Statistics and Probability Letters 179 (2021) 109209

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Testing for subsphericity when n and p are of different asymptotic order

Joni Virta [a]

[a] Department of Mathematics and Statistics, University of Turku, Finland

ARTICLE INFO

Article history:
Received 29 January 2021
Received in revised form 30 June 2021
Accepted 23 July 2021
Available online 3 August 2021

Keywords:
Dimension estimation
High-dimensional statistics
PCA
Sample covariance matrix
Wishart distribution

ABSTRACT

We extend a test of subsphericity to the high-dimensional Gaussian regime where the spikes diverge to infinity and $p/n \to \{0, \infty\}$. The test is used to derive a consistent estimator for the latent dimension of the model.

1. Introduction

The objective of principal component analysis (PCA), and dimension reduction in general, is to extract a low-dimensional signal from noise-corrupted observed data. The most basic statistical model for the problem is as follows. Assume that S_n is the sample covariance matrix of a random sample from a p-variate normal distribution whose covariance matrix has the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d > \sigma^2, \ldots, \sigma^2$ exhibiting "spiked" structure. The data can thus be seen to be generated by contaminating a random sample residing in a d-dimensional subspace with independent normal noise having the covariance matrix $\sigma^2 I_p$. This signal subspace can be straightforwardly estimated with PCA as long as one knows its dimension d which is, however, usually unknown in practice. Numerous procedures for determining the dimension have been proposed, see Jolliffe (2002) for a review and, e.g., Schott (2006), Nordhausen et al. (2016) for asymptotic tests and Beran and Srivastava (1985), Luo and Li (2016) for bootstrap- and permutation-based techniques. Simplest of these methods is perhaps the test of sub-sphericity based on the test statistics $T_{n,j} = m_{2,p-j}(S_n)/m_{1,p-j}(S_n)^2 - 1$, $j = 0, \ldots, p-1$, where $m_{\ell,r}(A)$ denotes the ℓth sample moment of the last r eigenvalues of the symmetric matrix A. Under the null hypothesis $H_{0k} : d = k$ that the signal dimension equals k, the limiting null distribution of $T_{n,k}$ is

$$\frac{1}{2}n(p-k)T_{n,k} \rightsquigarrow \chi^2_{\frac{1}{2}(p-k)(p-k+1)-1}, \tag{1}$$

as $n \to \infty$, see, e.g., Schott (2006). Hence, the dimension d can in practice be determined by testing the sequence of null hypotheses $H_{00}, H_{01}, \ldots$ and taking the estimate of d to be the smallest k for which H_{0k} is not rejected. By examining the power of the tests, Nordhausen et al. (2016) concluded that this procedure yields a consistent estimate of d.

The previous test assumes a fixed dimension p and, in the face of modern large and noisy data sets with great room for dimension reduction, it is desirable to extend the test to the high-dimensional regime where $p = p_n$ is a function of n and we have $p_n \to \infty$ as $n \to \infty$. This is discussed in Section 2 where our first main contribution, extending the test

E-mail address: joni.virta@utu.fi.

https://doi.org/10.1016/j.spl.2021.109209

based on (1) to the high-dimensional regime where either the sample size or the dimension asymptotically dominates the other, is also presented. Section 3 introduces our second main contribution, a power study of the test, using which we construct a consistent estimator for the true latent dimension. In Section 4 we demonstrate our results using simulations and a real data example and, in Section 5, we finally conclude with some discussion. The proofs of the technical results are collected to the supplementary Appendix B and some supporting figures and tables are gathered to Appendix C.

2. High-dimensional testing of subsphericity

The behaviour of most high-dimensional statistical procedures depends crucially on the interplay between n and p_n and the most common approach in the literature is to assume that their growth rates are proportional in the sense that $p_n/n \to \gamma \in (0, \infty)$ as $n \to \infty$, see, e.g., Yao et al. (2015). The limiting ratio γ is also known as the *concentration* of the regime. In Schott (2006), the test of subsphericity discussed in Section 1 is extended to this asymptotic regime under the following two assumptions (note that in Assumption 2 the signal dimension d is a constant not depending on n).

Assumption 1. The observations $x_1, \dots, x_n$ are a random sample from $\mathcal{N}_{p_n}(\mu_n, \Sigma_n)$ for some $\mu_n \in \mathbb{R}^{p_n}$ and some positive-definite $\Sigma_n \in \mathbb{R}^{p_n \times p_n}$.

Assumption 2. The eigenvalues of the matrix Σ_n are $\lambda_{n1} \geq \cdots \geq \lambda_{nd} > \sigma^2 = \cdots = \sigma^2$ for some $\sigma^2 > 0$. Moreover, the eigenvalues λ_{nk}, $k = 1, \dots, d$, satisfy $\lambda_{nk} \to \infty$.

In fact, Schott (2006) additionally required that the quantities $\lambda_{nk}/\mathrm{tr}(\Sigma_n)$ converge to positive constants summing to less than unity, but applying our Lemma 1 in the proof of their Theorem 4 reveals that this condition is unnecessary, see Appendix A for details. Hence, denoting by S_n the sample covariance matrix of the observations, under Assumptions 1 and 2 and $\gamma \in (0, \infty) \setminus \{1\}$ (see Appendix A for more details on the exclusion of the case $\gamma = 1$), Theorem 4 in Schott (2006) establishes that the test statistic,

$$T_{n,j} := \frac{m_{2,p_n-j}(S_n)}{m_{1,p_n-j}(S_n)^2} - 1,$$

satisfies $(n - d - 1)T_{n,d} - (p_n - d) \rightsquigarrow \mathcal{N}(1, 4)$ where d is the signal dimension. As remarked by Schott (2006), this limiting result is consistent with its low-dimensional equivalent (1) in the sense that, as $p \to \infty$, we have that $\{2/(p-d)\}\chi^2_{\frac{1}{2}(p-d)(p-d+1)-1} - (p-d) \rightsquigarrow \mathcal{N}(1, 4)$.

A crucial condition that allows the above limiting result is the divergence of the spike eigenvalues $\lambda_{n1}, \dots, \lambda_{nd}$ of the covariance matrix to infinity in Assumption 2. Indeed, usually the spikes are taken to be constant in the literature for high-dimensional PCA, see, e.g. Baik and Silverstein (2006), Johnstone and Paul (2018). However, requiring the spikes to diverge is rather natural and reflects the idea that only a few PCs are sufficient to recover a large proportion of the total variance even in high dimensions. See, for example, Yata et al. (2018), who use cross-data-matrices to detect spiked principal components with divergent variance, and the references therein.

As our first contribution, we extend the result of Schott (2006) outside of the regime $p_n/n \to \gamma \in (0, \infty)$, to the extreme cases $\gamma \in \{0, \infty\}$. The latter have been less studied in the high-dimensional literature, but see, for example, Karoui (2003), Birke and Dette (2005), Yata and Aoshima (2009), Jung and Marron (2009), the last of which consider the extreme asymptotic scenario where the dimension diverges to infinity but the sample size remains fixed. In our treatment of the case $\gamma = \infty$, we further require the additional condition that $p_n/(n\sqrt{\lambda_{nd}}) \to 0$ as $n \to \infty$, i.e., the dimension must not diverge too fast compared to the sample size and the magnitude of the spike λ_{nd} corresponding to the weakest signal. Assumptions of this form are rather common in high-dimensional PCA when the spikes are taken to diverge, see, e.g., Shen et al. (2016) who saw n, λ_{nk} and p_n as three competing forces affecting the consistency properties of PCA, n and λ_{nk} contributing information about the signals and p_n decreasing the relative share of information in the sample by introducing more noise to the model. The condition $p_n/(n\sqrt{\lambda_{nd}}) \to 0$ can thus be interpreted as requiring that even the weakest of the spike principal components has asymptotically strong enough signal to be detected.

The extension of the test to the previous regimes is given below in Theorem 1. The main line of proof is based on extending the work of Birke and Dette (2005), who considered sphericity tests for $\gamma \in \{0, \infty\}$, to testing of subsphericity. In this sense, our work is to Birke and Dette (2005) what Schott (2006) is to Ledoit and Wolf (2002), who studied tests of sphericity under $\gamma \in (0, \infty)$ and on whose work Schott (2006) based their proof.

Theorem 1. *Under Assumptions 1 and 2, if, as $n \to \infty$, either*

(i) $p_n/n \to 0$, *or,*
(ii) $p_n/n \to \infty$ *and* $p_n/(n\sqrt{\lambda_{nd}}) \to 0$, *then,*

$$(n - d - 1)T_{n,d} - (p_n - d) \rightsquigarrow \mathcal{N}(1, 4).$$

3. Power analysis and dimension estimation

A natural question is whether the test of subsphericity can be used to consistently estimate the latent dimension d under our model. In a low-dimensional setting, this is accomplished by chaining together tests for $H_{0k} : d = k$ for different values of k in some specific order. E.g., in forward testing one sequentially tests for $H_{00}, H_{01}, \ldots$ and takes as the estimate of d the smallest k for which H_{0k} is not rejected. Various other strategies are also available but in the high-dimensional setting where our working assumption is that the number of latent signals is small compared to the overall dimensionality (finite d vs. $p_n \to \infty$), the forward testing is likely to be the most economic choice. In the following we show that this strategy indeed leads, under suitable assumptions, to a consistent estimate of the dimension d in various high-dimensional regimes. Even though the equivalent of Theorem 1 for $\gamma \in (0, \infty) \setminus \{1\}$ was established already in Schott (2006), the following results are novel also in that case. We use the notation $g_{n,k} := (n - k - 1)T_{n,k} - (p_n - k)$, $k = 0, \ldots, p_n - 1$, for the test statistic.

Theorem 2. *Under Assumptions 1 and 2, if, as $n \to \infty$, either*

(i) $p_n/n \to \gamma \in [0, \infty) \setminus \{1\}$ and $p_n/\lambda_{nd}^2 \to 0$, or,
(ii) $p_n/n \to \infty$, $p_n/(n\sqrt{\lambda_{nd}}) \to 0$ and $p_n/(\sqrt{n}\lambda_{nd}) \to 0$, then,

we have, for each $k = 0, \ldots, d - 1$ and for all $M > 0$, that

$$\mathbb{P}(g_{n,k}/n \leq M) \to 0.$$

Theorem 2 shows that the test for H_{0k} is consistent under the alternative hypothesis that the true dimension $d > k$ (the power of the test in the case $d < k$ plays no role in the forward testing and, hence, is not studied here). As a corollary we then obtain the consistency of the forward testing.

Corollary 1. *Under the assumptions of Theorem 2, let c_n be any sequence of real numbers satisfying $c_n \to \infty$ and $c_n = \mathcal{O}(n)$ as $n \to \infty$. Then,*

$$\hat{d} := \min\{k = 0, \ldots, p_n - 1 : g_{n,k} \leq c_n\} \to_p d.$$

Choosing a sequence c_n for which the forward testing estimator $\hat{d}$ performs well in finite samples is a highly non-trivial task and, thus, we advocate using in practice the alternative estimator,

$$\hat{d} := \min\{k = 0, \ldots, p_n - 1 : |(g_{n,k} - 1)/2| \leq z_{1-\alpha/2}\}, \tag{2}$$

where $z_{1-\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution, see, e.g., Nordhausen et al. (2016) for a similar modification. The resulting procedure has asymptotically zero probability to underestimate the dimension (by Theorem 2) and carries the Type I error probability equal to α of overestimating the dimension (by Theorem 1).

Finally, we still briefly discuss the assumptions of Corollary 1 which, while stricter than in Theorem 1, can nevertheless be seen to be very natural. That is, regardless of the regime, the assumptions ask that the weakest of the signals is strong enough not to be masked by the noise (similarly as in part (ii) of Theorem 1). To gain a more concrete idea on the severity of the assumptions, let $p_n = cn^\alpha$ and $\lambda_{nd} = n^\beta$ for some $c \neq 1$ and $\alpha, \beta > 0$. Then, the feasible values of (α, β) form a polygon in $\mathbb{R}^2$ that is illustrated in the range $0 < \alpha \leq 2$ as the grey area in Fig. C.1 in Appendix C. The plot reveals the intuitive fact that the effect of the dimension on the minimal feasible growth rate for the signal is the stronger the faster p_n increases (the slope of the curve is for $\alpha > 1.5$ four times higher than for $\alpha \in (0, 1)$).

4. Numerical examples

We first demonstrate the result of Theorem 1 using simulated data. We consider four different settings, each of which assumes a sample of size n from $\mathcal{N}_{p_n}(0, \Sigma_n)$ where $\Sigma_n = \text{diag}(\lambda_{n1}, \ldots, \lambda_{nd}, 1, \ldots, 1)$. Note that this simplified form of the normal distribution (zero location, unit noise variance and diagonal covariance) is without loss of generality as our test statistic is location, scale and rotation invariant. The settings are as follows:

1. $d = 3$, $n = 216$, $p_n = n^{3/4}$, $\lambda_{n1} = 3n$, and $\lambda_{n2} = \lambda_{n3} = n^{1/2}$,
2. $d = 3$, $n = 216$, $p_n = n^{3/4}$, $\lambda_{n1} = 3n^{1/2}$, and $\lambda_{n2} = \lambda_{n3} = n^{1/4}$,
3. $d = 2$, $n = 36$, $p_n = n^{3/2}$, $\lambda_{n1} = 2n^2$ and $\lambda_{n2} = n^{3/2}$,
4. $d = 2$, $n = 36$, $p_n = n^{3/2}$, $\lambda_{n1} = 2n^2$ and $\lambda_{n2} = n^{1/4}$.

Settings 1 and 2 fall within the case $\gamma = 0$, and their only difference is in the growth rates of the spikes. Settings 3 and 4 explore the case $\gamma = \infty$, the former satisfying the conditions of Theorem 1 and the latter not (again the only difference between them is in the growth rates of the spikes). In each case, we compute 10000 replicates of the test statistic $g_{n,d} = (n - d - 1)T_{n,d} - (p_n - d)$ and plot the obtained histogram superimposed with the density of the limiting distribution $\mathcal{N}(1, 4)$.

Table 1
The subtables give the observed rejection rates for different null hypotheses over 10000 independent replicates under each of the four settings. Two different sample sizes are considered for each setting. The columns corresponding to the true dimension are shaded grey.

Setting 1

n	H_{02}	H_{03}	H_{04}
216	1.000	0.053	0.115
512	1.000	0.051	0.138

Setting 2

n	H_{02}	H_{03}	H_{04}
216	1.000	0.054	0.122
512	1.000	0.051	0.131

Setting 3

n	H_{01}	H_{02}	H_{03}
36	1.000	0.051	0.102
64	1.000	0.053	0.124

Setting 4

n	H_{01}	H_{02}	H_{03}
36	0.059	0.091	0.211
64	0.058	0.093	0.261

The results are shown in Fig. C.2 in Appendix C where we immediately make two observations: the convergence to the limiting distribution is (at least visually) rather fast in Settings 1–3, with the histograms exhibiting the Gaussian shape and being only slightly shifted to the left from their limiting density; in Setting 4 where the condition $p_n/(n\sqrt{\lambda_{nd}}) \to 0$ required by Theorem 1 is violated, the histogram has the correct shape and scale, but underestimates the location. The difference between the true mean and the mean of the replicates in Setting 4 is approximately 1.35 and testing (not shown here) reveals that the difference seems to stay roughly constant when n is increased. Based on this, it seems possible that, even when $p_n/(n\sqrt{\lambda_{nd}}) \nrightarrow 0$, the limiting distribution of $g_{n,d}$ could be made to equal $\mathcal{N}(1, 4)$ with a suitable additive correction term a_n, which vanishes, $a_n \to 0$ as $n \to \infty$, when the conditions of Theorem 1 are satisfied.

Next, we demonstrate how forward testing, as defined in (2), can be used to estimate the signal dimension d with a chain of tests for the null hypotheses $H_{0k} : d = k$. That is, we sequentially test $H_{00}, H_{01}, \ldots$ using, respectively, the test statistics $g_{n,0}, g_{n,1}, \ldots$ and take our estimate of the dimension to be the smallest k for which H_{0k} is not rejected. For each test, we use $\alpha = 0.05$, i.e., the two-sided 95% critical regions of the limiting $\mathcal{N}(1, 4)$-distribution. We consider Settings 1–4, but include an additional, larger sample size for each. Of the four settings, only the first and the third satisfy the assumptions of Corollary 1, see Fig. C.1 on how the four settings are located with respect to the "feasibility region" of the assumptions.

For simplicity, we report in Table 1 the rejection rates (over 10000 replicates) of the null hypotheses corresponding to the true dimension and the neighbouring dimensions only (the columns corresponding to the true dimension are shaded grey). In Settings 1 and 3 where the assumptions of Corollary 1 are satisfied, the test achieves rather accurately the nominal level at the true dimension and shows extremely good power at the smaller dimensions, as expected. Interestingly, the same conclusions are reached also in Setting 2 where the assumptions of Corollary 1 are not satisfied, implying that the assumptions, while sufficient, are not necessary for the consistency of the forward testing estimator. Finally, as expected, the procedure reaches neither a sufficient level nor power in Setting 4 where the conditions of Theorem 1 and Corollary 1 are not satisfied.

We conclude with an application of the procedure to the `phoneme` data in the R-package `ElemStatLearn`. The data consists of 4509 log-periodograms of length $p = 256$, each corresponding to a single utterance of one of several phonemes. For simplicity, we consider only the phoneme "sh" and, moreover, take only the first utterances of it by the first 64 speakers in the data set. This yields a data matrix with the dimensions $n = 64$ and $p = 256$, meaning that the experiment can be embedded, e.g., to either of the regimes $p_n = 4n$ and $p_n = n^{4/3}$. To gain some idea on the possible Gaussianity of the data, we ran univariate Shapiro–Wilk tests for each of the p variables using the Bonferroni correction and the significance level 0.05. Based on the tests, 4 out of the 256 variables were deemed as non-normal, implying that the assumption of Gaussianity might indeed be warranted in the current context.

We then applied the estimator (2) with $\alpha = 0.05$ to the data and obtained the estimate $\hat{d} = 14$, implying that there is indeed great room for dimension reduction in the data set. As an alternative, "naive" approach we also considered forward testing based on a sequence of tests of the form (1) that assume p to be fixed. It turned out that each of the tests was rejected (with $\alpha = 0.05$), giving the maximal estimate $\hat{d} = \min\{n, p\} = 64$. As the sample size is most likely too small for the fixed-dimension asymptotics to kick in (unlike for the high-dimensional asymptotics, which are in Table 1 seen to be good approximations already for sample sizes and dimensions comparable to the current situation), we conclude that ignoring the high-dimensional nature of the data led to a gross overestimation of the latent dimension.

5. Discussion

The main limitation of the presented results is the assumption of Gaussianity. This requirement could possibly be weakened by showing that the *universality phenomenon* applies to our scenario; in high-dimensional statistics, a result derived under the Gaussian assumption is said to exhibit universality if it continues to hold when the normal distribution is replaced with some other distribution that is close to it in some suitable sense, see Johnstone and Paul (2018) for a review. In the current situation concerning the limiting behaviour of second-order quantities, it seems reasonable to

conjecture that our main results continue to hold if the normal distribution is replaced with a distribution that shares its first four moments with the normal distribution. While the theoretical study of this claim goes beyond the scope of this work (our proofs rely heavily on various results for Wishart matrices), we nevertheless did quick experiments in Settings 1–4 described in Section 4, with the normal distribution replaced by the Laplace mixture $(1/2)\mathcal{L}(-\mu, b) + (1/2)\mathcal{L}(\mu, b)$ having the dispersion parameter $b = \sqrt{3/2} - 1$ and the mean $\mu = \sqrt{1 - 2b^2}$. The resulting distribution has identical moments with $\mathcal{N}(0, 1)$ up to the fourth one. The resulting rejection rates are shown in Table C.1 in Appendix C and indeed match very closely with those in Table 1, giving plausibility to the universality claim.

Acknowledgements

This work was supported by the Academy of Finland [335077]. The author would like to express his gratitude to the two anonymous reviewers whose comments helped greatly improve the manucsript.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2021.109209.

References

Baik, J., Silverstein, J.W., 2006. Eigenvalues of large sample covariance matrices of spiked population models. J. Multivariate Anal. 97 (6), 1382–1408.
Beran, R., Srivastava, M.S., 1985. Bootstrap tests and confidence regions for functions of a covariance matrix. Ann. Statist. 13 (1), 95–115.
Birke, M., Dette, H., 2005. A note on testing the covariance matrix for large dimension. Statist. Probab. Lett. 74 (3), 281–289.
Johnstone, I.M., Paul, D., 2018. PCA in high dimensions: An orientation. Proc. IEEE 106 (8), 1277–1292.
Jolliffe, I.T., 2002. Principal Component Analysis. Springer, second ed..
Jung, S., Marron, J.S., 2009. PCA consistency in high dimension, low sample size context. Ann. Statist. 37 (6B), 4104–4130.
Karoui, N.E., 2003. On the largest eigenvalue of Wishart matrices with identity covariance when n, p and $p/n \to \infty$. ArXiv preprint, math/0309355.
Ledoit, O., 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.
Luo, W., Li, B., 2016. Combining eigenvalues and variation of eigenvectors for order determination. Biometrika 103 (4), 875–887.
Nordhausen, K., Oja, H., Tyler, D.E., 2016. Asymptotic and bootstrap tests for subspace dimension. ArXiv preprint arXiv:1611.04908.
Schott, J.R., 2006. A high-dimensional test for the equality of the smallest eigenvalues of a covariance matrix. J. Multivariate Anal. 97 (4), 827–843.
Shen, D., Shen, H., Marron, J., 2016. A general framework for consistency of principal component analysis. J. Mach. Learn. Res. 17 (1), 5218–5251.
Yao, J., Zheng, S., Bai, Z., 2015. Large Sample Covariance Matrices and High-Dimensional Data Analysis. Cambridge University Press.
Yata, K., Aoshima, M., 2009. PCA consistency for non-Gaussian data in high dimension, low sample size context. Comm. Statist. Theory Methods 38 (16–17), 2634–2652.
Yata, K., Aoshima, M., Nakayama, Y., 2018. A test of sphericity for high-dimensional data and its application for detection of divergently spiked noise. Sequential Anal. 37 (3), 397–411.