li et al. (2020), Paindaveine et al. (2020a,b) and Bao et al. (2022) to cite only a few; see also the references therein. In the context of spiked models or in PCA in general, a very important testing problem is therefore the problem of testing the equality of the $p - q$ smallest eigenvalues $\mathcal{H}_{0q} : \lambda_q > \lambda_{q+1} = \ldots = \lambda_p$ of $\Sigma$. Under $\mathcal{H}_{0q}$, the smallest $p - q$ eigenvalues are equal so that they correspond to some noise and as a result selecting more than $q$ principal components is useless. Tests for $\mathcal{H}_{0q}$ can typically be used before selecting the number of
components to keep. The problem is not new: in [Bartlett (1950)], tests for $H_{0q}$ are used to determine the numbers of significant factors over a dataset measuring reading speed, reading power, arithmetic speed and arithmetic power over 140 children. Tests for $H_{0q}$ as a tool for checking suitability of a dataset to factor analysis can also be more recently seen in [Sahan et al. (2019)], where the goal was to asses if a psychological questionnaire was consistent. In the same spirit, tests for $H_{0q}$ were used in [Chakraborty et al. (2020)] to ensure that every PCA-based sub-indicator was 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 $\lambda_0$ arbitrarily large) still attracts a lot of attention: [Ledoit and Wolf (2002)], [Onatski et al. (2014)], [Tian et al. (2015)], [Li and Yao (2016)] and [Paindaveine and Verdebout (2016)] tackled the problem in the high-dimensional case. [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 while [Francq et al. (2017)] considered the problem in a time series context to cite a few. Fixing $q < p - 1$, the problem of testing the equality of the smallest
$p - q$ eigenvalues $H_{0q} : \lambda_q > \lambda_{q+1} = \ldots = \lambda_p$ has also been investigated a lot in the multivariate statistics literature for the reasons described above. Methods for determining the dimension of the signal 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) while other procedures for the problem have been investigated in Nadler (2010), Luo and Li (2016) and Nordhausen et al. (2022) to cite a few. Finally, high-dimensional tests have been studied in Schott (2006) and more recently in Virta (2021).

In the present paper, our objective is to provide tests for $H_{0q} : \lambda_q > \lambda_{q+1} = \ldots = \lambda_p$ that are able to detect both alternatives of stronger signals under which $\lambda_{q+1}, \ldots, \lambda_p$ are not equal and alternatives of weaker signals under which $\lambda_q$ and $\lambda_{q+1}$ are "too close to each other". Note that the tests we obtain in this work for $H_{0q}$ can easily be adapted to tests for other restrictions like $\lambda_{q_1} > \lambda_{q_1+1} = \ldots = \lambda_{q_2} > \lambda_{q_2+1}$ for some $q_1$ and $q_2$. To properly formalize the problem, we consider a triangular array context in which the $n$th line of the array consists in i.i.d. $p$-variate Gaussian vectors $X_{1n}, \ldots, X_{nn}$ with common covariance matrix $\Sigma^{(n)} = \beta \Lambda^{(n)} \beta'$, where $\beta$ is
orthogonal and $\Lambda^{(n)} := \text{diag}(\lambda_1^{(n)}, \ldots, \lambda_p^{(n)})$ is a diagonal matrix of positive ordered eigenvalues that may change with $n$. Within such sequences of experiments, we consider the (sequence of) hypotheses testing problems characterized by null hypotheses of the form

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

Under $H_{0q}^{(n)}$, 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 $n \to \infty$. The sequence of hypotheses testing problems characterized by null hypotheses of the form $H_{0q}^{(n)}$ are well adapted to the practical objective which is the detection of the signal dimension. Indeed, a rejection of $H_{0q}^{(n)}$ can indicate 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 moreover that, as shown in the sequel, 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 $\infty$ as $n \to \infty$; this makes the testing problem associated with $H_{0q}^{(n)}$ in (1.1) a very natural problem to tackle in the context. Alternatives to $H_{0q}^{(n)}$ (for $q \geq 1$) are of two different natures:

(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 $n \to \infty$; we throughout call them
alternatives of type I;

(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 \( \lambda_q^{(n)} = \ldots = \lambda_p^{(n)} \); we throughout call them alternatives of type II.

We begin the paper with a study of 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 alternatives of type I but behaves extremely 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 test \( \phi^{(n)} \) behaves dramatically in the sense that its limiting power \( \lim_{n \to \infty} E[\phi^{(n)}] \) is far below the asymptotic nominal level \( \alpha \). It directly follows that the test \( \phi^{(n)} \) is unable to detect alternatives of weaker signals (alternatives of type II). The two main contributions of the paper are as follows. First, we show that the test \( \phi^{(n)} \) enjoys some local and asymptotic optimality properties to detect alternatives of type I within the present triangular array context. Second, we obtain tests for the problem
that keep the local and asymptotic optimality properties of $\phi(n)$ against alternatives of type I and that are able to detect alternatives of type II. The idea underpinning our new tests finds its roots in the concept of preliminary test estimators studied in Saleh (2006) and Paindaveine et al. (2021). Our tests can be seen as preliminary test tests guided by the power enhancement principle recently studied in a high-dimensional setup by Fan et al. (2015) and Kock and Preinerstorfer (2019). We show via 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.

The rest of the paper is organized as follows. In Section 2, we set out several notations 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 illustrate our method on a gene expression dataset and discuss the problem of estimating the signal dimension. Further Monte-Carlo simulation results and technical details are contained in the supplementary material.
2. Testing the equality of eigenvalues

As explained in the Introduction, we consider in this paper triangular arrays of observations where the \( n \)th line of the array consists in i.i.d. observations \( X_{n1}, \ldots, X_{nn} \) with a common Gaussian distribution with mean zero (without loss of generality since in the Gaussian case, location and scatter parameters are “orthogonal” as explained for instance in Hallin et al. (2010)) and covariance matrix \( \Sigma^{(n)} \) that admits the spectral decomposition

\[
\Sigma^{(n)} = \beta \Lambda^{(n)} \beta' = \sum_{j=1}^{p} \lambda_{j}^{(n)} \beta_{j} \beta_{j}',
\]

where \( \beta = (\beta_{1}, \ldots, \beta_{p}) \) is an orthogonal matrix and \( \Lambda^{(n)} = \text{diag}(\lambda_{1}^{(n)}, \ldots, \lambda_{p}^{(n)}) \) is a diagonal matrix of finite positive (well-ordered) eigenvalues; throughout, \( \text{diag}(A_{1}, \ldots, A_{m}) \) stands for the block-diagonal matrix with blocks \( A_{1}, \ldots, A_{m} \). In the sequel, we write \( P_{\beta, \lambda^{(n)}} \) for this Gaussian triangular array hypothesis parametrized by \( \beta \) and \( \lambda^{(n)} := (\lambda_{1}^{(n)}, \ldots, \lambda_{p}^{(n)})' \).

Fixing \( 0 \leq q < p - 1 \) we consider the testing problem characterized by sequences of null hypotheses of the form \( H_{0q}^{(n)} \) 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 throughout tacitly assume that \( \lambda_{0}^{(n)} = \)
\( \lambda_1^{(n)} + 1 \). For testing the equality of the smallest roots of a covariance matrix, the classical Gaussian LRT \( \phi_{\text{LRT}}^{(n)} \) rejects the null hypothesis at the asymptotic level \( \alpha \) when \( d(p,q) := (p - q + 2)(p - q - 1)/2 \)

\[
L_q^{(n)} := -n \log \left\{ \prod_{j=q+1}^{p} \hat{\lambda}_j/( (p-q)^{-1} \sum_{j=q+1}^{p} \hat{\lambda}_j)^{p-q} \right\} > \chi^2_{d(p,q);1-\alpha}.
\]

(2.2)

where \( \chi^2_{\nu,\delta} \) stands for the quantile of order \( \delta \) of a chi-square distribution with \( \nu \) degrees of freedom and \( \hat{\lambda}_1, \ldots, \hat{\lambda}_p \) are the ordered eigenvalues of the empirical covariance matrix \( S^{(n)} := n^{-1} \sum_{i=1}^{n} X_i X_i' \); see for instance in Muirhead (1982). Another classical test \( \phi^{(n)} \) for the same problem rejects the null hypothesis at the asymptotic level \( \alpha \) when

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

(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 \( 1_p := (1, \ldots, 1)' \in \mathbb{R}^p \) and

\[
\lambda^{(n)} = (\lambda_1^{(n)}, \ldots, \lambda_q^{(n)}, \Delta_{p-q}^{(n)} 1_{p-q}')',
\]

(2.4)

where \( \lambda_1^{(n)} \geq \ldots \geq \lambda_q^{(n)} \geq \Delta_{p-q}^{(n)} \). Then \( L_q^{(n)} - T_q^{(n)} = o_p(1) \) as \( n \to \infty \) under

\( \beta_{\lambda^{(n)}} \) as \( n \to \infty \).
Lemma 1 shows that the Gaussian LRT $\phi_{LRT}^{(n)}$ and the test $\phi^{(n)}$ enjoy a similar asymptotic behavior under $P_{\beta,\lambda}^{(n)}$ with $\lambda^{(n)}$ as in (2.4). It directly follows from the definition of contiguity that their asymptotic behaviors also coincide under contiguous sequences. In particular, their local and asymptotic powers coincide under contiguous alternatives of type I. Moreover, since the result obtained in Lemma 1 does not depend on the asymptotic behavior of $(\lambda_q^{(n)} - \lambda_{p-q}^{(n)})$, the asymptotic behaviors of $\phi_{LRT}^{(n)}$ and $\phi^{(n)}$ also coincide under alternatives of type II. In the rest of the paper, asymptotic results we will draw for $\phi^{(n)}$ therefore also hold for $\phi_{LRT}^{(n)}$.

Our objective in the next two Sections is therefore to derive the asymptotic behavior of $\phi^{(n)}$ against both types of alternatives. We will need the following notation: as usual $\text{vec}(A)$ stands for the vector obtained by stacking the columns of a matrix $A$. Letting $A \otimes B$ stand for the Kronecker product between two matrices $A$ and $B$ ($A \otimes^2 := A \otimes A$), the commutation matrix $K_{k,\ell}$ such that $K_{k,\ell}(\text{vec} A) = \text{vec}(A')$ for any $k \times \ell$ matrix $A$, satisfies $K_{p,k}(A \otimes B) = (B \otimes A)K_{q,\ell}$ for any $k \times \ell$ matrix $A$ and $p \times q$ matrix $B$; see, e.g., Magnus and Neudecker (2007). In the sequel we put $K_k := K_{k,k}$.
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 \( P_{\beta,\lambda}^{(n)} \) such that the sequence \( \lambda^{(n)} \) provides alternatives of type II. Accordingly the covariance matrix \( \Sigma^{(n)} \) in (2.1) has eigenvalues \( \lambda^{(n)} = (\lambda_1^{(n)}, \ldots, \lambda_p^{(n)})' \) of the form

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

for some rates vector \( r^{(n)} := (r_1^{(n)}, \ldots, r_q^{(n)})' \) and some positive localisation parameters \( v := (v_1, \ldots, v_q)' \) such that (3.1) holds for all \( n \). More precisely, \( r_j^{(n)} \) \((j = 1, \ldots, q)\) 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 \( \mathcal{H}_{qy}^{(n)} \) 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 in the paper are clearly invariant with respect to scale transformations of the form \( (X_{n1}, \ldots, X_{nn}) \to (sX_{n1}, \ldots, sX_{nn}), \ s \in \mathbb{R}, \) so that 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) can be taken equal to 1 without loss of generality. As we will see below, the asymptotic behavior of \( T_{q}^{(n)} \) under \( P_{\beta,\lambda}^{(n)} \) with \( \lambda^{(n)} \) as in (3.1)
crucially depends on the various rates in $r^{(n)}$. We will assume that the rates vector

$$r^{(n)} = (r^{(n)}_1, \ldots, r^{(n)}_{s_1}, r^{(n)}_{s_1+1}, \ldots, r^{(n)}_{s_2}, r^{(n)}_{s_2+1}, \ldots, r^{(n)}_{s_3}, \ldots, r^{(n)}_{q})^\prime$$

contains four blocks; in block 1, the $r^{(n)}_j$’s are all equal to 1; in block 2, the $r^{(n)}_j$’s are $o(1)$ and such that $n^{1/2}r^{(n)}_j \rightarrow \infty$; in block 3, $r^{(n)}_j \equiv n^{-1/2}$ and in block 4, the $r^{(n)}_j$’s are $o(n^{-1/2})$. Of course, the blocks can potentially be empty; for instance $s_1 = 0$ coincides with a situation where the first block is empty while block 2 is empty if $s_2 - s_1 = 0$, etc. Under $H^{(n)}_{0q}$, blocks 3 and 4 are empty. The setup is illustrated in Figure 1 below.

Figure 1: Illustration of the way eigenvalues are separated in various blocks in the data generating process.

We have the following result.
Proposition 1. Let $r^{(n)}$ and $v$ be such that (3.1) holds and such that for

$0 \leq s_1 \leq s_2 \leq s_3 \leq q$, (i) $r_j^{(n)} \equiv 1$ for each $1 \leq j \leq s_1$, (ii) $r_j^{(n)} = o(1)$ with $n^{1/2} r_j^{(n)} \to \infty$, for each $s_1 < j \leq s_2$, (iii) $r_j^{(n)} = n^{-1/2}$, for each $s_2 < j \leq s_3$ and (iv) $r_j^{(n)} = o(n^{-1/2})$, for each $s_3 < j \leq q$. Let furthermore

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

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

$\text{vec}(Z(v_1, \ldots, v_{s_1})) \sim \mathcal{N}_{p^2}(0, (I_{p^2} + K_p)(\text{diag}(1 + v_1, \ldots, 1 + v_{s_1}, 1_{p-s_1}'))^\otimes)$. 

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

\[
\frac{1}{2} \left( \sum_{j=q+1}^{p} \ell_j^2 - (p-q)^{-1} \left( \sum_{j=q+1}^{p} \ell_j \right)^2 \right),
\]

(3.3)

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

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

See the supplementary material for a proof. Proposition 1 entails that the asymptotic behavior of $T_q^{(n)}$ crucially depends on the content of the various blocks in (3.2). In particular, under $\mathcal{H}_{0q}^{(n)}$, that is if $s_1 \leq s_2 = q$
\( (s_3 - s_2 = 0) \), meaning that the blocks 3 and 4 in (3.2) are empty (and therefore \( n^{1/2}r_q^{(n)} \to \infty \) as \( n \to \infty \)), \( \ell_{p-q} := (\ell_{q+1}, \ldots, \ell_p) \) are the \( p-q \) eigenvalues of the \((p-q) \times (p-q)\) matrix \( Z_{22} \) in Proposition 1. It is then easy to see that the resulting weak limit of \( T_q^{(n)} \) is chi-square 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 \( \infty \), that is under alternatives of type II, the test statistic \( T_q^{(n)} \) does not converge weakly to a chi-square random variable with \( d(p,q) \) degrees of freedom. Its asymptotic behavior is nevertheless completely characterized by Proposition 1. In Figure 2 below, we provide approximations of

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

for \( \alpha = 0.05 \), \( p = 8 \) and various values of \( q \) under triangular arrays of observations with covariance \( \Sigma^{(n)}(b) = \text{diag}((1 + n^{-b})1_q, 1_{p-q}) \) with \( b = 0, 1/4, 1/2, 1 \). For \( b < 1/2 \), the corresponding sequences of models belong to \( \mathcal{H}_{0q}^{(n)} \) while for \( b \geq 1/2 \), the sequences of models are alternatives of type II. The approximations of \( \lim_{n \to \infty} E[\phi^{(n)}] \) are based on 100,000 replications of the random variable in (3.3). Inspection of Figure 2 clearly reveals that the test \( \phi^{(n)} \) is asymptotically valid for the problem at hand but is blind to alternatives of type II. For \( b \geq 1/2 \), \( \lim_{n \to \infty} E[\phi^{(n)}] \) is far below the nominal level \( \alpha = 0.05 \). Monte-Carlo simulations provided in the “Further simula-
tions” 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 $r_n$ such that $n^{1/2}r_q^{(n)} \rightarrow \infty$ and not only in the classical $r_q^{(n)} \equiv 1$ case)? Second, it directly follows from the results above that the test $\phi^{(n)}$ clearly does not detect properly alternatives of type II; Figure 2 shows that the limiting power of $\phi^{(n)}$ against such alternatives can almost be zero. A natural question is therefore: “can we obtain tests that are able to detect such 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 the present section, our objective is to answer the first question raised at the end of the previous section: “does the test $\phi^{(n)}$ (and therefore $\phi_L^{(n)}$) enjoy some optimality properties against alternatives of type I?”. Consider the $(p - q)$-dimensional observations

$$Y_{ni} := (\beta_{q+1}, \ldots, \beta_p)'X_{ni}, \quad i = 1, \ldots, n,$$
Figure 2: Approximations of $\lim_{n \to \infty} E[\phi(n)]$ for $p = 8$ and various values of $q$ under triangular arrays of observations with covariance $\Sigma^{(n)}(b) = \text{diag}((1 + n^{-b})1_q, 1_{p-q})$. 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).
obtained by selecting the last $p - q$ components of the rotated sample $\beta'X_{n1}, \ldots, \beta'X_{nn}$ and define

$$S_Y^{(n)} := n^{-1} \sum_{i=1}^{n} Y_{ni} Y_{ni}' = (\beta_{q+1}, \ldots, \beta_p)'S^{(n)}(\beta_{q+1}, \ldots, \beta_p),$$

where $S^{(n)}$ (defined above (2.3)) is the empirical covariance matrix associated with the original sample. The $Y_{ni}$'s are i.i.d. with covariance matrix

$$\Sigma_Y^{(n)} := (\beta_{q+1}, \ldots, \beta_p)'\Sigma^{(n)}(\beta_{q+1}, \ldots, \beta_p).$$

(4.1)

An asymptotically maximin test for the null hypothesis of sphericity $H_0 : \Sigma_Y^{(n)} = \delta I_{p-q}$, with $\delta > 0$ against contiguous local alternatives of type I has been obtained in Hallin and Paindaveine (2006). A test $\phi^*$ is called maximin in the class $C_\alpha$ of level–$\alpha$ tests for a problem of testing some null hypothesis $H_0$ against $H_1$ if (i) $\phi^*$ has level $\alpha$ and (ii) the power of $\phi^*$ is such that

$$\inf_{P \in H_1} \mathbb{E}_P[\phi^*] \geq \sup_{\phi \in C_\alpha} \inf_{P \in H_1} \mathbb{E}_P[\phi].$$

Note in passing that if $\lambda^{(n)}$ belongs to $H_{0q}^{(n)}$, $\lambda^{(n)} + n^{-1/2}\ell$ 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 $\phi^{(n)}_{\beta}$, rejects the null hypothesis at the asymptotic level $\alpha$
when

\[ T_q^{(n)}(\beta) = \frac{n}{2} \left( \frac{p-q}{\text{tr}(S_Y^{(n)})} \right)^2 (\text{tr}((S_Y^{(n)})^2) - (p-q)^{-1}\text{tr}^2(S_Y^{(n)})) > \chi^2_{d(p,q);1-\alpha}. \]  

(4.2)

Of course, in practice, the eigenvectors \( \beta_{q+1}, \ldots, \beta_p \) are rarely specified so that they generally need to be estimated. The most natural estimators of \( \beta_{q+1}, \ldots, \beta_p \) in the present Gaussian context are the eigenvectors \( \hat{\beta}_{q+1}, \ldots, \hat{\beta}_p \) associated with the \( p-q \) smallest eigenvalues of

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

Below, \( \hat{\beta} := (\hat{\beta}_1, \ldots, \hat{\beta}_p) \) stands for the \( p \times p \) orthogonal matrix collecting the eigenvectors of \( S^{(n)} \). Plugging in these estimators in \( T_q^{(n)}(\beta) \) yields to the test statistic \( T_q^{(n)} \) in [2.3]. To study the potential asymptotic equivalence between \( T_q^{(n)}(\beta) \) and \( T_q^{(n)}(\hat{\beta}) \), we therefore clearly need to control the asymptotic cost of the substitution of \( \beta_{q+1}, \ldots, \beta_p \) by \( \hat{\beta}_{q+1}, \ldots, \hat{\beta}_p \). Still in the same model, letting

\[ E^{(n)} = \begin{pmatrix} E_{11}^{(n)} & E_{12}^{(n)} \\ E_{21}^{(n)} & E_{22}^{(n)} \end{pmatrix} := \hat{\beta}' \beta, \]

(4.3)

where \( E_{11}^{(n)} \) and \( E_{22}^{(n)} \) are respectively the \( q \times q \) upper diagonal and \((p-q) \times (p-q)\) lower diagonal blocks of \( E^{(n)} \), we have the following result.
Proposition 2. As \( n \to \infty \) under \( P^{(n)}_{\beta, \lambda} \) with \( \lambda^{(n)} \) as in (3.1),

(i) if \( n^{1/2} r_q^{(n)} \to \infty \) as \( n \to \infty \), \( n^{1/2} \text{diag}((r^{(n)})') E_{12}^{(n)} = O_P(1) \) as \( n \to \infty \)

and \( E_{22}^{(n)}(E_{22}^{(n)})' = I_{p-q} + o_P(1) \) as \( n \to \infty \);

(ii) if \( n^{1/2} r_q^{(n)} \to c < \infty \) as \( n \to \infty \), we have that \( E_{12}^{(n)} \) is not \( o_P(1) \) as \( n \to \infty \).

See 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 \( \phi^{(n)}_{\beta} \) and \( \phi^{(n)} \) asymptotically equivalent under \( H_{0q}^{(n)} \) (and therefore also under contiguous alternatives of type I)? The following result provides a positive answer.

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

See the supplementary material for a proof. Proposition 3 entails that the test \( \phi^{(n)} \) and the test \( \phi_{\beta}^{(n)} \) are asymptotically equivalent under the null hypothesis \( H_{0q}^{(n)} \) and therefore also under contiguous alternatives. This directly entails that the three tests \( \phi^{(n)} \), \( \phi_{\text{LRT}}^{(n)} \) and \( \phi_{\beta}^{(n)} \) enjoy the same asymptotic local power properties against the same contiguous alternatives of type...
In particular, they are locally and asymptotically maximin for the sphericity of $\Sigma_{Y}^{(n)}$ in (4.1) and therefore enjoy some local and asymptotic optimality property to detect alternatives of type I. Monte-Carlo simulations provided in the “Further simulations” section of the supplementary material confirm the various results obtained in the present section.

5. New tests

As shown in the previous sections, the test $\phi^{(n)}$ (and therefore $\phi_{LRT}^{(n)}$) enjoys some local and asymptotic optimality properties against alternatives of type I but is totally blind to alternatives of type II. This can obviously be very often problematic since the purpose of this test is in general to provide information on the dimension of the underlying signal. We propose in the present section tests that combine the properties of (i) being asymptotically equivalent to $\phi^{(n)}$ under $H_{0q}^{(n)}$ (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

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

(5.1)

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

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

(5.2)
is a natural test statistic to test the equality of \( \lambda^{(n)}_q \) and \( \lambda^{(n)}_{q+1} \). Note that in (5.1), we take the convention \( T_{0,1}^{(n)} \equiv +\infty \) so that for testing \( \mathcal{H}^{(n)}_{00} \), the tests \( \phi^{(n)}_{\text{new}} \) and \( \phi^{(n)} \) do coincide. The test \( \phi^{(n)}_{\text{new}} \) can be seen as a “preliminary test” test that rejects \( \mathcal{H}^{(n)}_{0q} \) for large values of \( T_{q}^{(n)} \) provided that \( T_{q,q+1}^{(n)} \) is large enough and that also rejects when \( T_{q,q+1}^{(n)} \) is too small. The idea underpinning the preliminary test test above finds its roots 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 \( P^{(n)}_{\beta,\lambda^{(n)}} \) with \( \lambda^{(n)} \) as in (3.1).

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

See the supplementary material for a proof. It directly follows from Proposition 4 above that \( \phi^{(n)}_{\text{new}} \) is asymptotically valid since under \( \mathcal{H}^{(n)}_{0q} \),

\[
\lim_{n \to \infty} E[\phi^{(n)}_{\text{new}}] = \alpha.
\]

Moreover, \( \phi^{(n)}_{\text{new}} \) inherits the local and asymptotic properties of \( \phi^{(n)} \) under contiguous alternatives of type I. As shown below through simulations and as expected, the test \( \phi^{(n)}_{\text{new}} \) shows far better power properties than \( \phi^{(n)} \) against alternatives of type II. Indeed, assume that \( \lambda^{(n)} \) 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, since
\[
\lim_{n\to\infty} P_{\beta,\lambda}^{(n)}(n)[T_{q,q+1} \geq \chi^2_{2,1-\gamma}] \leq \gamma \quad \text{for } \gamma \in (0, 1),
\]

\[
\lim_{n\to\infty} P_{\beta,\lambda}^{(n)}(n)[\phi_{\text{new}}^{(n)} = 1] \geq \lim_{n\to\infty} P_{\beta,\lambda}^{(n)}(n)[T_{q,q+1} \leq \chi^2_{2,1-\gamma}]
\geq 1 - \lim_{n\to\infty} P_{\beta,\lambda}^{(n)}(n)[T_{q,q+1} > \chi^2_{2,1-\gamma}]
\geq 1 - \gamma,
\]

so that small values of \( \gamma \) necessarily yield to a large asymptotic power of \( \phi_{\text{new}}^{(n)} \) against type II alternatives.

To illustrate the properties of the new tests, we performed Monte-Carlo simulations. We generated \( M = 2,000 \) independent samples of i.i.d. observations

\[
X^{(b,\tau)}_1, \ldots, X^{(b,\tau)}_n,
\]

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

\[
\Sigma(b, \tau) = \text{diag}(3, 1 + b, 1 + n^{-b}, 1, 1 - \frac{\tau}{\sqrt{n}}).
\]

We compared the classical test \( \phi^{(n)} \) performed at the asymptotic nominal level \( \alpha = .05 \) with three versions of the \( \phi_{\text{new}}^{(n)} \) test (all with \( \alpha = .05 \) in (5.1)) based on three different choices of \( \gamma \): \( \gamma = .9, \gamma = .5 \) and \( \gamma = .05 \). All the tests are performed for \( H_{03}^{(n)}(q = 3) \). The couples \((\tau, b) = (0, 0), (\tau, b) = \ldots\)
(0, 1/8) and (τ, b) = (0, 1/2) provide data generating processes that are under $H_{03}^{(n)}$ while all other couples provide data generating processes under the alternative. In particular, the values (τ, b) = (0, 1/2), (τ, b) = (0, 1) and (τ, b) = (0, 2) provide alternatives that are purely of type II while the couples (τ, b) with τ > 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. Inspection of Figures 3 and 4 reveals that the new tests $\phi_{\text{new}}^{(n)}$ 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 entails that for finite samples, the nominal level constraint holds essentially for (τ, b) = (0, 0) and (τ, b) = (0, 1/8) only. The situation improves as n becomes larger as shown in Figure 4. This is a finite sample effect since; as explained below Proposition 4, $\phi_{\text{new}}^{(n)}$ is asymptotically valid. For large values of $\gamma$, the same “continuity phenomenon” is more pronounced with a larger power enhancement. The new tests $\phi_{\text{new}}^{(n)}$ are outperforming $\phi^{(n)}$ in the detection of alternatives of type II as expected.
Figure 3: Empirical rejection frequencies of the classical $\phi^{(n)}$ performed at the asymptotic nominal level .05 and three versions of the $\phi^{(n)}_{new}$ test (all with $\alpha = .05$ in (5.1)) based on three different choices of $\gamma$: $\gamma = .9$ (denoted as new(.1)), $\gamma = .5$ (denoted as new(.5)) and $\gamma = .05$ (denoted as new(.95)). The sample size is $n = 500$. 
Figure 4: Empirical rejection frequencies of the classical $\phi^{(n)}$ performed at the asymptotic nominal level .05 and three versions of the $\phi^{(n)}$ test (all with $\alpha = .05$ in (5.1)) based on three different choices of $\gamma$: $\gamma = .9$ (denoted as new(.9)), $\gamma = .5$ (denoted as new(.5)) and $\gamma = .05$ (denoted as new(.95)). The sample size is $n = 10,000$. 
6. Estimation of the signal dimension and real data application

In this Section, we illustrate the usefulness of our method on a real dataset used in Cho et al. (1998) involving gene expressions. This dataset that consists in $n = 384$ gene expressions measured at $p = 17$ time points is available online through the URL 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 are able to perform a clustering of such datasets in order to recover the cell cycles. In Yeung and Ruzzo (2001), PCA is used as preliminary step to clustering to reduce the noise level contained in the data. The idea is therefore to first perform a PCA to remove the noise in the data and then perform a clustering based on the noise-free dataset. Deleting the noise is crucial in the Yeung and Ruzzo (2001) procedure.

We show here 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 $\mathcal{H}_{0q}^{(n)}$ does hold. 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 then defined as

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

(6.1)

with \( \hat{k} := p - 1 \) if the minimum above is not achieved. Using the test \( \phi^{(n)}_{\text{new}} \), we define here a new estimator of \( k \) as

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

(6.2)

for some positive sequence \( c^{(n)} \to \infty \) such that \( c^{(n)} = o(n) \) as \( n \to \infty \) and with \( \hat{k}_{\text{new}} := p - 1 \) if the minimum is not achieved. While a consistency result for \( \hat{k}_{\text{new}} \) is provided in the supplementary material, we compare here the small-sample properties of the estimators \( \hat{k} \) and \( \hat{k}_{\text{new}} \) through Monte-Carlo simulations before using them on the real data. We generated \( M = 2,000 \) independent samples of i.i.d. observations \( X_1^{(b,\tau^{(n)})}, \ldots, X_n^{(b,\tau^{(n)})} \) with a common \((p = 3)\)-dimensional Gaussian distribution with mean zero and covariance matrix \( \Sigma(b, \tau^{(n)}) = \text{diag}(1 + n^{-b}, 1, 1 - \tau^{(n)}) \). We simulated observations with \( \tau^{(n)} = 0, n^{-1/2}, .99 \) and \( b = 0, \frac{1}{2}, 1 \). At each replication, we computed three versions of the estimator \( \hat{k} \) in (6.1); one for each \( b_q^{(n)} \in \{\log(n), \chi^2_{d(p,q), .95}, n^{1/2}\}, q = 0, \ldots, p - 2 \). We also computed twelve versions
of the estimator $\hat{k}_{\text{new}}$ in (6.2); one for each couple $(b^{(n)}_q, c^{(n)})$ with $b^{(n)}_q \in \{\log(n), \chi^2_{d(p,q),.95}, n^{1/2}\}$ and $c^{(n)} \in \{\chi^2_{2:.05}, \chi^2_{2:.1}, \chi^2_{2:.95}, n^{1/2}\}$ respectively. We compared the various estimators to the true value of $k$ given by

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

In Figures 5, 6 and 7 we provide the frequencies (among the 2,000 replications) of good selection of $k$ for the various estimators. More explicitly, we computed the proportion of replications for which $\hat{k} = k$ and $\hat{k}_{\text{new}} = k$ respectively. Inspection of Figures 5, 6 and 7 reveals that the new selectors perform equivalently well as $\hat{k}$ when $b = 0$ (the two largest eigenvalues are sufficiently separated) and outperforms $\hat{k}$ when $b > 0$. This is in line with the fact that the tests associated with the selectors $\hat{k}_{\text{new}}$ perform better in detecting alternatives of type II. When $\tau = .99$, the two smallest eigenvalues are strongly separated and all the estimators are selecting perfectly the signal dimension.

In practice, the selection of $c^{(n)}$ and $b^{(n)}_q$ remains an issue; a similar issue was encountered in Virta and Nordhausen (2019) for the selection of the $b^{(n)}_q$ used to compute the classical estimator $\hat{k}$. Our recommendation is similar to Virta and Nordhausen (2019)’s one: use $b^{(n)}_q = \chi^2_{d(p,q);1-\alpha}$ and $c^{(n)} = \chi^2_{2;1-\alpha}$ as default choices for some reasonable $\alpha$. 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 \( \hat{k}_{\text{new}} \) is performing quite well. We therefore used it to estimate the signal dimension of the log-transformed dataset described in the beginning of the section and that consists in \( n = 384 \) gene expressions measured at \( p = 17 \) time points. Since Gaussianity can be questioned in this practical real data illustration, we used estimators based on robustified versions of our test statistics, namely the pseudo-Gaussian test statistics in the sense of Water-\( \text{naux} \) (1984) (see also Hallin et al. (2010)). Such pseudo-Gaussian test statistics involve 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 Water-\( \text{naux} \) (1984) for details), 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 computed (pseudo-Gaussian versions of) \( \hat{k} \) with \( b_q^{(n)} \in \{\log(n), \chi^2_{d(p,q),.95}, n^{1/2}\} \), \( q = 0, \ldots, p - 2 \) and twelve (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 \( c^{(n)} \in \{\chi^2_{2,.05}, \chi^2_{2,.1}, \chi^2_{2,.95}, n^{1/2}\} \) respec-
tively. In Figure 8, we provide the values taken by the various estimators. Inspection of Figure 8 reveals that, although the small eigenvalues look very close to each other, the dataset 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 contains no noise. Given the performances of the various estimators on simulated examples, our suggestion here is that every principal component should be considered significant and kept in the dataset if the goal is to explain the maximal amount of variance possible. The dataset being noiseless, any dimension reduction based over PCA will come at a cost in terms of relevant information. There is no denoising step to conduct here and if any further dimension reduction technique would be applied, it should be performed on the all dataset.
Figure 5: Proportion of good selection of \(k\) for three estimator \(\hat{k}\) (in red, denoted as “classic”) and for estimators \(\hat{k}_{\text{new}}\) (in blue) with \(b_q(n) \in \{\log(n), \chi_{d(p,q),.95}^2, n^{1/2}\}\) and \(c(n) \in \{\chi_{2,.05}^2, \chi_{2,.1}^2, \chi_{2,.95}^2, n^{1/2}\}\). The sample size is \(n = 1000\) and \(\tau(n) = 0\).
Figure 6: Proportion of good selection of $k$ for three estimator $\hat{k}$ (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 $\tau^{(n)} = n^{-1/2}$.
Figure 7: Proportion of good selection of $k$ for three estimator $\hat{k}$ (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 $\tau^{(n)} = .99$. 
Figure 8: On the top left, the eigenvalues of the log transformed gene expression dataset. Then the values (between zero and 16) taken by the estimators $\hat{k}$ (in red, denoted as “classic”) for $b_q^{(n)} \in \{\log(n), \chi_{d(p,q),.95}^2, n^{1/2}\}$ and $\hat{k}_{\text{new}}$ in blue for $b_q^{(n)} \in \{\log(n), \chi_{d(p,q),.95}^2, n^{1/2}\}$ and $c^{(n)} \in \{\chi_{2,.05}^2, \chi_{2,.1}^2, \chi_{2,.95}^2, n^{1/2}\}$.
7. Conclusions

In the present paper, we studied procedures for the testing problems characterized by null hypotheses of the form

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

We showed that $\phi^{(n)}$ (or equivalently $\phi^{(n)}_{\text{LRT}}$) enjoys some local and asymptotic optimality property against alternatives of type I. However the latter is blind against alternatives of type II. We therefore built tests for the problem that keep the local and asymptotic optimality properties of $\phi^{(n)}$ against alternatives of type I and that are able to detect alternatives of type II. We also showed through Proposition 2 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 $\infty$ as $n \rightarrow \infty$. This makes $\mathcal{H}_{0q}^{(n)}$ a very natural sequence of null hypotheses to test in order to perform inference on the signal dimension. We then used our new tests to build a new estimator of the signal dimension which is performing quite well as shown in a simulation study.

To conclude, 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 eigenvalues of empirical robust scatter matrices $R^{(n)}$ in setups such that the distribution of $U^{(n)} := (\Sigma^{(n)})^{-1/2}n^{1/2}(R^{(n)} -$
\( \Sigma^{(n)}(\Sigma^{(n)})^{-1/2} \) is spherically symmetric and such that the weak limit of \( \text{vec}(U^{(n)}) \) is Gaussian. As explained in the real data illustration, it is for instance very classical to correct Gaussian LRTs with 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 note that, throughout the paper, the dimension \( p \) has been fixed. It would be natural to extend the results we obtained in the present work to the high-dimensional case considered for instance in Forzani et al. (2017) and Virta (2021). This is beyond the scope of the present paper but this will be the subject of a 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.
Acknowledgements

Thomas Verdebout’s research is supported by an ARC grant of the Communauté Française de Belgique and a Projet de Recherche (PDR) from the Fonds National de la Recherche Scientifique (FNRS).

References


REFERENCES


REFERENCES

Ledoit, O. and M. Wolf (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., F. Han, and J. Yao (2020). Asymptotic joint distribution of extreme eigenvalues and
trace of large sample covariance matrix in a generalized spiked population model. *Ann.
Statist.* 48(6), 3138–3160.

Li, Z. and J. Yao (2016). Testing the sphericity of a covariance matrix when the dimension is

Luo, W. and B. Li (2016). Combining eigenvalues and variation of eigenvectors for order


Sons.

Nadler, B. (2010). Nonparametric detection of signals by information theoretic criteria: perfor-


REFERENCES


REFERENCES


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
REFERENCES

B-1050 Brussels, Belgium

E-mail: thomas.verdebout@ulb.be