On projection-based tests for spherical and compositional data

Juan A. Cuesta-Albertos, Antonio Cuevas and Ricardo Fraiman
Departamento de Matemáticas, Estadística y Computación, Univ. de Cantabria, Spain
Departamento de Matemáticas, Univ. Autónoma de Madrid, Spain
Departamento de Matemática, Univ. de San Andrés, Argentina and Centro de Matemática, Univ. de la República, Uruguay

Summary

A new class of nonparametric tests, based on random projections, is proposed. They can be used for several null hypotheses of practical interest, including uniformity for spherical (directional) and compositional data, sphericity of the underlying distribution and homogeneity in two-sample problems on the sphere or the simplex.

The proposed procedures have a number of advantages, mostly associated with their flexibility (for example, they also work to test “partial uniformity” in a subset of the sphere), computational simplicity and ease of application even in high-dimensional cases.

This paper includes some theoretical results concerning the behaviour of these tests, as well as a simulation study and a detailed discussion of a real data problem in astronomy.

Key words: Uniformity; Directional data; Sphericity; Compositional data

1 Introduction

1.1 Uniformity tests for directional data

According to the usual terminology, directional data are those whose sample space is the unit sphere, $S^{d-1}$, of the space $\mathbb{R}^d$ endowed with the usual euclidean norm, that is $S^{d-1} = \{ x : \| x \|_2 = 1 \}$, where $\| x \|_2 = (\sum_{i=1}^{d} x_i^2)^{1/2}$, with $x = (x_1, \ldots, x_d)^t$.

The books by Mardia and Jupp (2000) and Fisher (1993) are classical references on the subject. Of course the cases $d = 2$ (circular data) and $d = 3$ (spherical data) are particularly important. As Mardia and Jupp (2000, p. 1) point out, the circular data come often from the compass (e.g., wind directions) or the clock (e.g., arrival times of patients to a medical service). Also, astronomy is a continual source of spherical data: An example is discussed in Section 5 below.
The study of directional data requires a special statistical theory, in some sense “parallel” to the classical one, adapted in order to cope with the fact that in directional data the sample space is a manifold that differs in many respects from the euclidean space. Thus, some basic concepts as mean, variance and density estimators must be carefully re-defined. Likewise, most standard models of multivariate statistics, in particular the Gaussian model, are not straightforwardly extended to the directional case. However, we will be mainly concerned here with the uniform distribution whose definition in the spherical case poses no particular difficulty: It is just the probability distribution with constant density on the sphere. More precisely, in this paper we propose a new procedure for the usual task of testing uniformity, which in many cases arises as a first natural step in the statistical analysis, before going further in the search for data structure.

Two classical procedures for testing uniformity in directional data are Kuiper’s test (e.g., Mardia and Jupp, 2000, p. 99), valid for the case $d = 2$, and Rayleigh’s test, which is asymptotic but can be applied for directional samples of any dimension (e.g., Mardia and Jupp, 2000, p. 99 and p. 207). The first one is an “universal test”, consistent against all alternatives; it relies on ideas similar to those of the Kolmogorov-Smirnov (K-S) procedure. Rayleigh’s test is rather designed to detect unimodal alternatives (Fisher, 1993, p. 69) and it is not universally consistent. In fact, it is the likelihood ratio test for testing a simple hypothesis within the von Mises family. It can be adapted to the case that the mean direction of the alternative distribution is given in advance.

Another more recent proposal, which will be also considered in the comparisons below, is Giné’s (1975) test. It is based on an asymptotic approximation but it is universal and can be applied for any dimension $d$; see Mardia and Jupp (2000, p. 209).

1.2 A new proposal based on random projections

We present here a new family of uniformity tests for directional data based on the use of projections along random directions, as proposed in Cuesta, Fraiman and Ransford (2007) (CFR07, in the sequel). Roughly speaking, in that paper it is shown that the distribution of a random element (even taking values in an infinite-dimensional space) is determined by that of a one-dimensional projection along just a randomly chosen direction. For a precise statement see Theorem 4.1 in CFR07. Usually this approach leads to a conditional (on the obtained random direction) test. However it is particularly easy to use in the framework of directional data since, as we will see, the distribution of the projections under the null hypothesis of uniformity does not depend on the obtained direction and admits a closed simple expression. Thus, in this situation, we have an unconditional test.

The projection methodology is also useful in a number of problems, besides testing uniformity. In particular, we will discuss how to use it in order to test the sphericity of the underlying distribution around a given point as well as to test the homogeneity (i.e., the coincidence) of two distributions in a two-sample problem with directional data. Still, the paper is mainly oriented to the uniformity problem just to focus the discussion.
in a situation where the null distributions of the test statistics are explicitly known and particularly simple.

Our test procedures are especially competitive for high-dimensional directional data, that is for data on the unit sphere $S^{d-1}$, with $d \gg 3$. Moreover, they are universally consistent and do not involve any asymptotic approximation.

While it is clear that the high-dimensional directional data appear less frequently in practice than those on the circle or the unit sphere, they are also interesting in some practical situations; see for example Juan and Prieto (2001) where a test of uniformity for $S^{d-1}$-valued data is proposed in connection with a problem of outlier detection.

1.3 On directional and compositional data

A further application of high-dimensional data on the sphere arises in the study of problems involving proportions, that is, those situations in which the observed variable is of type $\mathbf{x} = (x_1, \ldots, x_d)^t \in \mathbb{R}^d$, with $x_j \geq 0$ and $\sum x_j = 1$, $x_j$ being, for example, the proportion of time devoted to the $j$-th activity by a worker or the proportion of the $j$-th component in the soil. These are the so-called “compositional data” which have received a considerable attention in the literature, especially motivated by their applications in geology and chemistry; see Aitchison and Egozcue (2005) for a recent survey. Stephens (1982) analyzes a direct connection between directional and compositional data given by the change of variable $\sqrt{x_j} = \xi_j$ which takes the vector of proportions $(x_1, \ldots, x_d)$ to a point $(\xi_1, \ldots, \xi_d)$ in the “positive part” of $S^{d-1}$. We explore here a different connection, based on the simple observation that the compositional data are just elements of the positive quadrant in the unit sphere of the space $\mathbb{R}^d$ endowed with the $L^1$ norm, $\|\mathbf{x}\|_1 = \sum |x_j|$. Hence, in a way, compositional data are also directional when viewed in the appropriate space. We will show below that this connection is in fact deeper than suggested by the formal replacement of $\|\cdot\|_2$ with $\|\cdot\|_1$: Our main results, Theorems 2.1 and 4.3 (which provide the support for projection–based tests of uniformity and homogeneity), cover also the case of compositional data.

1.4 Organization of the paper and notation

Our uniformity projection tests are introduced in Section 2. A simulation study, devoted only to the problem of testing uniformity on the unit sphere, appears in Section 3: Different versions of the projection tests are compared with Kuiper’s, Rayleigh’s and Giné’s test. Sphericity and homogeneity tests are studied in Section 4. A real data problem concerning the uniformity of the cometary orbits is discussed in Section 5.

We will assume that all the random variables are defined on the same rich enough probability space $(\Omega, \sigma, P)$. Given the random vectors $\mathbf{X}$ and $\mathbf{Y}$, $P_{\mathbf{X}}$ will denote the distribution generated by $\mathbf{X}$, the expression $\mathbf{X} \sim \mathbf{Y}$ will mean that $P_{\mathbf{X}} = P_{\mathbf{Y}}$, and $P[\mathbf{X} \in B|\mathbf{Y}]$ will denote the conditional probability that $[\mathbf{X} \in B]$ given $\mathbf{Y}$. 

3
2 The uniformity random projection tests for directional data

In this section we present the theoretical developments in which our procedures are based.

2.1 The Random Projection Tests based on a unique direction (RP1): Theoretical basis

Let $X_1, \ldots, X_n$ be a random sample of $S^{d-1}$-valued random variables. Denote by $U$ another random variable, independent from the $X_i$’s and uniformly distributed on $S^{d-1}$. Let $Y_1 = X'_1 U, \ldots, Y_n = X'_n U$ be the one-dimensional projections of the directional data $X_i$ on the random direction $U$.

As we will see, the distribution of the random variables $Y_i$ uniquely determines, with probability one (in a sense to be specified in Theorem 2.1 (a)), the distribution of $X_1$. Thus, almost surely, a consistent test for the null hypothesis

$H_0$: The distribution of $X_1$ is uniform on $S^{d-1}$

would be obtained by just testing

$H_0^*$: The distribution function of $Y_1$ is $F_0$,

where $F_0$ denotes the distribution function of the projection $Y_1 = X'_1 U$ under $H_0$. The hypothesis $H_0^*$ could be tested using a suitable goodness-of-fit test as for example the classical, universally consistent, Kolmogorov-Smirnov procedure.

Thus, our first proposal to test $H_0$ can be summarized as follows:

(a) Given the directional data (on $S^{d-1}$) $X_1, \ldots, X_n$, select at random (uniformly) a direction $U$ in $S^{d-1}$. A simple way to obtain $U$ with a standard software package is to take $U = Z/\|Z\|$, $Z$ being a $d$-dimensional standard Gaussian distribution.

(b) Compute the projections $Y_1 = X'_1 U, \ldots, Y_n = X'_n U$.

(c) Compute the empirical distribution $F_n$ of the $Y_i$ and the K-S statistic

$$D_n = \sup_x |F_n(x) - F_0(x)|.$$

(d) Reject $H_0^*$, and consequently $H_0$, at a significance level $\alpha$, whenever $D_n > C_{n,\alpha}$, where $C_{n,\alpha}$ denotes the appropriate critical value extracted from the distribution of the K-S statistic.

We will call this procedure the “RP1 test” (i.e., Random Projection test based on a unique direction). It can be seen as a particular case of a general methodology of (multivariate and functional) inference based on projections: See Cuesta-Albertos, Fraiman and Ransford (2006) and Cuesta-Albertos et al. (2007).
The distribution $F_0$ can be calculated explicitly (see Juan and Prieto (2001)) or can be easily approximated, with arbitrary precision, by drawing very large uniform samples on $S^{d-1}$. However, in the cases $d = 2$ and $d = 3$, the distribution function $F_0$ has a particularly simple form (see Proposition 2.2) which eases considerably the computations.

Our first theoretical result is the following theorem whose statement (a) provides the basis for the proposed uniformity test. Part (b) extends the result for compositional data.

**Theorem 2.1.** (a) Let $X$ be a $S^{d-1}$-valued random variable and $U$ be a random variable, independent from $X$, with uniform distribution on $S^{d-1}$. With probability one, the distribution of $X$ is uniquely determined by the distribution of the projection of $X$ on $U$, in the sense that, if $X_1$ and $X_2$ are $S^{d-1}$-valued random variables, independent from $U$ and the probability that $U$ takes a value $u$ such that $X_1^t u \sim X_2^t u$ is positive, then $X_1$ and $X_2$ are identically distributed.

(b) Result (a) is also true if the unit sphere $S^{d-1}$ corresponding to the euclidean norm is replaced with the unit $\| \cdot \|_1$-sphere $S^{d-1}_1$.

**Proof:** (a) Obviously, we can consider that $X$ is a bounded variable taking values on $\mathbb{R}^d$ and, without loss of generality, we can assume that $U = Z/\|Z\|$ where $Z$ is Gaussian. Thus, the condition on the moments of $X$ in Theorem 4.1 in CFR07 holds. However, taking into account that, up to a constant, the conditional distribution of $X^t U$ coincides with that of $X^t Z$ the result follows easily from Theorem 4.1 in CFR07.

(b) Similarly to the case (a), the proof is a direct consequence of the extension of Theorem 4.1 in CFR07 to the Banach spaces setup given in Cuevas and Fraiman (2007, Theorem 4) which covers the case of the non-Hilbert norm $\| \cdot \|_1$, that is, this result proves that the distribution of $X$ is determined from that of the dual element $X^t U$ where $U$ is uniform on $S^{d-1}_1$.

**Remark 2.1.1.** Let us emphasize on the importance of the randomness in Theorem 2.1. Clearly, if $X_1$ and $X_2$ are two $S^{d-1}$-valued random variables such that, for a fixed direction $u_0$ (for instance, $u_0 = (1, 0, \ldots, 0)^t$), $X_1^t u_0$ and $X_2^t u_0$ have the same distribution, there is no guarantee that $X_1$ and $X_2$ are identically distributed. The interest of Theorem 2.1 lies in the fact that, if we choose $u_0$ at random, we can be sure that $X_1$ and $X_2$ are identically distributed whenever $X_1^t u_0$ and $X_2^t u_0$ have the same distribution.

**Remark 2.1.2.** It is not essential that the distribution of $U$ is uniform. Other possible continuous distributions with support on $S^{d-1}$ could do the same job. However, we choose the uniform by obvious reasons of simplicity.

**Remark 2.1.3.** The RP1 procedure (and all the extensions considered below) can also be adapted, with obvious modifications, to test hypotheses of uniformity on some subsets of $S^{d-1}$. For example, if the support of $X$ is the “Northern” (upper) hemisphere $S^{d-1}_N$, and we want to test the uniformity of $X$ in $S^{d-1}_N$, we may perform the RP1 test evaluating
the distribution of a random projection under \( H_0 \). This evaluation can be carried out by Monte Carlo simulation or by direct calculation, but taking into account that, in general, the distribution of the projection will depend on the chosen direction.

In a similar way we can test the hypothesis of homogeneity (equal distribution) of two \( S^{d-1} \)-valued random variables (see Section 4).

The following proposition includes some known results on the distributions of the projections which can be checked by direct calculation. They are included here for the sake of completeness and will be used later. Further results about projections and uniform distributions on the sphere appear in Brown, Cartwright and Eagleson (1986).

**Proposition 2.2.** 1. If \( X \) is uniform on \( S^{d-1} \) and \( u_0 \in S^{d-1} \), then the distribution of \( Y = X^t u_0 \) does not depend on \( u_0 \).

2. In the case \( d = 2 \) the distribution function \( F_0 \) is
   \[
   F_0(t) = \frac{\pi - \cos(t)}{\pi}, \text{ for all } t \in (-1, 1),
   \]
   where \( \cos(t) \) denotes the only value \( t \in [0, \pi] \) whose cosine is \( t \).

3. In the case \( d = 3 \) the distribution \( F_0 \) is uniform in the interval \([-1, 1]\).

### 2.2 The Random Projection Tests based on \( k \) directions (RP\( k \))

The projection direction \( U \) in the RP1 test is randomly chosen, which suffices to perfectly identify the distribution of \( X \) with probability one and obtain a consistent test. However, under the alternative, this also entails the risk of getting an unfortunate direction \( u_0 \) where there is no much difference between the distribution of the projection of \( X \) along \( u_0 \) and \( F_0 \). Thus, with a certain positive probability, we can have a test of low power (conditional to \( U = u_0 \)). A natural thing to do in this respect is to consider several independent random directions \( U_1, \ldots, U_k \) along which the original data could be successively projected. Then, we could perform the RP1 test for each of these directions and to combine appropriately the results to get an overall test. More precisely, denote by \( P_1, \ldots, P_k \) the \( p \)-values of the RP1 tests performed with the original data projected along the directions \( U_1, \ldots, U_k \), respectively. As these tests are not in general independent, we could adopt the classical Bonferroni procedure. This would lead to reject \( H_0 \) at a level \( \alpha \) whenever \( \min_j P_j \leq \alpha/k \).

However, it is well known that Bonferroni procedure is quite conservative. An alternative is to use \( \tau_n = \min_j P_j \) itself as a test statistic, provided that its null distribution is known or can be numerically approximated. According to Berk and Jones (1978), this test enjoys some optimality properties. In our case we have used Monte Carlo simulations to approximate its conditional distribution (given the obtained directions \( U_1, \ldots, U_k \)) under the null hypothesis.

We will call this method the “RP\( k \) test”.

6
Remark 2.2.1. A further version of our procedure would arise by projecting on a number \( k \) (given in advance) of directions chosen in a “random-systematic” way, as follows: The first direction \( \mathbf{u}_1 \) is randomly chosen but the remaining ones are systematically selected from \( \mathbf{u}_1 \) in order to cover a wide range of different directions to project, thus taking benefit of the consistency provided by the random choice but avoiding the possible “bad luck” in the selection of redundant directions. We have checked by simulation, different versions of this mixed, random-systematic procedure but we do not include them here in order to keep the focus on the main ideas of the random projection methodology.

3 A simulation study

In this section we will study the performance of the different tests \( RP_k \) through a simulation study, restricting ourselves to the most common cases \( d = 2 \) and \( d = 3 \). The high-dimensional situation \( d > 3 \) will be considered and motivated in Section 4, in connection with the problem of testing sphericity and homogeneity.

The details of the simulation study are as follows:

Tests to be compared: In all cases, the aim is to check the null hypothesis of uniformity for the underlying distribution which generates the data. In the case of circular data (\( d = 2 \)) we compare the random projection tests \( RP_1, RP_2, RP_3, RP_5, RP_{10}, \) and \( RP_{25} \) and the classical Kuiper’s and Rayleigh’s tests mentioned in the introduction. For Kuiper’s test we have used the implementation included in the R package Circular, developed by Lund and Agostinelli (2006).

As for Rayleigh’s test, we use the improved version of this test described in Mardia and Jupp (2000, p. 207): The test statistic is

\[
\left( 1 - \frac{1}{2n} \right) T + \frac{1}{2n(d+2)} T^2,
\]

where \( T = dn\|\bar{X}\|_2^2, \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \). The asymptotic distribution of this statistic under the uniform distribution is \( \chi^2_d \) and the approximation error is of order \( O(n^{-2}) \).

For spherical data (\( d = 3 \)) we employ the random projection tests \( RP_1, RP_{5}, RP_{10}, RP_{25}, RP_{50}, \) and \( RP_{100} \). These are compared, again, with Rayleigh test but Kuiper’s test (which is not easily extended for data on the sphere \( S^{d-1} \) with \( d > 2 \)) is in this case replaced with Giné’s test. This test can be used in different versions. We have chosen the implementation based on the statistic \( F_n \) defined in Mardia and Jupp (2000) p. 209.

In the choice of the tests for comparison, the idea is to assess in each case (\( d = 2, 3 \)) the performance of our \( RP_k \) tests against two of the most popular “classical” competitors. For \( d = 2 \) the choice of Kuiper’s test seems quite natural, in view of its universal consistency and relative performance when compared with other tests; see e.g. Mardia and Jupp (2000, p. 115). Giné’s procedure plays a similar role in the high dimensional cases. Finally, the choice of Rayleigh’s test looks quite natural, as it is maybe the most popular
uniformity test in circular statistics and can be used in any dimension. Moreover, it
enjoys interesting optimality properties against von Mises alternatives (see Mardia and

Underlying distributions: All the considered tests are used under the null hypothesis and
under several alternative distributions, namely:

1. The model $M_1$: This is a sort of location projected normal model, which can be
seen as a particular case of the so-called angular Gaussian or offset normal; see
Mardia and Jupp (2000, p. 46). The distributions in this model are indexed by a
real parameter $b$. In the case $d = 2$ the data are obtained by projecting on the
unit circumference $S^1$ observations drawn from the bivariate Gaussian distribution
$N_2((b, b)^t, I_2)$, where $I_d$ denotes the $d$-dimensional identity matrix. In other words,
if a random variable $Z^b$ is $N_2((b, b)^t, I_2)$, then the distribution of $Z^b/\|Z^b\|$ is $M_1(b)$.
Analogously, in the three-dimensional case the data are obtained by projecting on
the unit sphere $S^2$ observations drawn from the Gaussian distribution $N_3((b, b, b)^t, I_3)$.

2. The model $M_2$: This is another sub-model (a sort of scale normal projected model)
of the offset normal. In this case, the data are generated by projecting on the unit
circumference the random points drawn from the standard bi-variate Gaussian dis-
tribution multiplied by the matrix $B_b = \begin{pmatrix} 1 & b \\ b & 1 \end{pmatrix}$, where, again, $b$ is a real parameter.
This amounts to project on $S^1$ the random observations generated from a centered
Gaussian distribution with $b^2I_2 + B_{2b}$ as covariance matrix.
In the $d$-dimensional case (with $d \geq 3$) the definition of model $M_2$ is analogous: the
matrix $B_b$ is replaced with another $d \times d$ matrix with 1 in the main diagonal and $b$
outside. We are now concerned with the cases $d = 2, 3$ but in Section 4 below we
will also use the model $M_2$ in high-dimensional problems with $d > 3$.

3. The von Mises model: This is maybe the most popular model in directional statistics.
The von Mises density on $S^1$ and $S^2$ (see, e.g., Mardia and Jupp (2000), pp. 36 and
167) depends on a mean direction $\mu$ (= 0 and $(1, 0)^t$, respectively, in the simulations
below in dimensions $d = 2$ and $d = 3$ respectively) and a concentration parameter
$\kappa > 0$. The uniform distribution appears as a limit as $\kappa \to 0$.

Number of runs, significance level, sample sizes: The outputs in Tables 1 and 2 below
(which correspond to $d = 2$ and $d = 3$, respectively) show the proportion of rejections
(power) over 10000 replications for distributions belonging to the models $M_1$, $M_2$ (values
of $b = 0.2, 0.4, 0.8$) and von Mises (values of $\kappa = 0.2, 0.4, 0.8$). The intended significance
level is in all cases 0.05. On the other hand, in both cases $d = 2$ and $d = 3$, the columns
with heading $H_0$ analyze the behavior under the null and should be used to check the
performance of the different tests to keep the intended type I error (0.05) under control.

The considered sample sizes are $n = 25, 50, 100$. 
Software: The simulation experiments have been implemented using the R language (R Development Core Team, 2006). The corresponding codes are available from the authors. We have used the R packages MASS (Venables and Ripley, 2002) and the above mentioned Circular, (Lund and Agostinelli, 2006). In the generation of the von Mises distribution on the sphere $S^{d-1}$ with $d > 2$ (see also Section 4 below) we have used the algorithm proposed by Ulrich (1984).

Table 1.- Power of some tests on $S^1$ under the null hypothesis and several alternatives.

<table>
<thead>
<tr>
<th>Test</th>
<th>$H_0$</th>
<th>$b = .2$</th>
<th>$b = .4$</th>
<th>$b = .8$</th>
<th>$b = .2$</th>
<th>$b = .4$</th>
<th>$b = .8$</th>
<th>$\kappa = .2$</th>
<th>$\kappa = .4$</th>
<th>$\kappa = .8$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>n = 25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RP1</td>
<td>.0516</td>
<td>.1194</td>
<td>.3450</td>
<td>.7131</td>
<td>.0682</td>
<td>.1625</td>
<td>.6737</td>
<td>.0738</td>
<td>.1494</td>
<td>.4000</td>
</tr>
<tr>
<td>RP2</td>
<td>.0535</td>
<td>.1415</td>
<td>.4365</td>
<td>.8755</td>
<td>.0776</td>
<td>.1905</td>
<td>.8135</td>
<td>.0781</td>
<td>.1712</td>
<td>.5056</td>
</tr>
<tr>
<td>RP3</td>
<td>.0531</td>
<td>.1486</td>
<td>.4750</td>
<td>.9383</td>
<td>.0791</td>
<td>.1945</td>
<td>.8722</td>
<td>.0822</td>
<td>.1831</td>
<td>.5495</td>
</tr>
<tr>
<td>RP5</td>
<td>.0517</td>
<td>.1559</td>
<td>.5037</td>
<td>.9731</td>
<td>.0777</td>
<td>.2014</td>
<td>.9216</td>
<td>.0813</td>
<td>.1862</td>
<td>.5854</td>
</tr>
<tr>
<td>RP10</td>
<td>.0496</td>
<td>.1578</td>
<td>.5195</td>
<td>.9872</td>
<td>.0769</td>
<td>.2091</td>
<td>.9538</td>
<td>.0827</td>
<td>.1891</td>
<td>.6061</td>
</tr>
<tr>
<td>RP25</td>
<td>.0476</td>
<td>.1568</td>
<td>.5189</td>
<td>.9880</td>
<td>.0753</td>
<td>.2106</td>
<td>.9609</td>
<td>.0797</td>
<td>.1814</td>
<td>.6088</td>
</tr>
<tr>
<td>Kuiper</td>
<td>.0489</td>
<td>.1258</td>
<td>.3907</td>
<td>.9540</td>
<td>.0804</td>
<td>.2252</td>
<td>.9700</td>
<td>.0753</td>
<td>.1527</td>
<td>.4809</td>
</tr>
<tr>
<td>Rayleigh</td>
<td>.0497</td>
<td>.1851</td>
<td>.5945</td>
<td>.9937</td>
<td>.0497</td>
<td>.0576</td>
<td>.0747</td>
<td>.0877</td>
<td>.2075</td>
<td>.6703</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n = 50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RP1</td>
<td>.0499</td>
<td>.2038</td>
<td>.5415</td>
<td>.8459</td>
<td>.0947</td>
<td>.3132</td>
<td>.8824</td>
<td>.0960</td>
<td>.2467</td>
<td>.5932</td>
</tr>
<tr>
<td>RP2</td>
<td>.0489</td>
<td>.2415</td>
<td>.6913</td>
<td>.9562</td>
<td>.1021</td>
<td>.3623</td>
<td>.9656</td>
<td>.1068</td>
<td>.2920</td>
<td>.7508</td>
</tr>
<tr>
<td>RP3</td>
<td>.0518</td>
<td>.2595</td>
<td>.7555</td>
<td>.9877</td>
<td>.1051</td>
<td>.3930</td>
<td>.9895</td>
<td>.1122</td>
<td>.3165</td>
<td>.8214</td>
</tr>
<tr>
<td>RP5</td>
<td>.0502</td>
<td>.2796</td>
<td>.8131</td>
<td>.9987</td>
<td>.1116</td>
<td>.4286</td>
<td>.9979</td>
<td>.1167</td>
<td>.3394</td>
<td>.8808</td>
</tr>
<tr>
<td>RP10</td>
<td>.0494</td>
<td>.2853</td>
<td>.8379</td>
<td>1</td>
<td>.1187</td>
<td>.4606</td>
<td>.9999</td>
<td>.1183</td>
<td>.3564</td>
<td>.9074</td>
</tr>
<tr>
<td>RP25</td>
<td>.0505</td>
<td>.2920</td>
<td>.8500</td>
<td>1</td>
<td>1.1258</td>
<td>.4841</td>
<td>1</td>
<td>1.1214</td>
<td>.3597</td>
<td>.9120</td>
</tr>
<tr>
<td>Kuiper</td>
<td>.0477</td>
<td>.2175</td>
<td>.7181</td>
<td>.9999</td>
<td>.1238</td>
<td>.4960</td>
<td>1</td>
<td>.0964</td>
<td>.2735</td>
<td>.8238</td>
</tr>
<tr>
<td>Rayleigh</td>
<td>.0482</td>
<td>.3280</td>
<td>.8913</td>
<td>1</td>
<td>.0524</td>
<td>.0542</td>
<td>.0697</td>
<td>.1280</td>
<td>.4035</td>
<td>.9431</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n = 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RP1</td>
<td>.0552</td>
<td>.3565</td>
<td>.7051</td>
<td>.9349</td>
<td>.1705</td>
<td>.5685</td>
<td>.9931</td>
<td>.1544</td>
<td>.4290</td>
<td>.7511</td>
</tr>
<tr>
<td>RP2</td>
<td>.0561</td>
<td>.4502</td>
<td>.8847</td>
<td>.9887</td>
<td>.1884</td>
<td>.6979</td>
<td>.9991</td>
<td>.1834</td>
<td>.5305</td>
<td>.9063</td>
</tr>
<tr>
<td>RP3</td>
<td>.0549</td>
<td>.4952</td>
<td>.9449</td>
<td>.9994</td>
<td>.1961</td>
<td>.7616</td>
<td>.9999</td>
<td>.1938</td>
<td>.5757</td>
<td>.9624</td>
</tr>
<tr>
<td>RP5</td>
<td>.0527</td>
<td>.5226</td>
<td>.9793</td>
<td>1</td>
<td>.2057</td>
<td>.8193</td>
<td>1</td>
<td>.2007</td>
<td>.6124</td>
<td>.9921</td>
</tr>
<tr>
<td>RP10</td>
<td>.0506</td>
<td>.5398</td>
<td>.9918</td>
<td>1</td>
<td>.2143</td>
<td>.8538</td>
<td>1</td>
<td>.2014</td>
<td>.6327</td>
<td>.9974</td>
</tr>
<tr>
<td>RP25</td>
<td>.0505</td>
<td>.5509</td>
<td>.9922</td>
<td>1</td>
<td>.2280</td>
<td>.8741</td>
<td>1</td>
<td>.2057</td>
<td>.6416</td>
<td>.9981</td>
</tr>
<tr>
<td>Kuiper</td>
<td>.0471</td>
<td>.4180</td>
<td>.9627</td>
<td>1</td>
<td>.2268</td>
<td>.8856</td>
<td>1</td>
<td>.1590</td>
<td>.5178</td>
<td>.9899</td>
</tr>
<tr>
<td>Rayleigh</td>
<td>.0487</td>
<td>.5909</td>
<td>.9957</td>
<td>1</td>
<td>.0526</td>
<td>.0493</td>
<td>.0714</td>
<td>.2228</td>
<td>.7062</td>
<td>.9992</td>
</tr>
</tbody>
</table>

The conclusions of these simulations could be summarized as follows:

(a) In the case of circular data, $d = 2$ (Table 1), Rayleigh test is the winner in both offset model M1 and von Mises model. The RP$k$ tests based on a large number of directions (RP10 and RP25) rank second, outperforming Kuiper test (especially in the von Mises model).

(b) Still in the circular case (Table 1), under model M2, the Rayleigh test exhibits a total failure; in fact, it is almost “blind” against this alternative. Kuiper test is the winner under this model with a very slight advantage over RP25. Thus the overall performance of the RP methods is quite satisfactory.
Table 2 - Power of some tests on $S^2$ under the null hypothesis and several alternatives

<table>
<thead>
<tr>
<th>Test</th>
<th>$H_0$</th>
<th>Model M1</th>
<th>Model M2</th>
<th>von Mises model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$b = .2$</td>
<td>$b = .4$</td>
<td>$b = .6$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$n = 25$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RP1</td>
<td>0.0519</td>
<td>.1383</td>
<td>.3498</td>
<td>.6545</td>
</tr>
<tr>
<td>RP5</td>
<td>0.0498</td>
<td>.1760</td>
<td>.5757</td>
<td>.9708</td>
</tr>
<tr>
<td>RP10</td>
<td>0.0513</td>
<td>.1912</td>
<td>.6383</td>
<td>.9558</td>
</tr>
<tr>
<td>RP25</td>
<td>0.0535</td>
<td>.2046</td>
<td>.6876</td>
<td>.9986</td>
</tr>
<tr>
<td>RP50</td>
<td>0.0513</td>
<td>.2010</td>
<td>.6858</td>
<td>.9991</td>
</tr>
<tr>
<td>RP100</td>
<td>0.0504</td>
<td>.1994</td>
<td>.6810</td>
<td>.9994</td>
</tr>
<tr>
<td>Giné</td>
<td>0.0506</td>
<td>.2163</td>
<td>.7251</td>
<td>.9997</td>
</tr>
<tr>
<td>Rayleigh</td>
<td>0.0483</td>
<td>.2300</td>
<td>.7474</td>
<td>.9998</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$n = 50$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RP1</td>
<td>0.0480</td>
<td>.2139</td>
<td>.5145</td>
<td>.7925</td>
</tr>
<tr>
<td>RP5</td>
<td>0.0460</td>
<td>.3249</td>
<td>.8665</td>
<td>.9972</td>
</tr>
<tr>
<td>RP10</td>
<td>0.0510</td>
<td>.3665</td>
<td>.9248</td>
<td>1</td>
</tr>
<tr>
<td>RP25</td>
<td>0.0561</td>
<td>.3986</td>
<td>.9515</td>
<td>1</td>
</tr>
<tr>
<td>RP50</td>
<td>0.0557</td>
<td>.4072</td>
<td>.9592</td>
<td>1</td>
</tr>
<tr>
<td>RP100</td>
<td>0.0516</td>
<td>.3959</td>
<td>.9556</td>
<td>1</td>
</tr>
<tr>
<td>Giné</td>
<td>0.0495</td>
<td>.4290</td>
<td>.9692</td>
<td>1</td>
</tr>
<tr>
<td>Rayleigh</td>
<td>0.0484</td>
<td>.4415</td>
<td>.9731</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$n = 100$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RP1</td>
<td>0.0506</td>
<td>.3597</td>
<td>.6673</td>
<td>.9034</td>
</tr>
<tr>
<td>RP5</td>
<td>0.0518</td>
<td>.5946</td>
<td>.9813</td>
<td>1</td>
</tr>
<tr>
<td>RP10</td>
<td>0.0470</td>
<td>.6520</td>
<td>.9982</td>
<td>1</td>
</tr>
<tr>
<td>RP25</td>
<td>0.0465</td>
<td>.6724</td>
<td>.9996</td>
<td>1</td>
</tr>
<tr>
<td>RP50</td>
<td>0.0468</td>
<td>.6884</td>
<td>.9998</td>
<td>1</td>
</tr>
<tr>
<td>RP100</td>
<td>0.0532</td>
<td>.6971</td>
<td>.9999</td>
<td>1</td>
</tr>
<tr>
<td>Giné</td>
<td>0.0445</td>
<td>.7578</td>
<td>.9998</td>
<td>1</td>
</tr>
<tr>
<td>Rayleigh</td>
<td>0.0453</td>
<td>.7707</td>
<td>.9999</td>
<td>1</td>
</tr>
</tbody>
</table>

(c) The situation is a bit different in the case of spherical data ($d = 3$, Table 2), where the methods based on many directions (RP50 and RP100) are the winners (with slight advantage over Giné’s test) under the von Mises model. As in case $d = 2$, Rayleigh is the winner under M1 with a slight advantage over Giné’s test, which ranks second not very far from RP50 and RP100. Again the Rayleigh test fails under M2, where Giné’s test wins with a considerable advantage.

(d) The $H_0$ columns show that the RP$k$ tests succeed in keeping the significance level, with a similar performance to that of the considered competitors.

(e) All in all, it seems that the tests RP25, for $d = 2$, and RP100 (for $d = 3$) are competitive with a reasonable overall performance. Obviously, the RP tests based on few directions are not to be recommended. They are included in the study only for the sake of a more complete understanding on the behavior of the RP method.
4 Projection tests for sphericity and homogeneity

In this section we show how to use the random projections methodology in two further testing problems. The first one is to test sphericity around a fixed point. The second one is the homogeneity problem, that is to test whether two or more distributions coincide. In both cases the purpose is to present all the required ideas though, due to space constraints, we include just some practical illustrations rather than extensive simulations.

4.1 Testing sphericity around a fixed point

A random vector \( Z = (Z_1, \ldots, Z_d)^t \) in \( \mathbb{R}^d \) is said to be spherical, or to have a spherical distribution around the point \( \mu_X \), if for any matrix \( A \) in the group of the orthogonal \( d \times d \) matrices \( O(d) \) (i.e. matrices such that \( AA^t = A^tA = I_d \)), the vector \( A(Z - \mu_X) \) has the same distribution as \( Z - \mu_X \).

Along this section, it will be assumed that the point \( \mu_X \) is known so that we may assume that \( \mu_X = 0 \). Thus in what follows “spherical” will mean “spherical around 0”.

It is well-known that if \( Z \) is spherical and it has an absolutely continuous distribution with density \( f \), then \( f \) must be necessarily of the form \( f(x) = h(\|x\|) \), for some univariate function \( h \). Also, if \( Z \) has finite second-order moments, the sphericity assumption entails that the covariances matrix \( \Sigma \) of \( Z \) fulfills

\[
\Sigma = \sigma^2 I_d, \tag{1}
\]

where \( \sigma^2 \) is the common variance of the \( Z_i \). Note that, if \( Z \) is normally distributed, sphericity is in fact equivalent to (1).

See Fang, Kotz and Ng (1990) for a detailed account on spherical distributions.

There are several sphericity tests widely used in applied statistics. A very popular, classical choice is Bartlett’s test. It is based on the use of a (asymptotic) chi-square goodness of fit methodology to test the null hypothesis that the correlation matrix is \( I_d \). See, e.g., Lattin et al. (2003, p. 109) for an elementary account of this test and its usefulness in principal component analysis. Another related well-known procedure, where the null hypothesis is also (1), is Mauchly’s test; see Timm (2002, p. 140).

These tests require the existence of second-order moments (which is a condition external to the sphericity assumption). Moreover, they cannot be applied in high-dimensional problems when the dimension \( d \) is typically larger than the sample size \( n \) since they are usually based on the estimation of the correlation matrix \( R \) (in particular Bartlett’s test is based on \( \log(\det(R)) \)) and the corresponding estimators will fail when \( n < d \) (or sometimes where \( n \) is close to \( d \)).

The closeness between sphericity and uniformity provides a simple procedure to test sphericity, without any moment assumptions, in those high-dimensional cases. More
specifically, it is easy to prove (see, e.g., Theorem 2.3 in Fang, Kotz and Ng (1990)) that if \( \|Z\| \) is almost surely positive and \( Z \) is spherical, then \( Z/\|Z\| \) is uniformly distributed on \( S^{d-1} \). So we could think of replacing the null hypothesis \( H_0 : Z \) has a spherical distribution, with \( \bar{H}_0 : Z/\|Z\| \) is uniformly distributed on \( S^{d-1} \). While these hypotheses are not equivalent (as the second one is less restrictive) in most practical cases the violations of sphericity will arise from the non-fulfillment of \( \bar{H}_0 \). Note also that the classical tests of sphericity (Bartlett, Mauchly,...) are non-universal as well, since they test when the covariance matrix is diagonal, which is neither equivalent to the sphericity. Thus, any uniformity test suitable for high-dimensional data (in particular, our projection-based tests) will work in practice to test sphericity.

However, we next show how the random projections methodology can be used to get a specific consistent sphericity test. The whole procedure relies again on the above commented Theorem 4.1 of CFR07. In view of the assumptions in that theorem, our test will require that the underlying distribution has finite moments of any order. As a counterpart, its is consistent among those distributions fulfilling these conditions and can be applied for any dimension and sample size.

The main result in this section (Theorem 4.3) is based on Propositions 4.1 and 4.2. The first one relates the sphericity with a randomly selected matrix on \( O(d) \). Proposition 4.2 could have some independent interest.

**Proposition 4.1.** Let \( \lambda \) be a probability measure on the orthogonal group \( O(d) \), that do not concentrate mass on any proper subgroup. Let \( X \) be a random vector on \( \mathbb{R}^d \). Then, \( X \) is spherical if and only if \( \lambda(\{A \in O(d) \mid AX \sim X\}) > 0 \).

**Proof:** The “only if” part is obvious. Concerning the “if” part, let \( \Gamma = \{A \in O(d) : AX \sim X\} \). Observe that

if \( A, B \in \Gamma, AB \in \Gamma \), since \( BX \sim X \) implies \( ABX \sim AX \sim X \),

if \( A \in \Gamma \), then \( A^{-1} \in \Gamma \), since \( AX \sim X \) implies \( X = A^{-1}AX \sim A^{-1}X \),

which implies that \( \Gamma \) is a subgroup of \( SO(d) \).

Therefore, we have that \( \Gamma = O(d) \) because \( \Gamma \) is a group and \( \lambda \) does not concentrate mass on any proper subgroup of \( O(d) \).

Proposition 4.1 implies that if \( X \) is not spherical and we choose at random an orthogonal matrix \( A \), then \( AX \not\sim X \), for almost all \( A \). Moreover, it is possible to take as \( \lambda \) in this lemma the Haar (uniform) measure on \( O(d) \).

**Proposition 4.2.**

1. If \( \lambda \) is the Haar measure on \( O(d) \), \( A \) is a random matrix with distribution \( \lambda \), and \( U \) is a random vector uniformly distributed on \( S^{d-1} \), then

(a) The distribution of \( AU \) is uniform on \( S^{d-1} \).
(b) The random vectors \( \mathbf{U} \) and \( \mathbf{AU} \) are independent.

2. Reciprocally, if \( \mathbf{U}_1 \) and \( \mathbf{U}_2 \) are two independent random vectors uniformly distributed on \( S^{d-1} \), then there exists a unique random matrix on \( \mathcal{O}(d) \) such that \( \mathbf{U}_2 = \mathbf{AU}_1 \) and the distribution of \( \mathbf{A} \) does not concentrate mass on any proper subgroup of \( \mathcal{O}(d) \).

Proof: Let \( B_1, B_2 \) be Borel sets in \( S^{d-1} \). Obviously, the surface measure of the sets \( B_1 \) and \( \mathbf{A}^T B_1 \) coincide. From here, and the independence between \( \mathbf{U} \) and \( \mathbf{A} \) we obtain that

\[
P[\mathbf{AU} \in B_1] = \int P\left[ \mathbf{U} \in \mathbf{A}^T B_1 | \mathbf{A} \right] dP = \int P[\mathbf{U} \in B_1]dP = P[\mathbf{U} \in B_1],
\]

which proves (a).

Now, from the independence between \( \mathbf{A} \) and \( \mathbf{U} \) and the fact that, if \( h_1, h_2 \in S^{d-1} \), then \( \mathbf{A}h_1 \sim \mathbf{A}h_2 \), we have that

\[
P[\mathbf{U} \in B_1, \mathbf{AU} \in B_2] = \int_{B_1} P[\mathbf{Ah} \in B_2 | \mathbf{U} = h]dP_U(h)
\]

\[
= \int_{B_1} P[\mathbf{Ah} \in B_2]dP_U(h) = P[\mathbf{Ah} \in B_2]P[\mathbf{U} \in B_1],
\]

where \( e_1 \) is any vector in \( S^{d-1} \). A similar reasoning, would give us that \( P[\mathbf{Ae}_1 \in B_2] = P[\mathbf{AU} \in B_2] \) and (b) follows from (2).

Concerning 2, we have that for every \( h_1, h_2 \in S^{d-1} \), there exists a unique \( A_{h_1, h_2} \in \mathcal{O}(d) \) such that \( \mathbf{h}_2 = (A_{h_1, h_2}) \mathbf{h}_1 \). Since \( \mathbf{A} \) is a continuous function of \( \mathbf{h}_1 \) and \( \mathbf{h}_2 \), we have that the map \( \omega \rightarrow \mathbf{A}(\omega) := A_{U_1(\omega), U_2(\omega)} \) is measurable and, by construction, \( \mathbf{U}_2 = \mathbf{AU}_1 \).

If \( P_\mathbf{A} \) were concentrated in a proper subgroup of \( \mathcal{O}(d) \), then the distribution of \( \mathbf{U}_2 \) given that \( \mathbf{U}_1 = \mathbf{u} \) would be concentrated on the transformation of \( \mathbf{u} \) by this subgroup and it would not be uniform on \( S^{d-1} \) which contradicts the assumptions.

Theorem 4.3. Let \( \mathbf{X} \) be a random vector with values in \( \mathbb{R}^d \) such that \( m_n = \int \|x\|^n dP_\mathbf{X}(x) < \infty \), for all \( n \) and \( \sum_n m_n^{-1/n} = \infty \). Denote by \( \lambda \) the Haar measure on \( S^{d-1} \). Then \( \mathbf{X} \) has a spherical distribution if and only if

\[
\lambda \ast \lambda \left[ \{(h_1, h_2) : h_1^T \mathbf{X} \sim h_2^T \mathbf{X}\} \right] > 0
\]

Proof: Let \( \mathbf{A} \) be a random matrix, independent from \( \mathbf{X} \) with the Haar distribution on \( \mathcal{O}(d) \). According to the Proposition 4.1, the sphericity of \( \mathbf{X} \) is equivalent to that \( \mathbf{AX} \sim \mathbf{X} \).

Let \( \mathbf{U} \) be a random vector uniformly distributed on \( S^{d-1} \) and independent from \( \mathbf{A} \) and \( \mathbf{X} \). Theorem 4.1 in CFR07 implies that \( \mathbf{AX} \sim \mathbf{X} \) if and only if

\[
P_\mathbf{U} \left[ h : h^T \mathbf{AX} \sim h^T \mathbf{X} \right] > 0.
\]
But this is equivalent to say that
\[
0 < P_U \left[ h : P[(U'A)X|U = h] = P[U'X|U = h] \right].
\] (5)

Obviously, \( P_{h'X} = P[(U'A)X|U = h] \) and from 1.(b) in Proposition 4.2 we obtain the independence between \( U'A \) and \( U \). Thus, (5) is equivalent to
\[
0 < P_U \left[ h : U'AX \sim h'tX \right].
\]

Finally, by 1.(a) in Proposition 4.2, we know that the distribution of \( U'A \) is uniform and, then we finally have that (4) is equivalent to (3).

\[\text{Remark 4.3.1.}\] If we replace (3) by
\[
\lambda \times \lambda \left[ \{(h_1, h_2) : h_1'X \not\sim h_2'X \} \right] = 0,
\]
Theorem 4.3 states the quite intuitive fact that if the distribution of \( X \) is spherical and \( h_1, h_2 \in S^{d-1} \) then \( h_1'X \) and \( h_2'X \) are identically distributed.

But statement (3) takes advantage of Theorem 4.1 in CFR07 in order to provide a result which is easily translated to construct an sphericity test as shown next.

Theorem 4.3 suggests the following procedure to test sphericity: Given a random sample \( X_1, \ldots, X_N \) from a random vector \( X \), take \( n \) a positive integer, such that \( n < N \) and choose \( h_1, h_2 \in S^{d-1} \) independent with uniform distribution.

Now, consider the independent samples of real random variables \( \{h_1'X_1, \ldots, h_1'X_n\} \) and \( \{h_2'X_{n+1}, \ldots, h_2'X_N\} \). By Theorem 4.3, testing the sphericity of \( X \) is equivalent to test that the distribution who produced both samples is the same. Thus, we can test the sphericity of \( X \) just testing (for instance, with the KS-test) the hypothesis of homogeneity of both samples. This procedure will be denoted RP1-S. Note that it does not require to know in advance the distributions \( P_X \) nor \( P_{h'X} \). Likewise, this procedure can be adapted to employ \( k \) pairs of projection directions \( h_1, h_2 \) (instead of just one) using the standard Bonferroni device, i.e., to reject \( H_0 \) at a level \( \alpha \) whenever it is rejected for at least one of the \( k \) tests performed at level \( \alpha/k \). In this case we cannot use the minimum \( p \)-value as a statistic test itself (as indicated in Subsection 2.2) since the null hypothesis is composed (that is, it is fulfilled for more than one distribution). Indeed while all the the individual \( p \)-values are uniformly distributed under the null hypothesis, their joint law will depend in general on the underlying spherical distribution and so the minimum \( p \)-value will not be in general distribution free. This procedure will be denoted RP\( k \)-S.

In our view there are mainly two situations where these tests are especially suitable, when compared with the classical choices (e.g., Bartlett’s test) based on covariance matrices: First, the high dimensional cases with \( d \sim n \); second, those cases where the underlying distribution is not spherical but its projection on the unit sphere is still uniformly distributed. As we will show in the next subsection such examples arise in a quite simple way, even in \( \mathbb{R}^2 \).
4.2 Some Monte Carlo comparisons for sphericity tests

We have compared our projection RP1-S and RP10-S tests with Bartlett’s test in high dimensional situations \((d = 50, 100)\) using two sample sizes: \(n = 50, 100\). More specifically, we have evaluated, along 2000 runs, the empirical level (i.e. the rejection proportion under the null hypothesis of uniformity) of Bartlett’s test and that of the RP\(k\)-S tests. We have also computed the respective empirical powers under the offset model \(M2\) with a parameter value \(b = 0.5\). The results are given in Table 3 below. The entry NA means that there were errors when computing \(\log(\det(R))\) in Bartlett’s test.

Table 3.- Power of the RP1-S, RP10-S and Bartlett’s test under the null hypothesis and model \(M2\) with \(b = 0.5\) in \(\mathbb{R}^d\), \(d = 50, 100\), and sample sizes \(n = 50, 100\)

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Test</th>
<th>(H_0) (n = 50)</th>
<th>(H_0) (n = 100)</th>
<th>Model (M2), (b = 0.5) (n = 50)</th>
<th>Model (M2), (b = 0.5) (n = 100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(d = 50)</td>
<td>RP1-S</td>
<td>0.028</td>
<td>0.034</td>
<td>0.384</td>
<td>0.598</td>
</tr>
<tr>
<td></td>
<td>RP10-S</td>
<td>0.016</td>
<td>0.024</td>
<td>0.466</td>
<td>0.959</td>
</tr>
<tr>
<td></td>
<td>Bartlett</td>
<td>NA</td>
<td>0.143</td>
<td>NA</td>
<td>1</td>
</tr>
<tr>
<td>(d = 100)</td>
<td>RP1-S</td>
<td>0.034</td>
<td>0.041</td>
<td>0.366</td>
<td>0.578</td>
</tr>
<tr>
<td></td>
<td>RP10-S</td>
<td>0.021</td>
<td>0.020</td>
<td>0.558</td>
<td>0.973</td>
</tr>
<tr>
<td></td>
<td>Bartlett</td>
<td>1</td>
<td>NA</td>
<td>1</td>
<td>NA</td>
</tr>
</tbody>
</table>

As it could be foreseen, Bartlett test simply does not work in these situations since it is unable to keep the level and shows an erratic behaviour, associated with the numerical instabilities in the estimation of a large correlation matrix from too few sample observations. The RP1-S tends to be rather conservative. This is likely due to the use of the KS test in the final stage. The power under the alternative hypothesis is moderate for \(k = 1\) but increases appreciably for \(k = 10\), even under the present low rates \(n/d\).

Another situation in which the RP\(k\)-S tests could be particularly useful is in those cases where the lack of uniformity cannot be detected via an uniformity test since the projections of the data on the unit sphere are uniformly distributed. Such distributions are very easy to construct. For example we could consider a model under which the data come from a standard normal distribution, but those belonging to a given cone \(C\) are multiplied for a positive constant \(b\). We have done 2000 simulations using this model (which we call \(M3\)) in the case \(d = 2\), with \(C = \{(x_1, x_2) : x_1 < 0\}\). The results are shown in Table 4.

Table 4.- Power of the RP1-S, RP10-S and Bartlett’s tests under the null hypothesis and model \(M3\) with \(b = 1, 3, 6, 8\) in \(\mathbb{R}^2\), and sample sizes \(n = 100, 200\)

<table>
<thead>
<tr>
<th>Test</th>
<th>(H_0) (n = 100)</th>
<th>Model M3, (n = 100)</th>
<th>Model M3, (n = 200)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(b = 3)</td>
<td>(b = 6)</td>
<td>(b = 8)</td>
</tr>
<tr>
<td>RP1-S</td>
<td>0.041</td>
<td>0.176</td>
<td>0.294</td>
</tr>
<tr>
<td>RP10-S</td>
<td>0.025</td>
<td>0.216</td>
<td>0.600</td>
</tr>
<tr>
<td>Bartlett</td>
<td>0.053</td>
<td>0.083</td>
<td>0.098</td>
</tr>
</tbody>
</table>
4.3 Testing homogeneity. The case of compositional data

Our last proposal has to do with the application of the random projection methodology to compositional data (see Section 1.3 above).

As previously mentioned, the main point is that the compositional data can also be seen as “circular data” by just changing the $\|\cdot\|_2$-sphere by the $\|\cdot\|_1$-sphere $S_{d-1}$ (or rather its positive face). The basic result (Theorem 2.1) which allows us to characterize distributions in the euclidean sphere remains true for $S_{d-1}$.

As a consequence, one could also define uniformity tests for compositional data along similar lines to those indicated in Section 2. A technical difference is the fact that, unlike the circular case, the null distributions are not expected to show “closed” relatively simple forms in the compositional case. Therefore it should be approximated by Monte Carlo procedures. Otherwise, the method would be analogous.

However, we will not focus here on the uniformity test. We will rather briefly use the case of compositional data to illustrate another use of the random projection methodology, namely the development of homogeneity tests. In view of Theorem 2.1 the idea is quite simple: In a two sample problem with compositional data, we will decide that both samples come from the same distributions if their respective one-dimensional projections along a random direction are equally distributed. This can be decided by using the standard distribution-free Kolmogorov-Smirnov homogeneity test (KS2). Of course, the idea can also be adapted to be used with several independent random directions, instead of just one. In this case we use (as in the sphericity test) the Bonferroni procedure to combine the tests performed along different directions, instead of using the test based on the minimum p-value. Again the reason is that the null hypothesis of homogeneity is composed and, in general, the null distribution of the minimum p-value will depend on the common distribution of both samples.

As a practical illustration, let us consider a data set analyzed in Aitchison (1986, p. 22). It arises as a consequence of an ecological experiment during which “plots of land of equal area were inspected and the parts of each plot which were thick or thin in vegetation and dense or sparse in animals were identified. From this field work the areal proportions of each plot were calculated for the four mutually exclusive and exhaustive categories: thick-dense, thick-sparse, thin-dense, thin-sparse. These sets of proportions are recorded for 50 plots from each of two different regions A and B”. We quote from the help file corresponding to the data set AnimalVegetation in the R package compositions, by van den Boogaart et al. (2006).

The second row of Table 5 shows the p-values (after Bonferroni’s correction) obtained in one execution of the RP-homogeneity test with several numbers of independent random directions. The third row contains the proportions of rejections of the null (homogeneity)
hypothesis over 500 executions of the test (at level 0.05), with different random choices of the projection directions.

As a consequence of this analysis we must say that no evidence against homogeneity has been found. The choice of \( k \) does not essentially affect the conclusion.

<table>
<thead>
<tr>
<th>( k )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p )-values</td>
<td>.52</td>
<td>.52</td>
<td>.60</td>
<td>.57</td>
<td>.68</td>
</tr>
<tr>
<td>Rejections</td>
<td>.064</td>
<td>.054</td>
<td>.008</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

5 A real data example: Are the comet orbits uniformly distributed?

5.1 Some background

It is a well-known fact that the planes defined by the orbits of the planets in the Solar System are all nearly coincident with the ecliptic (the plane of the Earth’s orbit). The only partial exception is Pluto, which in fact is no longer considered as a true planet, according to the recent decision of the International Astronomical Union.

Such an “almost coincidence” in the orbits has intrigued the scientists for a long time. Thus D. Bernoulli (in the 1730’s) wondered if this fact could happen “by chance”. To put this question in a statistical framework, we could consider that the planet orbits are defined by the corresponding normal vectors. So they reduce to a sample of nine points in the unit sphere \( S^2 \). Therefore, one could think of using a uniformity test on the sphere. This has been done, e.g., by Mardia and Jupp (2000, p. 209) obtaining a strong evidence against the uniformity hypothesis. Whereas, as these authors rightly point out, it is not clear what is the random variable from which the planet orbits are a i.i.d. sample, this result has at least the interest of suggesting an analogous, but more ambitious question regarding the possible uniformity of comet orbits.

It is believed that most of the comets are produced in the Kuiper Belt or in the Oort Cloud following a nowadays unpredictable mechanism. Thus, their orbits could be seen as multiple realizations (in fact we have thousands of them) of an experiment whose output is not deterministic.

Comet orbits are always conical sections. So they are either elliptical (which lead to periodic comets), parabolic or hyperbolic (non-periodic comets). The orbit plane is usually determined by two angles, namely the inclination (denoted by \( i \)) which is the angle between the normal vector of the orbit and the normal vector of the ecliptic, and the so-called longitude of the ascending node, often denoted by \( \Omega \); see, e.g., Teets and Whitehead (1998) for a definition of \( \Omega \). The expression of the normal vector to the orbit plane in terms of \( i \) and \( \Omega \) is

\[
R = (\sin(i) \sin(\Omega), -\sin(i) \cos(\Omega), \cos(i))^t.
\]
Thus, the orbit of a comet can be seen as an observation from the $S^2$-valued random variable $R$. We are interested in testing the null hypothesis that $R$ is uniformly distributed on $S^2$ from a random sample $R_1, \ldots, R_n$ of this variable.

The paper by Jupp et al. (2003, see also references therein) tackles this problem with a particular emphasis on the statistical aspects. The analysis focuses on a data set of 658 single-apparition long-period cometary orbits from the catalog of Marsden and Williams (1993). Such data provide a strong statistical evidence against the uniformity assumption for the distribution of $R$ in this type of cometary orbits. Such a conclusion is somewhat surprising as it contradicts the isotropy properties which could be expected a priori. According to the explanation (supported by a careful data analysis) proposed by Jupp et al. (2003), the lack of uniformity is largely due to an observational bias which favours the observation of cometary orbits near to the ecliptic. As a solution, these authors provide a new probabilistic model on the unit sphere which incorporates an observational window defined by the inequality $|\sin(i)| \leq \epsilon$. This model assumes that the probability density of $R$ inside the observational window is uniform and its value outside (i.e. for $|\sin(i)| > \epsilon$) is $(2/\pi\epsilon)\arcsin(\epsilon/\sin(i))$. The parameter $\epsilon$ can be estimated by maximum likelihood: The estimation found in Jupp et al. (2003) is $\hat{\epsilon} = .84$. The proposed model shows a satisfactory goodness of fit to the data set at hand.

5.2 The data

We give here a further look to this problem, using now a different database, freely available from the NASA website, http://ssd.jpl.nasa.gov/sbdb_query.cgi#x. The following steps describe exactly how to obtain the data we have analyzed. The terminology is that used in the NASA web page:

STEP 1

Object group: Select All. Object kind: Tick Comets. Numbered state: Choose Unnumbered (the comets receive a permanent number prefix only after the second perihelion passage). Limit to selected orbit classes: In the box named Comet Orbit Classes select all, except Parabolic and Hyperbolic. Limit by other characteristics: In the box Select orbital parameter choose period(years). In the next box, operator, select $\geq$. In the box on the right write 200. Then tick the box Add.

STEP 2

In Object Fields select object internal database ID, and object full name/designation. In Object and Model Parameter Fields select inclination and longitude of the ascending node.

The above procedure led to a data file which, by December 14, 2007, included the values of $(i, \Omega)$ for 211 comets. An inspection of such data showed that four of them
(in the places 12 to 15) were identical up to the fourth decimal position. They were all labelled “Great September Comet”. Our statistical analysis has been applied to the 208 data obtained by suppressing three of these repeated observations.

An exploratory data analysis, similar to that in Jupp et al. (2003) has been performed in order to visually detect any clear deviations from uniformity.

![Histograms of the variables “inclination” (i) and “longitude of the ascending node” (Ω).](image)

Figure 1. Histograms of the variables “inclination” (i) and “longitude of the ascending node” (Ω).

Similarly to the case of the data studied in Jupp et al. (2003), the distribution of i shows some deviations from uniformity near the extremes of its range. However, as shown in the following subsection, such deviations do not provide enough statistical evidence against uniformity.

5.3 Results and conclusions

Giné’s and Rayleigh’s tests, as well as the random projection tests RP50 and RP100 (based on $k = 50$ and 100 random directions, respectively) have been performed with these data. The corresponding $p$-values are given in Table 6.

<table>
<thead>
<tr>
<th></th>
<th>RP50</th>
<th>RP100</th>
<th>Giné</th>
<th>Rayleigh</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$-value</td>
<td>.248 (.089)</td>
<td>.250 (.049)</td>
<td>.060</td>
<td>.210</td>
</tr>
</tbody>
</table>

Table 6.- $p$-values of different uniformity tests for the NASA comet data set.

In the case of the RP tests, the indicated values correspond to the average over six independent runs (based on different sets of random directions) applied to the same data. The values in parentheses are the respective standard deviations. It is worth mentioning that, in the case of RP50, the minimum $p$-value obtained (over the six runs) has been 0.11. In the case of RP100 this minimum value has been 0.18.

As a conclusion, we can say that no statistical evidence has been found against the
uniformity hypothesis.

The \( p \)-value of Giné’s test is the closest one to provide statistical evidence (but yet without enough support) against uniformity. This could be perhaps interpreted in terms of a higher sensitivity of this test with respect to the mild observational bias shown in the left histogram of Figure 1.

ACKNOWLEDGEMENTS

We are very grateful to César Sánchez-Sellero for a useful discussion on multiple contrasts and, in particular, for bringing to our attention the paper by Berk and Jones (1978).

This work has been partially supported by Spanish grants MTM2005-08519-C02-02 and PAIJC1 VA102/06 (J. Cuesta-Albertos) and MTM2007-66632 (A. Cuevas and R. Fraiman).

REFERENCES


