Journal of Multivariate Analysis 159 (2017) 18–38

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Likelihood ratio test for partial sphericity in high and ultra-high dimensions

Liliana Forzani [a], Antonella Gieco [a], Carlos Tolmasky [b,*]

[a] Universidad Nacional del Litoral, Santa Fe de la Vera Cruz, Argentina

[b] Institute for Mathematics and Its Applications and MCFAM, University of Minnesota, Minneapolis, MN 55455, United States

ARTICLE INFO

Article history:
Received 5 July 2016
Available online 18 April 2017

AMS subject classifications:
62H15
62G10
62G20

Keywords:
Sample covariance matrix
Spiked population model
High-dimensional statistics
Principal component analysis
Random matrix theory

ABSTRACT

We consider, in the setting of p and n large, sample covariance matrices whose population counterparts follow a spiked population model, i.e., with the exception of the first (largest) few, all the population eigenvalues are equal. We study the asymptotic distribution of the partial maximum likelihood ratio statistic and use it to test for the dimension of the population spike subspace. Furthermore, we extend this to the ultra-high-dimensional case, i.e., $p > n$. A thorough study of the power of the test gives a correction that allows us to test for the dimension of the population spike subspace even for values of the limit of p/n close to 1, a setting where other approaches have proved to be deficient.

1. Introduction

In many applications involving high-dimensional data, a few of the dimensions contain most of the relevant information. Identifying how many dimensions should be kept in the analysis is of paramount importance in representing and modeling data efficiently. Even though this issue has attracted much attention from practitioners as well as researchers, there is still no clear consensus on how to proceed in a systematic way. Among practitioners, a popular approach amounts to checking how many of the transformed variables explain a large part of the variance in the data and little (if any) attention is paid to the nature of what is discarded. An exception to this simplified approach is presented in [20], in which the authors compare the bulk of the eigenvalues to the typical bulk found in random matrix theory.

Systems of this sort, in which a small number of variables contain all the relevant information, appear in various fields. In an effort to understand these types of systems, Johnstone [11] introduced the spiked population model. In this model, all the population eigenvalues are equal to 1 except for a few fixed, larger eigenvalues that carry the relevant information. The behavior of the sample eigenvalues of the spiked population model in the high-dimensional case has been thoroughly studied in the past decade; see, e.g., [3,2,19]. In a remarkable result, Baik et al. [2] proved that the asymptotic behavior of the sample eigenvalues experiences a phase transition. If a population eigenvalue from the spike is not big enough, its value cannot be recovered from the samples: the estimated eigenvalue gets pulled towards the bulk, the noisy section of the matrix. On the other hand, if the spike population eigenvalue is bigger than a certain threshold, its value can be recovered from the limit of the estimates, which are, however, biased.

* Corresponding author.
E-mail addresses: liliana.forzani@gmail.com (L. Forzani), antogieco@gmail.com (A. Gieco), tolmasky@ima.umn.edu (C. Tolmasky).

http://dx.doi.org/10.1016/j.jmva.2017.04.001
0047-259X/

The same question about how many components should be kept was long ago answered in the traditional p fixed, n growing paradigm (here p indicates the dimension of the data $\mathbf{X}$ and n indicates the sample size). One of the most common tests assumes that the data follow a normal distribution and uses the maximum likelihood ratio statistics $LRT_d = L_d/L_p$, where L_d indicates the maximum likelihood under the null hypothesis (that d components should be kept) while L_p is the maximum likelihood under the full model [15]. This maximum likelihood ratio test is used sequentially, starting with $d = 0$ and estimating d as the first hypothesized value that is not rejected. In the fixed p and n growing paradigm, under the null hypothesis, $\ln(LRT_d)$ has a known asymptotic distribution—a fact used by Bartlett [4] and by Lawley [13] to build the rejection region of the test. Another common approach, which has the advantage of requiring no subjective judgments, is based on the application of information theoretic criteria. Wax and Kailath [26] presented an estimator in this direction using the minimum description length (MDL) principle [21,22]. In both cases, sequential testing or information criteria, a crucial ingredient is the knowledge of the asymptotic distribution of the maximum likelihood ratio statistic under the null hypothesis.

In the high-dimensional case, the dimensionality of the data can be relatively large compared to the sample size and traditional statistical theory cannot be easily adapted. Under the assumption that there exist $q_0 < p < n$ fixed components, Kritchman and Nadler [12] considered the MDL estimator developed in [26]. They show that MDL fails to detect the signal at low signal-to-noise ratios and hence underestimates the signal at small sample sizes; they then present a new estimator that improves the detection rate. Nevertheless, they only prove the consistency of their estimator under the scenario in which p is fixed and $n \to \infty$.

One of the contributions of our paper is the study of the asymptotic distribution of the partial maximum likelihood ratio statistic for the case in which $p, n \to \infty, p/n \to y \in (0, 1)$. This allows us to present a sequential test to determine the dimension of the population spike subspace. Also, as a bonus, it paves the way to correct the penalty term in Wax–Kailath's MDL estimator of the true dimension and then prove its consistency in this high-dimensional scenario.

We also address the problem for $p > n$. In some applications one can find situations in which the number of variables exceeds the number of observations ($y > 1$). Suppose that we have multiple time series and, given a window in time, we look for a small number of factors that contain most of the relevant information. In principle, we could take a big window (large n) to estimate the covariance matrix. Financial time series, for example, change frequently (they could even be non-stationary) leading us to believe that bigger time windows do not help in the understanding of the current structure. To attack a situation of this sort we would need to develop a similar test for the case $p \geq n, p/n \to y \in [1, \infty)$. In this case the maximum likelihood ratio statistic is not defined; see [7]. However, we motivate a new definition by switching the rows and columns in the data matrix. We find its asymptotic distribution and extend the definition and consistency of the MDL criteria to this case. It should be noted that the case $d = 0$ was already done by Srivastava [23].

This paper is organized as follows: Section 2 presents the asymptotic distribution of the maximum likelihood ratio statistic which is used in Section 3 to define the sequential test. Section 4 illustrates the results using simulated scenarios. The power of the test is found in Section 5. Finally, Section 6 builds on the analysis from Section 5 to improve on the way to estimate the true dimension in a consistent way and Section 7 concludes. All proofs are relegated to Appendix A.

The following notation and definitions will be used in our exposition. For positive integers m and n, $\mathbb{R}^{m\times n}$ stands for the class of all matrices of dimension $m \times n$. For a square matrix $\mathbf{A}$, $|\mathbf{A}|$ indicates its determinant. We will use the operator $\mathrm{vec} : \mathbb{R}^{m\times n} \to \mathbb{R}^{mn}$ which vectorizes an arbitrary matrix by stacking its columns. Let $\mathbf{A} \otimes \mathbf{B}$ denote the Kronecker product of matrices $\mathbf{A}$ and $\mathbf{B}$. We will use $\mathbf{S} \sim \mathcal{W}_p(m, \mathbf{\Sigma})$ to denote that $\mathbf{S}$ follows a Wishart distribution with m degrees of freedom and scale matrix $\mathbf{\Sigma}$, i.e., $\mathbf{S} = \mathbf{X}^\top\mathbf{X}$ where $\mathbf{X} \in \mathbb{R}^{m\times p}$ has independent rows following a normal distribution with mean 0 and covariance matrix $\mathbf{\Sigma}$. We write $\chi^2(f)$ for the chi-square distribution with f degrees of freedom. The multivariate Gamma function is defined as $\Gamma_p(x) = \pi^{p(p-1)/4}\prod_{j=1}^{p}\Gamma\{x - 1/2(j-x)\}$ for a complex number x with $\mathrm{Re}(x) > 1/2(p-1)$, where $\Gamma(x)$ is the ordinary Gamma function; see p. 62 of [15].

2. Asymptotic distribution of the maximum likelihood ratio statistic for partial sphericity

For $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, with $\mathbf{X} \in \mathbb{R}^p$, the *sphericity test* is given by

$$\mathcal{H}_0 : \mathbf{\Sigma} = \sigma^2\mathbf{I}_p \quad \text{vs.} \quad \mathcal{H}_a : \mathbf{\Sigma} \neq \sigma^2\mathbf{I}_p \tag{1}$$

with unknown σ. The maximum likelihood ratio test statistic to test the null hypothesis (1) was first derived by Mauchly [14] as the power $n/2$ of

$$LRT_0 = |\widehat{\mathbf{\Sigma}}|\{\mathrm{tr}(\widehat{\mathbf{\Sigma}})/p\}^{-p}, \tag{2}$$

where $\widehat{\mathbf{\Sigma}}$ is the sample covariance matrix of the data $\mathbf{X}_1, \ldots, \mathbf{X}_n$, defined as $\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top/(n-1)$. Gleser [7] shows that the maximum likelihood ratio statistic exists only when $p \leq n - 1$ and that the test with the rejection region $\{LRT_0 \leq c_\alpha\}$ (where c_α is chosen so that the test has a significance level of α) is unbiased. The choice of c_α follows from the classical asymptotic result (see [15], Theorem 8.3.7) to the effect that under $\mathcal{H}_0$ with p fixed

$$-(n-1)\rho\ln(LRT_0) \rightsquigarrow \chi^2(f)$$