Journal of Multivariate Analysis 139 (2015) 79–91

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

High-dimensional tests for spherical location and spiked covariance

Christophe Ley [a], Davy Paindaveine [a,*], Thomas Verdebout [b]

[a] Département de Mathématique and ECARES, Université Libre de Bruxelles, Avenue F.D. Roosevelt, 50 - ECARES, CP 114/04 and Campus Plaine, CP 210, bvd du triomphe, B-1050 Bruxelles, Belgium

[b] EQUIPPE and INRIA, Université Lille III, Domaine Universitaire du Pont de Bois, BP 60149, F-59653 Villeneuve d'Ascq Cedex, France

ARTICLE INFO

Article history:
Received 12 April 2014
Available online 10 March 2015

AMS subject classifications:
62H11
62H15

Keywords:
Directional statistics
High-dimensional data
Location tests
Principal component analysis
Rotationally symmetric distributions
Spherical mean

ABSTRACT

This paper mainly focuses on one of the most classical testing problems in directional statistics, namely the spherical location problem that consists in testing the null hypothesis $\mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ under which the (rotational) symmetry center $\boldsymbol{\theta}$ is equal to a given value $\boldsymbol{\theta}_0$. The most classical procedure for this problem is the so-called Watson test, which is based on the sample mean of the observations. This test enjoys many desirable properties, but its asymptotic theory requires the sample size n to be large compared to the dimension p. This is a severe limitation, since more and more problems nowadays involve high-dimensional directional data (e.g., in genetics or text mining). In the present work, we derive the asymptotic null distribution of the Watson statistic as both n and p go to infinity. This reveals that (i) the Watson test is robust against high dimensionality, and that (ii) it allows for (n, p)-asymptotic results that are universal, in the sense that p may go to infinity arbitrarily fast (or slowly) as a function of n. Turning to Euclidean data, we show that our results also lead to a test for the null that the covariance matrix of a high-dimensional multinormal distribution has a "$\boldsymbol{\theta}_0$-spiked" structure. Finally, Monte Carlo studies corroborate our asymptotic results and briefly explore non-null rejection frequencies.

1. Introduction

The technological advances and the ensuing new devices to collect and store data lead nowadays in many disciplines to data sets with very high dimension p, often larger than the sample size n. Consequently, there is a need for inferential methods that can deal with such high-dimensional data, and this has entailed a huge activity related to high-dimensional problems in the last decade. One- and multi-sample location problems have been investigated in [23,22,9,24,25], among others. Since the seminal paper by Ledoit and Wolf [15], problems related to covariance or scatter matrices have also been thoroughly studied by several authors; see, e.g., [10,16,18,13]. In particular, the problem of testing for sphericity has attracted much attention.

In this paper, we are interested in high-dimensional *directional* data, that is, in data lying on the unit hypersphere

$$\mathcal{S}^{p-1} = \left\{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}} = 1\right\},$$

with p large. Such data occur when only the direction of the observations and not their magnitude matters, and are extremely common, e.g., in magnetic resonance [11], gene-expression [1], and text mining [2]. Inference for high-dimensional

* Corresponding author.
E-mail address: dpaindav@ulb.ac.be (D. Paindaveine).

http://dx.doi.org/10.1016/j.jmva.2015.02.019
0047-259X/

directional data has already been considered in several papers. For instance, Banerjee and Ghosh [3,4] and Banerjee et al. [2] investigate clustering methods in this context. Most asymptotic results available, however, have been obtained as p goes to infinity, with n fixed. This is the case of almost all results in [26,28,30,11]. To the best of our knowledge, the only (n, p)-asymptotic results available can be found in [11,8,7,19]. However, Dryden [11] imposes the stringent condition that $p/n^2 \to \infty$ when studying the asymptotic behavior of the classical pseudo-FvML location estimator (FvML here refers to *Fisher–von Mises–Langevin* distributions; see below). Cai and Jiang [8] and Cai et al. [7] consider various (n, p)-asymptotic regimes in the context of testing for uniformity on the unit sphere, but the tests to be used depend on the regime considered which makes practical implementation problematic. Finally, Paindaveine and Verdebout [19] propose tests that are robust to the (n, p)-asymptotic regime considered; their tests, however, are sign procedures, hence are not based on sufficient statistics—unlike the much more classical pseudo-FvML procedures.

In the present paper, we intend to overcome these limitations in the context of the spherical location problem, one of the most fundamental problems in directional statistics. The natural distributional framework for this problem is provided by the class of *rotationally symmetric distributions* (see Section 2), that is a semiparametric model, indexed by a finite-dimensional (location) parameter $\boldsymbol{\theta} \in \mathcal{S}^{p-1}$ and an infinite-dimensional parameter F. The spherical location problem is the problem

$$\begin{cases} \mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \\ \mathcal{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0, \end{cases}$$

where $\boldsymbol{\theta}_0$ is a given unit vector and F remains unspecified. The classical test for this problem is the so-called Watson test, based on the sample mean of the observations; see [29]. This test enjoys many desirable properties, and in particular is a *pseudo-FvML* procedure: in other words, it achieves optimality under FvML distributions, yet remains valid (in the sense that it meets the asymptotic nominal level constraint) under extremely mild assumptions on F.

Unfortunately, nothing is known about the validity of the Watson test in the high-dimensional setup, which, in view of the growing number of high-dimensional directional data to be analyzed, is a severe limitation. Therefore, the aim of this paper is to investigate this issue. We derive the (n, p)-asymptotic null properties of the Watson test. Our results require minimal distributional assumptions and allow for virtually any rotationally symmetric distributions. Even better: in contrast with earlier asymptotic investigations of high-dimensional pseudo-FvML procedures, our asymptotic results are "universal" in the sense that they only require that p goes to infinity as n does (p may go arbitrarily fast (or slowly) to infinity as a function of n). Moreover, as an interesting by-product, we show that our procedures can be used to test the null hypothesis that the covariance matrix of a high-dimensional multinormal distribution is "$\boldsymbol{\theta}_0$-spiked", meaning that it is of the form $\boldsymbol{\Sigma} = \sigma^2(\mathbf{I}_p + \lambda\boldsymbol{\theta}_0\boldsymbol{\theta}_0')$ for some $\sigma^2 > 0$ and some $\lambda \geq 0$ (here, $\boldsymbol{\theta}_0 \in \mathcal{S}^{p-1}$ is fixed); see, e.g., [14] or the quite recent Onatski et al. [18] where this covariance structure has been used as an alternative to sphericity.

The outline of the paper is as follows. In Section 2, we define the class of rotationally symmetric distributions and introduce the Watson test for spherical location. In Section 3, we propose a standardized Watson test statistic and derive its asymptotic null distribution in the high-dimensional setting. We also prove that, in some cases, it is asymptotically equivalent to a sign test statistic. In Section 4, we show that the standardized Watson test further allows to test for a spiked covariance structure in high-dimensional multinormal distributions. Monte Carlo studies are conducted in Section 5, while an Appendix collects the proofs.

2. Rotational symmetry and the Watson test

The distribution of the random p-vector $\mathbf{X}$, with values on the unit hypersphere $\mathcal{S}^{p-1}$, is *rotationally symmetric* about location $\boldsymbol{\theta}(\in \mathcal{S}^{p-1})$ if $\mathbf{OX}$ is equal in distribution to $\mathbf{X}$ for any orthogonal $p \times p$ matrix $\mathbf{O}$ satisfying $\mathbf{O}\boldsymbol{\theta} = \boldsymbol{\theta}$; see [21]. Rotationally symmetric distributions are characterized by the location parameter $\boldsymbol{\theta}$ and an infinite-dimensional parameter, the cumulative distribution function F of $\mathbf{X}'\boldsymbol{\theta}$, hence they are of a semiparametric nature. The rotationally symmetric distribution associated with $\boldsymbol{\theta}$ and F will be denoted as $\mathcal{R}(\boldsymbol{\theta}, F)$ in the sequel. The most celebrated members of this family are the Fisher-von Mises-Langevin distributions, corresponding to

$$F_{p,\kappa}(t) = c_{p,\kappa} \int_{-1}^{t} (1 - s^2)^{(p-3)/2} \exp(\kappa s)\, ds \quad (t \in [-1, 1]),$$

where $c_{p,\kappa}$ is a normalization constant and $\kappa(>0)$ is a *concentration* parameter (the larger the value of κ, the more concentrated about $\boldsymbol{\theta}$ the distribution is); see [17] for further details.

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from $\mathcal{R}(\boldsymbol{\theta}, F)$ and consider the problem of testing the null hypothesis $\mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against the alternative $\mathcal{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0 \in \mathcal{S}^{p-1}$ is fixed and F remains unspecified. At first sight, the rotational symmetry assumption may appear quite restrictive. Note however that it contains the null hypothesis of uniformity on the sphere, which itself contains the null hypothesis of sphericity for Euclidean data (since the uniform distribution on the sphere may be obtained by projecting spherical distributions on the sphere), a null that has been the topic of numerous papers in high-dimensional statistics.

Letting $\bar{\mathbf{X}} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$, the classical test for the problem above rejects the null for large values of the Watson statistic

$$W_n := \frac{n(p-1)\bar{\mathbf{X}}'(\mathbf{I}_p - \boldsymbol{\theta}_0\boldsymbol{\theta}_0')\bar{\mathbf{X}}}{1 - \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i'\boldsymbol{\theta}_0)^2}. \tag{2.1}$$