Statistica Sinica Preprint No: SS-2022-0347	
Title	Poisson Kernel-Based Tests for Uniformity on the d-Dimensional Sphere
Manuscript ID	SS-2022-0347
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0347
Complete List of Authors	Yuxin Ding, Marianthi Markatou and Giovanni Saraceno
Corresponding Authors	Marianthi Markatou
E-mails	markatou@buffalo.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

POISSON KERNEL-BASED TESTS FOR UNIFORMITY ON THE D-DIMENSIONAL SPHERE

Yuxin Ding, Marianthi Markatou, Giovanni Saraceno

Department of Biostatistics, University at Buffalo,

State University of New York

Abstract: Tests for uniformity of distribution for data vectors on the d-dimensional hypersphere are proposed. The tests are U-statistic and V-statistic estimates of the quadratic distance between the hypothesized, under the null, uniform distribution on the sphere and the empirical cumulative distribution function. We introduce a class of *diffusion* kernels and study in detail a special member of this class, the Poisson kernel, on which our proposed tests of uniformity are based. We obtain the Karhunen-Loève decomposition of the kernel, connect it with its degrees of freedom, and hence with the power of the test via a tuning parameter, *the diffusion parameter*. We propose an algorithm that allows one to select the tuning parameter, and study the connection between the Poisson kernel-based tests and the Sobolev tests. We then study the performance of the proposed tests in terms of level and power, for a number of alternative distributions. Our simulations show that the proposed methods are powerful and outperform the Rayleigh, Giné, Ajne and Bingham test procedures in the case of multimodal alternatives. We apply the new methods to test uniformity of data on the orbits of comets obtained from the NASA website.

Key words and phrases: Diffusion kernels, directional data, exit on the sphere distribution,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

multimodal alternatives, Poisson kernel, spherical data, testing for uniformity.

1. Introduction

In many applications of interest, data are represented as unit vectors in a high-dimensional space or as points on a hyper-sphere. The area of directional statistics deals with data that belong to the unit hypersphere $S^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|^2 = 1\}$ of $\mathbb{R}^d$. Many non-directional datasets can be usefully re-expressed in the form of directions and analyzed as such data (Golzy and Markatou (2020)). For example, in gene expression analysis, standardized gene expressions that have mean zero and variance 1 can be interpreted as directional data. This standardization is applicable when one is interested in gene expression variation under different conditions (Dortet-Bernadet and Wicker (2008)).

Assessing the presence of uniformity is one of the important initial modeling questions related to the analysis of data on the sphere. This question is formalized as a test of uniformity on S^{d-1}. There is a considerable amount of work on testing for uniformity and tests for dimensions $d \geq 2$ can be found in the literature. García-Portugués and Verdebout (2018) provide a reasonably exhaustive overview of uniformity tests on the hypersphere.

One aspect that is not well studied in the literature is the performance of tests of uniformity in the presence of multiple modes in the data. In this paper, we propose tests of uniformity that exhibit high power in the presence of multiple-modal spherical data,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

and compare their performance with that of other well-known tests for uniformity. The proposed tests are based on an important class of distance kernels, the class of diffusion kernels, the most prominent example of which is the normal kernel. Diffusion kernels generalize key features of the normal kernel to other sample spaces, enabling one to create useful distances for spherical or binary sequence data. Our tests are U- and V-statistic estimates of quadratic distances and are based on a special diffusion kernel, the Poisson kernel. We will discuss briefly the quadratic distance framework that is used to assess model fit in section 2.

Our contributions are as follows.

1. We introduce a class of special kernels, the class of *diffusion kernels*, which are tunable and allow easy computation.

2. We discuss a special diffusion kernel, the Poisson kernel and its associated densities. Just as Brownian motion generates the normal distribution as the natural homogeneous diffusion model for the Euclidean space, the Poisson kernel appears to be its natural generalization to the sample space S^{d-1}. We derive the eigen-decomposition of the d-dimensional Poisson kernel, compute its centered version and its degrees of freedom. We show that the degrees of freedom of the kernel are a function of the dimension d.

3. We then use the centered kernel to construct *tests of uniformity* that are U- and V-statistic estimates of quadratic distances. We investigate the connection

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

between the Poisson kernel-based tests and the class of Sobolev tests, showing the equivalence between the two classes. We provide an algorithm that allows fast computation of the test statistic and study its level and power and compare the proposed tests with the Rayleigh, Bingham, Giné and Ajne tests via simulation, illustrating the superior performance of the new proposals in the presence of multiple modes.

The paper is organized as follows. Section 2 offers a short literature review of two fundamental strands of literature on which this paper is based. The first corresponds to quadratic distances (Lindsay and Qu (2003); Lindsay et al. (2008)). The second refers to tests of uniformity that are proposed in the literature. We focus on tests that are easily computable and classic in their use, i.e. well-accepted in the statistical literature. Section 3 introduces the class of *diffusion kernels* and studies a special case of these kernels, the Poisson kernel. Sections 4 and 5 present the proposed tests and their performance in terms of level and power for a variety of alternatives, as a function of the number of modes, the concentration parameter and the sample size. Alternatives that we consider include von Mises-Fisher distributions with various values of the concentration parameter, mixtures of Poisson kernel-based densities with different number of modes and mixtures of von Mises-Fisher distributions. Discussion and recommendations are presented in section 6. The online supplement includes proofs of our theoretical results, and further details on our experiments.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

2. Literature review

We begin with a brief review of a framework within which we base our proposed tests of uniformity, and then we continue with relevant literature on tests for uniformity.

2.1 Quadratic distances

Let χ be a sample space and let $du(s)$ be the canonical uniform measure on this space. For example, this could be the Lebesgue measure, the counting measure or the spherical volume measure depending on the application. The fundamental building block of a statistical distance is the function $K(s,t)$, a bounded, symmetric, non-negative definite kernel defined on $\chi \times \chi$. The quadratic distance between two probability distributions F and G is then defined as

$$d_K(F,G) = \int\int K_G(x,y)d(F-G)(x)d(F-G)(y),$$

where G is a distribution whose goodness of fit we wish to assess. An important example of a quadratic distance is Pearson's Chi-squared statistic (Lindsay et al. (2008); Markatou et al. (2017)).

If $x_1, x_2, ..., x_n$ is a random sample with empirical distribution function $\hat{F}$, then one can construct the quadratic distance between the data and the model as $d(\hat{F}, G)$ and write $d_K(\hat{F},G) = \int\int K_{cen}(x,y)d\hat{F}(x)d\hat{F}(y)$, where $K_{cen}(x,y)$ is the *centered kernel* with respect to the distribution G defined as $K_{cen}(x,y) = K(x,y) - K(G,y) - K(x,G) + K(G,G)$, $K(x,G) = \int K(x,y)dG(y)$, $K(G,G) = \int\int K(x,y)dG(x)dG(y)$ (see Lindsay

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

et al. (2008), p.989-990, for further discussion on kernel centering).

A fundamental concept associated with the kernel K is the concept of *degrees of freedom*. The degrees of freedom (DOF) under measure G, of a kernel K are defined as $DOF(K) = \frac{[tr_G(K)]^2}{tr_G(K^2)} = \frac{(\sum \lambda_j)^2}{\sum \lambda_j^2}$, where λ_j are the eigenvalues of the kernel K under measure G. A very important feature of the methodology we discuss here is that we can determine these quantities without the need to find the full spectral decomposition.

Our tests are based on a special kernel, called the Poisson kernel. It is defined on vectors of length one, so it can be applied to directional data as well as to data expressed by standardized vectors. We will discuss the Poisson kernel in section 3. We next present a brief review of tests of uniformity that we use to compare our proposed tests in terms of power and level.

2.2 Tests of uniformity: a brief review

Testing for uniformity is a classical problem that dates back to Bernoulli (1735). The literature contains various tests of uniformity (see Cuesta-Albertos et al. (2009) and García-Portugués et al. (2020)). Classical tests for uniformity include, among others, Rayleigh's test (Rayleigh (1919)), Ajne's (Ajne (1968)), Bingham's (Bingham (1974)) and Giné's (Giné (1975)) tests. Section S1 of the online supplement offers a description of these tests as we use those to carry out comparisons with the newly proposed tests of uniformity in the presence of multiple data modes. The reason we use Rayleigh, Ajne, Bingham and Giné tests is because of ease of computation, particularly when the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

samples are very large, and the data are of higher dimensions.

The aformentioned tests belong to the large class of Sobolev tests introduced in Giné (1975). Considering the importance of this class of tests for uniformity, we briefly introduce the Sobolev tests following the construction in Jupp (2008). The relationship of these tests with the proposed Poisson kernel-based distance tests is investigated in Section 4.1.

According to Giné (1975), Sobolev tests are defined by constructing a mapping $t : \mathcal{X} \to L^2(\mathcal{X}, u)$, square-integrable real-valued functions on $\mathcal{X}$ with respect to the uniform measure u, based on the eigenfunctions of the Laplacian operator. For $k \geq 1$, let E_k be the space of eigenfunctions corresponding to the non-zero eigenvalue λ_k, $\{f_i\}$ be a orthonormal basis of E_k and $d_k = \dim E_k$. Let $\{a_k\}$ be a sequence of real numbers such that

$$\sum_{k=1}^{\infty} a_k^2 d_k < \infty, \tag{2.1}$$

then the mapping is defined as

$$t(x) = \sum_{k=1}^{\infty} a_k t_k(x) \quad \text{with} \quad t_k(x) = \sum_{i=1}^{d_k} f_i(x) f_i.$$

Given a sample $X_1, \ldots, X_n$, the resulting Sobolev test statistic is given by

$$T_n = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \langle t(X_i), t(X_j) \rangle,$$

or equivalently

$$T_n(\{a_k\}) = \frac{1}{n} \sum_{k=1}^{\infty} a_k^2 \sum_{i=1}^{n} \sum_{j=1}^{n} \langle t_k(X_i), t_k(X_j) \rangle, \tag{2.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

where $\langle f, g \rangle$ denotes the inner product on $L^2(\mathcal{X}, u)$.

3. Definition of general diffusion kernels

An important class of distance kernels is the class of diffusion kernels. A prominent example of this class is the normal kernel, the key features of which can be generalized to other sample spaces, enabling one to create useful distances. We concentrate on a special diffusion class of kernels, the Poisson kernel class. Section S2 of the online supplement discusses the Poinsson kernel's normalized version as a density function on the d-dimensional sphere.

An approach to defining diffusion kernels on continuous spaces is based on concepts from mathematical physics and in particular on the use of the heat equation. Lafferty and Lebanon (2005) used the heat equation on statistical manifolds to define diffusion kernels. This approach can be useful for constructing kernels with desirable physical properties. Here we take a more pragmatic approach and define kernels through the properties we seek. We offer as an example of a diffusion kernel, the normal kernel but other examples generalize the key features of the Gaussian kernel to other sample spaces. Recall that the normal kernel satisfies the equation

$$K_{t_1+t_2}(x, y) = \int K_{t_1}(x, s) K_{t_2}(s, y) ds, \tag{3.1}$$

where we use t_1, t_2 to denote the associated variance parameters. We call the above

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

equation the *diffusion equation*, and we define a diffusion kernel class through the mathematical properties we would like this class to have as follows.

Definition 1. A family of symmetric kernels $K_t(r,s)$ defined on $\chi \times \chi$, where χ is a sample space, with parameter $t \in (0,\infty)$ will be called *a diffusion kernel family* with respect to a measure γ if the following properties hold.

1. The kernel is nonnegative valued, that is, $K_t(r,s) \geq 0$
2. The kernel satisfies the diffusion equation (3.1) in the time parameter.
3. The kernel satisfies the following equations

$$\int K_t(r,s)d\gamma(s) = 1, \text{ and } \int K_t(r,s)d\gamma(r) = 1,$$

that is, it is a probability density under measure $d\gamma$ in either argument.

Note that the diffusion equation implies that there exists a square root kernel $K_{t/2}$ in the same family of kernels. The implication of this statement is that the kernels are always conditionally non-negative definite, and hence they generate quadratic distances. See Lindsay et al. (2014) for a definition of square root kernels.

Our interest in diffusion kernels as defined here is motivated by the fact that these kernels are tunable and thus allow easy computation.

Definition 2. We say a kernel $K(x,y)$ is a canonical diffusion kernel if it has a representation of the form

$$K(x,y) = 1 + e^{-t}\xi_1(x)\xi_1(y) + e^{-2t}\xi_2(x)\xi_2(y) + \cdots$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

This kernel representation provides an eigenanalysis of the kernel with eigenvalues given by $e^{-kt}, k \in \{0, 1, 2, \cdots\}$ and eigenfunctions $\xi_i, i = 1, 2, 3, \cdots$. We call this representation a *geometric spectral decomposition*. As we will see below, the Poisson kernel in dimensions $d = 2$ exhibits this structure.

A specific diffusion kernel that we introduce and discuss in detail below is the Poisson kernel. In what follows, we study this kernel in the d-dimensional case and use it to construct tests for uniformity. Here, we show how the two-dimensional case can be written in the form presented in definition 2.

3.1 Poisson kernel in dimension d=2

Lindsay et al. (2008) defined the Poisson kernel in dimension $d = 2$ as follows:

$$P_\rho(\theta, \phi) = \frac{1 - \rho^2}{1 - 2\rho cos(\theta - \phi) + \rho^2}, \tag{3.2}$$

where $0 < \rho < 1$ and $0 \leq \theta, \phi < 2\pi$; ϕ is a fixed angle or direction and can be thought of as a location parameter and ρ is a concentration parameter. The sample space is the interval $[0, 2\pi)$ and the baseline measure is the uniform distribution on $[0, 2\pi)$.

The functions $\left\{1, \sqrt{2}cos\theta, \sqrt{2}sin\theta, \sqrt{2}cos(2\theta), \sqrt{2}sin(2\theta), \cdots\right\}$ form a basis under the L_2 distances. This is easy to see if we can show that $\frac{1}{2\pi}\int_0^{2\pi} \gamma_i\gamma_j d\theta = 0$ if $i \neq j$ and $\frac{1}{2\pi}\int_0^{2\pi} \gamma_i\gamma_j d\theta = 1$ if $i = j$, where γ_i is the i^{th} function from the above set. But this is indeed the case, since the corresponding integrals all equal 0 when $i \neq j$, and equal 1 when $i = j$. Furthermore, we can write the kernel $P_\rho(\theta, \phi) = \sum \lambda_i \gamma_i^T(\theta)\gamma_i(\phi)$, where

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

$\gamma_i(\theta) = (\sqrt{2}cos(i\theta), \sqrt{2}sin(i\theta))^T$, with $\gamma_i^T(\theta)\gamma_i(\theta) = 2cos(i\theta)cos(i\phi) + 2sin(i\theta)sin(i\phi)$, hence the kernel is given by the expression

$$P_\rho(\theta, \phi) = 2\sum_{i=0}^{\infty} \lambda_i \left\{cos(i\theta)cos(i\phi) + sin(i\theta)sin(i\phi)\right\}, \tag{3.3}$$

and with $\lambda_i = e^{-i\eta} = \rho^i$, $\rho = e^{-\eta}$, η is the smoothing parameter, the aforementioned formulas return the Poisson kernel in the two-dimensional case.

3.2 Poisson kernel in dimension d

We define now the general form of the d-dimensional kernel.

Let $\mathcal{B}^d = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| < 1\}$ and let $\mathcal{S}^{d-1}$ represent the unit sphere, the boundary of $\mathcal{B}$. Then, the d-dimensional Poisson kernel is defined as

$$P(\mathbf{x}, \mathbf{y}) = \frac{1 - \|\mathbf{x}\|^2}{\|\mathbf{x} - \mathbf{y}\|^d}, \qquad \mathbf{y} \in \mathcal{S}^{d-1}, \mathbf{x} \in \mathcal{B}^d.$$

First, we observe that this is not a symmetric kernel with the right integration properties. However, if we define a kernel on $\mathcal{S}^{d-1} \times \mathcal{S}^{d-1}$ by the expression

$$P_\rho(\mathbf{u}, \mathbf{v}) = P(\rho\mathbf{u}, \mathbf{v}) = \frac{1 - \rho^2}{(1 + \rho^2 - 2\rho(\mathbf{u} \cdot \mathbf{v}))^{d/2}}, \tag{3.4}$$

where $\mathbf{u}, \mathbf{v} \in \mathcal{S}^{d-1}$, we obtain a symmetric nonnegative kernel that integrates to one with respect to the uniform measure on the d-dimensional unit sphere.

Next, define the Poisson integral of a differentiable function f as $P[f](\mathbf{x}) = \int_{\mathcal{S}^{d-1}} f(\boldsymbol{\zeta})P(\mathbf{x}, \boldsymbol{\zeta})d\sigma(\boldsymbol{\zeta})$, where σ is the unique Borel probability measure on $\mathcal{S}^{d-1}$, the unit sphere, that is rotation invariant for every set $E \subset \mathcal{S}^{d-1}$. This integral

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

will be used in some of the proofs when discussing the Poisson kernel as a density function.

We need the following lemma.

Lemma 1. *Let $P(\mathbf{x},\mathbf{y}), P(\mathbf{z},\mathbf{y})$ be Poisson kernels defined on $\mathcal{S}^{d-1}$, the d-dimensional unit sphere. Then*

$$\int P(\mathbf{x},\mathbf{y})P(\mathbf{z},\mathbf{y})d\sigma(\mathbf{y}) = \frac{1-\|\mathbf{x}\|^2\cdot\|\mathbf{z}\|^2}{(1-2\mathbf{x}\cdot\mathbf{z}+\|\mathbf{x}\|^2\cdot\|\mathbf{z}\|^2)^{d/2}}.$$

Proposition 1. *Let $P_{\rho_1}(\mathbf{x},\boldsymbol{\mu}), P_{\rho_2}(\mathbf{z},\boldsymbol{\mu})$ be two Poisson kernel densities. Then*

$$\int P_{\rho_1}(\mathbf{x},\boldsymbol{\mu})P_{\rho_2}(\boldsymbol{\mu},\mathbf{z})d\sigma(\boldsymbol{\mu}) = P_{\rho_1\rho_2}(\mathbf{x},\mathbf{z}),$$

where $\sigma(\cdot)$ denotes the uniform measure on the sphere, $\mathbf{x},\mathbf{z}\in\mathcal{S}^{d-1}, 0<\rho_1<\rho_2<1$.

Proof. When $\|\mathbf{x}\|=1, \|\boldsymbol{\mu}\|=1, \|\mathbf{x}-\rho\boldsymbol{\mu}\| = 1-2\rho\mathbf{x}\cdot\boldsymbol{\mu}+1 = \|\rho\mathbf{x}-\boldsymbol{\mu}\|$. Write $P_{\rho_1}(\mathbf{x},\boldsymbol{\mu}) = P(\rho_1\mathbf{x},\boldsymbol{\mu})$ and $P_{\rho_2}(\boldsymbol{\mu},\mathbf{z}) = P(\rho_2\boldsymbol{\mu},\mathbf{z})$. Then apply the aforementioned lemma to obtain:

$$\begin{aligned}\int_{\mathcal{S}} P(\rho_1\mathbf{x},\boldsymbol{\mu})P(\rho_2\boldsymbol{\mu},\mathbf{z})d\sigma(\boldsymbol{\mu}) &= \frac{1-(\rho_1\rho_2)^2\|\mathbf{x}\|^2\cdot\|\boldsymbol{\mu}\|^2}{(1-2\rho_1\mathbf{x}\cdot\mathbf{z}+\|\rho_1\mathbf{x}\|^2\cdot\|\mathbf{z}\|^2)^{d/2}}\\ &= P_{\rho_1\rho_2}(\mathbf{x},\mathbf{z}).\end{aligned}$$

□

This property is called the convolution property of the Poisson kernel. If the time parameter $t=-\log\rho$, this property gives the same additive convolution closure property with the one provided for example, by the normal distribution.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

3.3 Eigendecomposition of the d-dimensional Poisson kernel: fundamental ideas

In this section, we show that this decomposition can be carried out for dimension $d > 2$ by using polynomials known as zonal harmonics and elements from harmonic analysis. Just as Brownian motion generates the normal distribution as the natural homogeneous diffusion model for the Euclidean space, the Poisson kernel appears to be its natural generalization on the sample space $\mathcal{S}^{d-1}$. Section S2 of the supplementary material briefly reviews the connection between the Poisson kernel based densities and Brownian motion. Some of the notation used below is presented in Section S3.1 of the supplementary material.

We now present the main results of this section. To establish their proof we present some lemmas below, the proof of which is included in the online supplement, section S3.

Lemma 2. *Let $Z_m(\mathbf{u}, \boldsymbol{\zeta})$, $\mathbf{u} \in \mathcal{B}^d$, $\boldsymbol{\zeta} \in \mathcal{S}^{d-1}$ be a zonal harmonic of degree m. Then*

$$\int_{\mathcal{S}^{d-1}} Z_m(\mathbf{u}, \boldsymbol{\zeta}) d\sigma(\boldsymbol{\zeta}) = 0,$$

where σ is the normalized surface measure on the sphere $\mathcal{S}^{d-1}$.

Lemma 3. *Let $Z_m(\mathbf{u}, \boldsymbol{\zeta})$, be a zonal harmonic. Then*

$$\int Z_m^2(\mathbf{u}, \boldsymbol{\zeta}) d\sigma(\boldsymbol{\zeta}) d\sigma(\mathbf{u}) = d_{d,m}, \forall \mathbf{u} \in \mathcal{B}^d, \boldsymbol{\zeta} \in \mathcal{S}^{d-1}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

Lemma 4. *Let $Z_p(\mathbf{u},\boldsymbol{\zeta})$, $Z_q(\mathbf{u},\boldsymbol{\zeta})$ be two zonal harmonics with $p \neq q$. Then*

$$\int_{\mathcal{S}^{d-1}} Z_p(\mathbf{u},\boldsymbol{\zeta}) Z_q(\mathbf{u},\boldsymbol{\zeta}) d\sigma(\mathbf{u}) d\sigma(\boldsymbol{\zeta}) = 0, \forall \mathbf{u} \in \mathcal{B}^d, \boldsymbol{\zeta} \in \mathcal{S}^{d-1} \text{ and } \forall p \neq q.$$

Theorem 1. *For every $d \geq 2$ the Poisson kernel $P(\mathbf{x},\boldsymbol{\zeta})$, $\mathbf{x} \in \mathcal{B}^d$, $\boldsymbol{\zeta} \in \mathcal{S}^{d-1}$, $\mathbf{x} = r\mathbf{u}$, $\mathbf{u} \in \mathcal{S}^{d-1}$ has the following spectral decomposition.*

The eigenspace corresponding to the eigenvalue r^m is $\mathcal{H}_m(\mathcal{S}^{d-1})$. This space has dimension $d_{d,m}$ that is given as follows:

$$d_{d,m} = \begin{cases} \binom{d+m-1}{d-1} - \binom{d+m-3}{d-1} & , m \geq 2 \\ \binom{d+m-2}{d-2} + \binom{d+m-3}{d-2} & , m \geq 1. \end{cases}$$

The projection operator onto $\mathcal{H}_m(\mathcal{S}^{d-1})$ is $Z_m(\mathbf{x},\boldsymbol{\zeta})$, the zonal harmonic of degree m with pole $\boldsymbol{\zeta}$. Therefore, the spectral decomposition of the Poisson kernel is

$$P(r\mathbf{u},\boldsymbol{\zeta}) = \sum r^m Z_m(\mathbf{u},\boldsymbol{\zeta}). \tag{3.5}$$

Moreover, $Z_m(\mathbf{u},\boldsymbol{\zeta})$ has the form

$$Z_m(\mathbf{u},\boldsymbol{\zeta}) = (d+2m-2) \cdot \sum_{k=0}^{[m/2]} (-1)^k \frac{d(d+2)\cdots(d+2m-2k-4)}{2^k k!(m-2k)!} \cdot (\mathbf{u}\cdot\boldsymbol{\zeta})^{m-2k},$$

where $\cdot$ indicates the inner product.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

Proof of Theorem.

The eigenequation is

$$\int_{\mathcal{S}^{d-1}} P(\mathbf{x},\boldsymbol{\zeta})q_m(\boldsymbol{\zeta})d\sigma(\boldsymbol{\zeta}) = \int_{\mathcal{S}^{d-1}} r^m q_m(\boldsymbol{\zeta})Z_m(\mathbf{u},\boldsymbol{\zeta})d\sigma(\boldsymbol{\zeta}),$$

for any $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x} = r\mathbf{u}$ and $q_m(\boldsymbol{\zeta}) \in \mathcal{H}_m(\boldsymbol{\zeta})$, $K(\mathbf{x},\boldsymbol{\zeta})$ denotes the Poisson kernel.

The above equation gives

$$\int_{\mathcal{S}^{d-1}} P(r\mathbf{u},\boldsymbol{\zeta})q_m(\boldsymbol{\zeta})d\sigma(\boldsymbol{\zeta}) = r^m q_m(\boldsymbol{\zeta}).$$

Therefore, lemmas 1,2 and 3 and this last relation indicate that the eigenvalues are r^m, with multiplicity m, and the eigenfunctions are the harmonic polynomials $q_m \in \mathcal{H}_m(\mathcal{S}^{d-1})$. Notice that $q_m(\boldsymbol{\zeta}) = Z_m(\mathbf{u},\boldsymbol{\zeta})$ given in the statement of the theorem. This concludes the proof. □

3.4 Centering the Poisson kernel

In what follows we present the centered Poisson kernel, where the centering is with respect to the normalized uniform measure on the sphere, denoted by σ.

Proposition 2. *Let $P_\rho(\mathbf{x},\mathbf{y})$ be a Poisson kernel defined on the d-dimensional sphere, $\mathcal{S}^{d-1}$. Then, if $\sigma(.)$ denotes the normalized measure on $\mathcal{S}^{d-1}$, the centered with respect to $\sigma(.)$ kernel, is given as*

$$K_{cen}(\mathbf{x},\mathbf{y}) = P_\rho(\mathbf{x},\mathbf{y}) - 1. \tag{3.6}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

Proof. Recall that

$$K_{cen}(\mathbf{x},\mathbf{y}) = P_\rho(\mathbf{x},\mathbf{y}) - \int P_\rho(\mathbf{x},\mathbf{y})d\sigma(\mathbf{y}) - \int P_\rho(\mathbf{x},\mathbf{y})d\sigma(\mathbf{x}) + \iint P_\rho(\mathbf{x},\mathbf{y})d\sigma(\mathbf{x})d\sigma(\mathbf{y}).$$

Proposition 1.2 on p.14 of Axler et al. (2001) shows $\int_{\mathcal{S}^{d-1}} P_\rho(\mathbf{x},\mathbf{y})d\sigma(\mathbf{y}) = 1$, hence $\int_{\mathcal{S}^{d-1}}\int_{\mathcal{S}^{d-1}} P_\rho(\mathbf{x},\mathbf{y})d\sigma(\mathbf{x})d\sigma(\mathbf{y}) = 1$. Therefore, the centered kernel is given as

$$K_{cen}(\mathbf{x},\mathbf{y}) = P_\rho(\mathbf{x},\mathbf{y}) - 1.$$

□

Proposition 3. *The degrees of freedom of the d-dimensional centered Poisson kernel* $K_{cen}(\mathbf{x},\mathbf{y})$ *with respect to the uniform measure* σ *is*

$$DOF(K_{cen}) = \left(\frac{1+\rho}{1-\rho}\right)^{d-1}\left\{\frac{\left(1+\rho-(1-\rho)^{d-1}\right)^2}{1+\rho^2-(1-\rho^2)^{d-1}}\right\}.$$

Furthermore, when $\rho \to 1$, $DOF(K_{cen}) \to \infty$ *and when* $\rho \to 0$, $DOF(K_{cen}) \to d$, *the dimension of the data vector.*

Proof. See section S3 of the online supplementary material. □

As an example, when $d = 2$, the degrees of freedom are

$$DOF(K_{cen}) = 2\left(\frac{1+\rho}{1-\rho}\right).$$

It is easily seen in this case, that when $\rho \to 0, DOF(K_{cen}) \to 2$. Similar results hold when $d = 3$. In this case, $DOF(K_{cen}) = \left(\frac{1+\rho}{1-\rho}\right)^2 \cdot \frac{(3-\rho)^2}{3-\rho^2}$ and as $\rho \to 0, DOF(K_{cen}) \to \frac{9}{3} = 3$, the dimension of the data.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

Figure S1 presents the relationship between the DOF, the concentration parameter (tuning parameter) ρ and the dimension d of the data. Notice that, for fixed dimension d, the DOF increase as the concentration parameter increases. The same pattern is noticed when ρ is kept fixed and the dimension of the data increases (see figure S2).

4. Distance based tests for uniformity and their distributions

We develop *kernel based tests for uniformity* on the sphere using the Poisson kernel. Our tests are based on U-statistic and V-statistic estimates of the quantity $D(F,G) = \int\int P_\rho(\mathbf{x},\mathbf{y})d(F-G)(\mathbf{x})d(F-G)(\mathbf{y})$, where ρ is the tuning parameter in $(0,1)$, G corresponds to the uniform measure and F is a distribution defined on $\mathcal{S}^{d-1}$.

To test the null hypothesis of uniformity on the d-dimensional sphere we propose the following two test-statistics.

Let

$$U_n = \frac{2}{n(n-1)}\sum_{i=2}^{n}\sum_{j=1}^{i-1} K_{cen}(\mathbf{x}_i,\mathbf{x}_j),$$

be a U-statistic estimate of $d(\hat{F},G)$, with $\hat{F}$ corresponding to the empirical cumulative distribution function, and $K_{cen}(\mathbf{x}_i,\mathbf{x}_j)$is given in Proposition 2. Then, the first test statistic is given as

$$T_n = \frac{U_n}{\sqrt{Var(U_n)}},$$

$$Var(U_n) = \frac{2}{n(n-1)}\left[\frac{1+\rho^2}{(1-\rho^2)^{d-1}} - 1\right].$$

The second test statistic we propose is a V-statistic estimate of $d(\hat{F},G)$ and it is

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

given as

$$S_n = nV_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} K_{cen}(\mathbf{x}_i, \mathbf{x}_j).$$

Under a prespecified null hypothesis the asymptotic distribution of the S_n statistic is an infinite combination of independent chi-squared random variables with one degree of freedom, each weighted by $\lambda_j, j = 1, 2, 3, ...,$ the eigenvalues of the Poisson kernel. Following Lindsay et al. (2008) we approximate the distribution $\sum \lambda_j \chi_1^2$ by the distribution $c \cdot \chi^2_{DOF}$, where $c = trace(K^2_{cen})/trace_G(K_{cen}), G$ is uniform, K_{cen} is the centered Poisson kernel. Hence, $c = \frac{(1+\rho^2)-(1-\rho^2)^{d-1}}{(1+\rho)^d-(1-\rho^2)^{d-1}}$ is the constant multiplying the χ^2_{DOF} Satterthwaite approximation of the aforementioned linear combination of independent χ_1^2 random variables, with DOF defined as above.

The corresponding result for the statistic T_n is that it follows a standard normal distribution.

4.1 Relationship with the class of Sobolev tests

We now discuss the relationship between our tests based on the Poisson kernel and the class of Sobolev tests proposed by Giné (1975) and Jupp (2008).

Proposition 4. *The V-statistic version of the Poisson kernel-based tests and the Sobolev tests for uniformity are equivalent on* $\mathcal{S}^{d-1}$. *In particular, if the sequence* $\{a_k^2\}$ *is chosen to be equal to the sequence of eigenvalues of the spectral decomposition of the Poisson kernel, then the Sobolev tests coincide with the V-statistic version of the*

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

Poisson kernel-based tests based on the uncentered kernel.

Proof. Consider the case $d = 2$. Given the Poisson kernel and its spectral decomposition in equations (3.2) and (3.3), and the centered kernel in equation (3.6), the Poisson kernel-based V-statistic can be written as

$$S_n = \frac{2}{n} \sum_{k=0}^{\infty} \rho^k \sum_{i=1}^{n} \sum_{j=1}^{n} \cos k(\mathbf{x}_i - \mathbf{x}_j) - n.$$

On $\mathcal{S}^1$, the functions $\{\sqrt{2}\cos(k\theta), \sqrt{2}\sin(k\theta)\}$ also constitute an orthonormal basis for E_k. Then, by equation (2.2), the Sovolev tests are given as

$$T_n(\{a_k\}) = \frac{2}{n} \sum_{k=1}^{\infty} a_k^2 \sum_{i=1}^{n} \sum_{j=1}^{n} \cos k(\mathbf{x}_i - \mathbf{x}_j).$$

Considering that condition (2.1) is satisfied by the sequence $\{a_k\}$, the obtained tests statistics are equivalent. Notice that, if $a_k^2 = \rho^k$, then the Sobolev test T_n coincides the V-statistic version of the kernel-based distance tests based on the uncentered Poisson kernel.

Consider now the general d-dimensional case, with $d > 2$. According to Jupp (2008), for $\mathbf{x}, \mathbf{y} \in \mathcal{S}^{d-1}$ we have that

$$\langle t_k(\mathbf{x}), t_k(\mathbf{y}) \rangle = \left(1 + \frac{k}{\alpha}\right) C_k^{\alpha}(\langle \mathbf{x}, \mathbf{y} \rangle),$$

with $\alpha = d/2 - 1$, and C_k^{α} denotes the Gegenbauer polynomial of degree k. Then, by equation (2.2), the Sobolev tests are given as

$$T_n(\{a_k\}) = \frac{1}{n} \sum_{k=1}^{\infty} a_k^2 \left(1 + \frac{k}{\alpha}\right) \sum_{i=1}^{n} \sum_{j=1}^{n} C_k^{\alpha}(\langle \mathbf{x}_i, \mathbf{x}_j \rangle).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

Given the Poisson kernel in equation (3.4), its spectral decomposition given in equation (3.5) involves the zonal harmonics. By Theorem 1.2.6 (pag.9) in Dai and Xu (2013), for $d \geq 3$, the Zonal harmonics can be expressed in terms of the Gegenbauer polynomials as $Z_m(x, y) = \frac{m+\alpha}{\alpha} C_m^{\alpha}(\langle \mathbf{x}, \mathbf{y} \rangle)$ with $\alpha = (d-2)/2$. The spectral decomposition of the Poisson kernel can be rewritten as

$$P_\rho(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{\infty} \rho^k \left(\frac{k+\alpha}{\alpha} \right) C_k^{\alpha}(\langle \mathbf{x}, \mathbf{y} \rangle).$$

Hence, the Poisson kernel-based V-statistic can be expressed as

$$S_n = \frac{1}{n} \sum_{k=1}^{\infty} \rho^k \left(\frac{k+\alpha}{\alpha} \right) \sum_{i=1}^{n} \sum_{j=1}^{n} C_k^{\alpha}(\langle \mathbf{x}_i, \mathbf{x}_j \rangle) - n.$$

□

The above proposition shows that the V-statistic version of the Poisson kernel-based tests for uniformity on $\mathcal{S}^{d-1}$, with $d \geq 2$, are equivalent to the Sobolev tests, and the sequence $\{a_k\}$ plays a similar role as the sequence of eigenvalues $\{\lambda_k\}$. This is ensured by condition (2.1), which holds for the $\{\lambda_k\}$. When $a_k^2 = \rho^k$ for $d = 2$, the resulting Sobolev tests have the same expression of the Poisson kernel-based V-statistic tests based on the non-centered kernel. Pycke (2010) has considered this choice and further details can be found in the associated paper.

Remark 1. We note the following. Jupp (2008) proposes a data-driven method to select the sequence a_k; it essentially selects the first $\hat{k}$ values of the sequence $\{a_k\}$ to

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

be equal 1 and the remaining zero. We use the entire sequence $\{a_k\}$, since $a_k^2 = \rho^k$, for $k = 1, 2, \ldots$.

5. Empirical results

We designed a simulation study to evaluate the performance of the proposed tests with the following goals in mind. The first goal refers to understanding the performance of the proposed tests in terms of level and power. The second goal pertains to comparing the new suggested procedures with other, existing in the literature tests, and specifically with the tests proposed by Rayleigh, Bingham, Ajne and Giné. The rationale for these comparisons is that the aforementioned procedures are used in practice and are easy to compute in any dimension.

5.1 Data generation and level computations

Data were generated from a d-dimensional uniform distribution that resides on the d-dimensional unit sphere. The R package "Directional" was used to generate the data ("rsop" function). We generated data on the 2, 3, 4 and 6-dimensional sphere of sample size 100, 500 and 1000 and computed the significance level of our tests and the comparison tests as the proportion of rejections of the null hypothesis of uniformity among the number of the 5000 tests statistics computed. Algorithm S1 of the supplementary material describes in detail how this computation was performed.

The cutoff values of the tests statistics for testing uniformity are computed either

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

using the asymptotic distribution of the statistics, or empirically. Specifically, for tests such as Rayleigh and Bingham, a simple asymptotic distribution exists and is a chi squared with d degrees of freedom for the Rayleigh test and a chi squared with $(d-1)(d+1)/2$ degrees of freedom for the Bingham test. Thus, we use the 95^{th} quantile of these distributions as the cutoff, indicating the value of the test statistic beyond which the null hypothesis of uniformity is rejected. The test statistics proposed by Ajne and Giné do not have such simple distributions. For those, the cutoff is computed empirically using Algorithm S1.

The cutoff of our S_n statistic is obtained by multiplying the χ^2_{DOF} cutoff with $c = \frac{(1+\rho^2)-(1-\rho^2)^{d-1}}{(1+\rho)^d-(1-\rho^2)^{d-1}}$. For the T_n statistic the cutoff is determined empirically, as the 95^{th} quantile of the empirical distribution of the test statistic.

Table S1 of the online supplement presents the results of the level calculation as a function of the dimension, sample size, and in the case of proposed tests tuning parameter ρ. All tests seem to control well the level at the nominal value of 0.05, for all sample sizes, dimensions and test studied.

In the next section we present the results of our study with respect to the power of the tests. We first discuss the selection of the tuning parameter that enters the calculation of T_n, S_n statistics. The tuning parameter is intimately connected with the degrees of freedom, and hence power, of the test statistics.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

Selection of the tuning parameter for the distance-based statistics: In our simulation study, the maximum power for the S_n and T_n statistics is obtained by a grid search that aims to find the value of ρ that produces the maximum power. The algorithm we use to find the tuning parameter ρ is presented below.

Algorithm 1: Algorithm to search for the optimal ρ for T_n and S_n statistics

1. For each MC replication, compute the T_n, S_n statistics for different tuning parameters ρ from 0.01 to 0.99 with step increase of 0.01;
2. For each value of ρ, determine the power of the T_n, S_n statistics;
3. The ρ value that corresponds to the maximum power of the T_n or S_n statistics is the optimal ρ.

Tenreiro (2019) studies the selection of tuning parameters that appear in certain goodness-of-fit tests. His methods are similar to the grid method we use; understanding however the exact connection is a subject for future work. Figure 1 presents the power of the proposed tests for testing uniformity as a function of the tuning parameter. The distribution of the data is PKBD($\rho = 0.2$) and the dimension is equal to 3. Sample size is set to 100 and the number of MC replications is 1000. The figure shows that there is an interval of values of the tuning parameter for which maximum power is obtained. For example, for the S_n statistic, the value 0.16 returns maximum power of 0.838. The DOF that correspond to this value is 5. On the other hand, T_n reaches maximum power of 0.836 when the tuning parameter is 0.03, which corresponds to 3 DOF.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

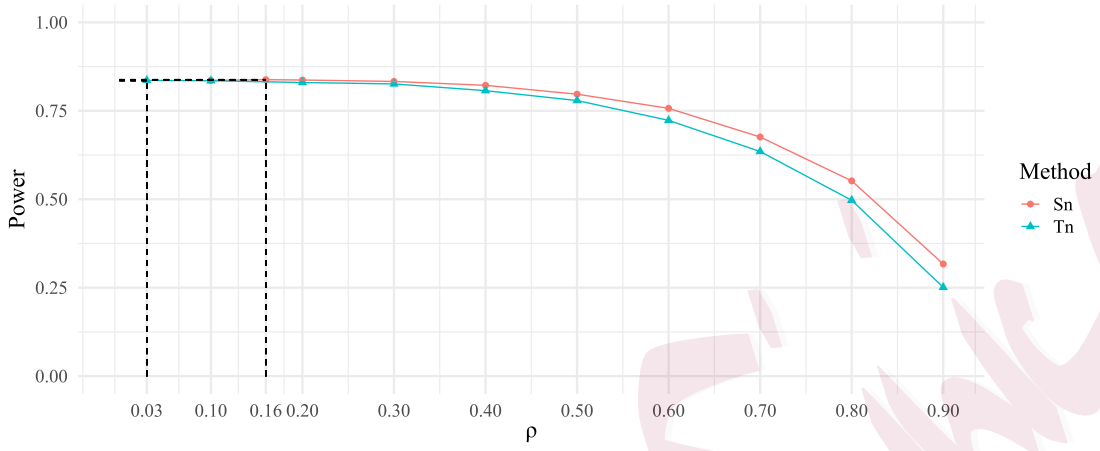

Figure 1: Power of S_n and T_n statistics testing uniformity against the PKBD distribution. The sample size is 100, and $\rho = 0.2$. Dimension of the data is 3.

5.2 Power against von Mises-Fisher distribution

The von Mises-Fisher distribution or Langevin distribution is a multivariate distribution defined on $\mathcal{S}^{d-1}$ and used in the analysis of directional data (Jupp (2008)).

We use the package "movMF" in R version 3.5.3 to generate data from a single von Mises-Fisher distribution (Hornik and Grün (2014)) with dimensions $1, 2, 3, 6$ or 10 and sample size from 50 to 1000, depending on the dimension. The direction of the mean is $(0, 0, ...0, 1)$.

We also study the performance of the tests when data are generated from a mixture of 4 and 8 components (dimension 2) and 6 and 14 components (dimension 3) of the von Mises-Fisher distribution. The concentration parameter κ varies in each scenario. For each case, we compute the test statistics and their power using 2000 MC replications.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

The tuning parameter ρ for S_n and T_n takes values from 0.01 to 0.99 with a 0.01 step increase.

The data distribution is von Mises-Fisher distribution with one mode: The null hypothesis is that of uniformity, while data follow a unimodal von Mises-Fisher distribution with a specified concentration parameter $\kappa \in \{1, 4\}$. Table 1 presents the power of each assessed test statistic in the 2-dimensional or 6-dimensional cases with different concentration parameter κ. The power associated with the S_n and T_n statistics is the maximum power obtained over a grid of values of the tuning parameter ρ. All six tests have good performance in terms of power when the sample size and the value of κ increase. The S_n and T_n tests are competitive with the Ajne and Rayleigh tests. The Bingham and Giné test statistics have lower powers than the other four tests when the concentration parameter is lower, that is, when the data is nearly uniformly distributed.

The data distribution is a mixture of several von Mises-Fisher distributions: Table 2 presents the power of the tests in different scenarios. We study the cases of 4-component and 8-component mixture of von Mises-Fisher distributions of dimension $d = 2$ with larger concentrations of 10, 20 or 45 and sample size of 100 and 500 respectively. The direction of the mean vectors of the 4 von Mises-Fisher distributions are set as $(1, 0)$, $(0, 1)$, $(-1, 0)$ and $(0, -1)$, and the directions of the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

Table 1: Evaluation of tests for uniformity against the von Mises-Fisher distribution in terms of power for 2-dimensional and 6-dimensional data. S_n, T_n are computed using the tuning parameter that produces the maximum power.

Dimension	Number of the modes	κ	Sample size	Bingham	Rayleigh	Ajne	Giné	S_n	T_n
2	1	1	100	0.260	1	1	0.240	1	1
2	1	4	100	1	1	1	1	1	1
6	1	1	100	0.124	0.866	0.866	0.068	0.872	0.868
6	1	4	100	1	1	1	1	1	1
6	1	1	500	0.258	1	1	0.158	1	1
6	1	4	500	1	1	1	1	1	1

mean vectors for the 8 von Mises-Fisher distributions are $(1,0)$, $(0,1)$, $(-1,0)$, $(0,-1)$, $(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}})$, $(-\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}})$, $(-\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}})$, and $(\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}})$. In the 3-dimensional cases, data follow a mixture of 6 or 14 von Mises-Fisher distributions with sample sizes 100 and 500. The direction of the mean vectors of the 6 von Mises-Fisher distributions are the orthogonal vectors $(1,0,0)$, $(0,1,0)$, $(0,0,1)$, $(-1,0,0)$, $(0,-1,0)$, $(0,0,-1)$ and the direction of the mean vectors of the 14 von Mises-Fisher distributions are set as $(1,0,0)$, $(0,1,0)$, $(0,0,1)$, $(-1,0,0)$, $(0,-1,0)$, $(0,0,-1)$, $(\frac{1}{\sqrt{3}},\frac{1}{\sqrt{3}},\frac{1}{\sqrt{3}})$, $(\frac{1}{\sqrt{3}},\frac{1}{\sqrt{3}},-\frac{1}{\sqrt{3}})$, $(\frac{1}{\sqrt{3}},-\frac{1}{\sqrt{3}},\frac{1}{\sqrt{3}})$, $(\frac{1}{\sqrt{3}},-\frac{1}{\sqrt{3}},-\frac{1}{\sqrt{3}})$, $(-\frac{1}{\sqrt{3}},\frac{1}{\sqrt{3}},\frac{1}{\sqrt{3}})$, $(-\frac{1}{\sqrt{3}},\frac{1}{\sqrt{3}},-\frac{1}{\sqrt{3}})$, $(-\frac{1}{\sqrt{3}},-\frac{1}{\sqrt{3}},\frac{1}{\sqrt{3}})$, and

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

$(-\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}})$. The results presented in table 2 clearly indicate the superior performance of the new tests.

5.3 Power against Poisson kernel-based density

To generate data from a Poisson kernel-based density (PKBD) we use the "acceptance-rejection"algorithm introduced by Golzy and Markatou (2020). We study the cases when the underlying distribution is one PKBD (see section S4.2 of the online supplement), and a mixture of two or more PKBDs.

The data distribution is a mixture of two Poisson kernel-based densities: In this simulation experiment, we generate data from the mixture of two PKBDs, with mixing weights equal to 0.5. The combination of concentration parameters (ρ_1, ρ_2) we use is as follows: $(0.1, 0.1)$, $(0.2, 0.2)$, $(0.4, 0.4)$ and $(0.8, 0.8)$ with the sample size varying from 100 to 1000. This selection indicates that the modes of the mixture are more pronounced as the value of the concentration parameter increases. The number of MC replications is 1000 for smaller sample sizes or 200 for larger sample sizes.

We aim to investigate the performance of the tests in terms of power when the angle between the two mean vectors of the PKBDs varies from 0° to 360° by a step of 10°.

Generally, when the concentration parameter becomes larger, that is the underlying distribution of the data is farther from the uniform distribution, the performance of all

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

tests improves. When the dimension and/or the sample size increases, Giné, Bingham, S_n and T_n tests have apparent increases in power. Figures 2 and 3 plot the power of the tests versus the degree of the angle between the two mean vectors when the dimension of the data equals 6 and 10, the sample size varies from 200 to 1000, and the number of MC replications is 1000 or 200. The smoothed power lines are plotted using the b-spline smoothing method in the R package "ggplot2". When the two mean vectors are in opposite directions, that is the angle between the mean vectors equals 180°, the Ajne and Rayleigh tests perform poorly across all sample sizes and dimensions studied. The Bingham and Giné tests perform best when the angle between the two mean vectors is relatively large, i.e. between 140° and 200°. Even in this case, the performance of the T_n and S_n test statistics is equivalent, if not better, than the Bingham and Giné tests for increased sample size. Bingham's and Giné's tests perform poorly when the sample size is relatively small and the concentration parameter of the underlying distribution that generated the data is small, i.e. $\rho \to 0$. The power of these tests increases when sample size and ρ increase and exhibit a periodic property in terms of obtaining low-high power alternatively when the two mean vectors are orthogonal (the angle has degree of 90° or 270°). Our proposed tests exhibit low power when the sample size is small, $\rho \to 0$, and the degree of the angle between the two mean vectors is around 180°. However, even in this case, the power of the tests increases considerably when the sample size increases.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

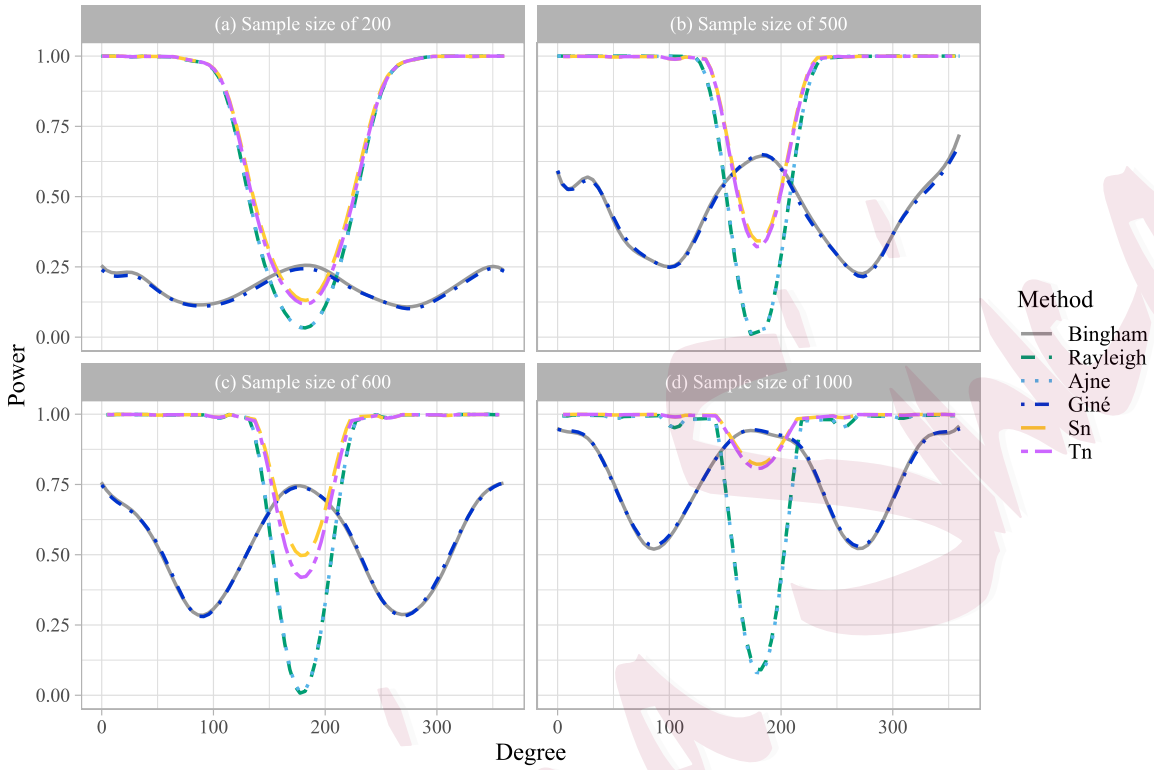

Figure 2: Evaluation of tests for uniformity against a mixture of two PKBDs in terms of power, when the degree of the angle between the two mean vectors of the two PKBD distributions varies. The sample size is 200, 500, 600 and 1000. Dimension of the data is 6 and $\rho_1 = \rho_2 = 0.2$.

The data distribution is a multi-component mixture of Poisson kernel-based densities: We also study the power of tests for uniformity when the underlying distribution is a mixture of more than two PKBDs.

For 2-dimensional data, we study the cases where the data are generated from mixtures of 4 PKBDs and 8 PKBDs. The concentration parameter for data generation is 0.8 for all PKBDs, and the sample sizes are set as 100 and 500, 200 and 500 respectively.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

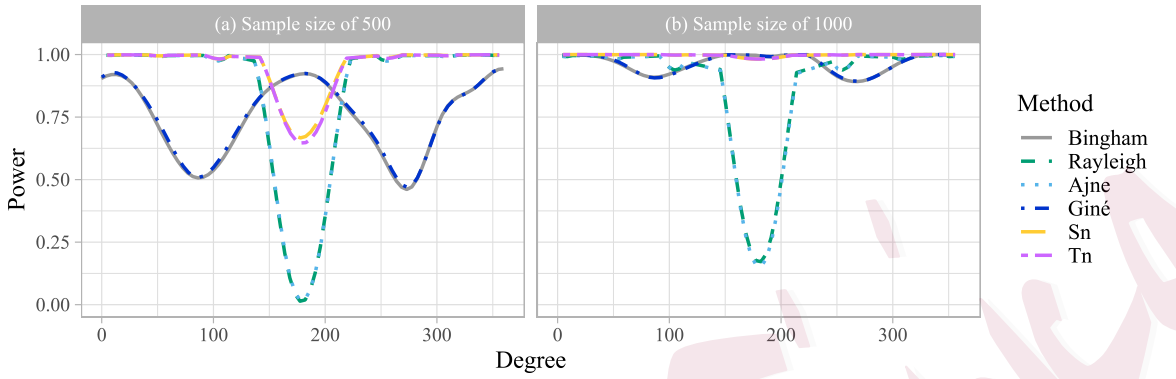

Figure 3: Evaluation of tests for uniformity against a mixture of two PKBDs in terms of power, when the degree of the angle between the two mean vectors of the two PKBD distributions varies. The sample size is 500 and 1000. Dimension of the data is 10 and $\rho_1 = \rho_2 = 0.2$.

The directions of mean vectors for the 4 PKBDs are set as $(1,0)$, $(0,1)$, $(-1,0)$ and $(0,-1)$, and the directions of the mean vectors for the 8 PKBDs are $(1,0)$, $(0,1)$, $(-1,0)$, $(0,-1)$, $(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}})$, $(-\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}})$, $(-\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}})$, and $(\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}})$. The number of MC replications is 1000. Table 3 and figure 4 present the evaluation results in terms of power for all six tests. The performance of S_n and T_n tests is superior than all other four tests in both smaller and larger sample sizes and gain power rapidly when the sample size increases.

Figure 5 shows the power of different tests when the alternative distribution is a mixture of 3 PKBDs. Two mean vectors are in opposite direction while the third one has an angle of 0° to 360° with the first mean vector. In this scenario, $\rho_1 = 0.2, \rho_2 = 0.2, \rho_3 = 0.3$, dimension is 10 and sample size is set as 100 or 200. S_n and T_n tests outperform all other tests.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

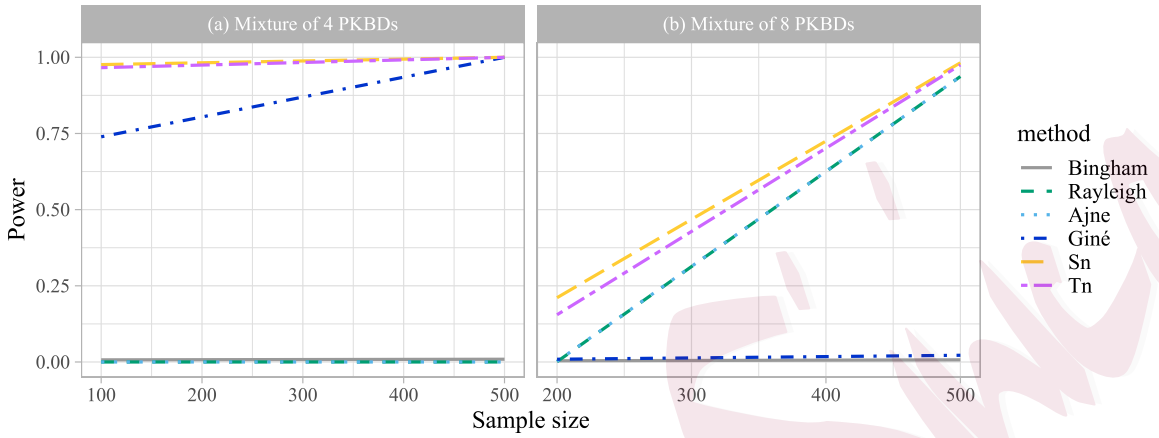

Figure 4: Evaluation of tests for uniformity against the mixture of 4 or 8 PKBDs in terms of power, when the sample size increases. Dimension of the data is 2 and the concentration parameter ρ of each PKBD is 0.8.

5.4 Example

Cuesta-Albertos et al. (2009) introduced a "comet orbits" dataset that is freely available from the NASA website (`http://ssd.jpl.nasa.gov/sbdb_query.cgi#x`) and applied the tests proposed by Giné and Rayleigh. The authors' results show no statistical evidence of rejection of the null hypothesis of uniformity.

We obtained the comet orbits dataset from the NASA website following the procedure described in Cuesta-Albertos et al. (2009). The raw dataset includes the inclination of the orbital plane from the ecliptic (denoted by i) and longitude of the ascending node (denoted by Ω) for 444 comet orbits (accessed on May 15, 2020). The analysis uses 439 data points. The distribution of the sample data in terms of i and Ω

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

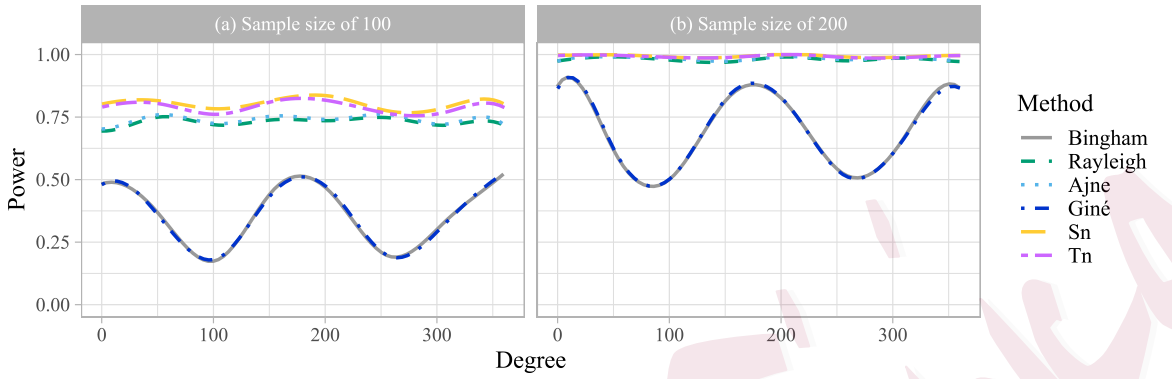

Figure 5: Evaluation of tests of uniformity against a mixture of 3 PKBD distributions in terms of power with mean vectors μ_1, μ_2 and μ_3. The vectors μ_1 and μ_2 are in opposite directions and the degree of the angle between μ_1 and μ_3 varies. Dimension of the data is 10, $\rho_1 = \rho_2 = 0.2, \rho_3 = 0.3$.

is visualized in figure 6. The directed unit normal vector to the orbital plane is given by $(sin\Omega sini, -cos\Omega sini, cosi)'$. Hence in the 3-dimensional space, the directed unit normals of 439 data points are distributed as in figure 7.

Ajne, Giné, Rayleigh, Bingham, S_n and T_n tests are performed on the comet dataset. Ajne, Giné and Rayleigh's tests do not reject the null hypothesis of uniformity, while Bingham, S_n and T_n statistics can detect the small violation of uniformity. Choosing the tuning parameter value corresponds to choosing the DOF. When the tuning parameter $\rho = 0.01$, $DOF = 3$, and S_n, T_n reject the hypothesis of uniformity; larger values of ρ correspond to larger degrees of freedom, so as the tuning parameter increases, the DOF increase. This indicates that the power for detecting slight deviations from uniformity increases with increased DOF, and the null hypothesis of uniformity is rejected.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

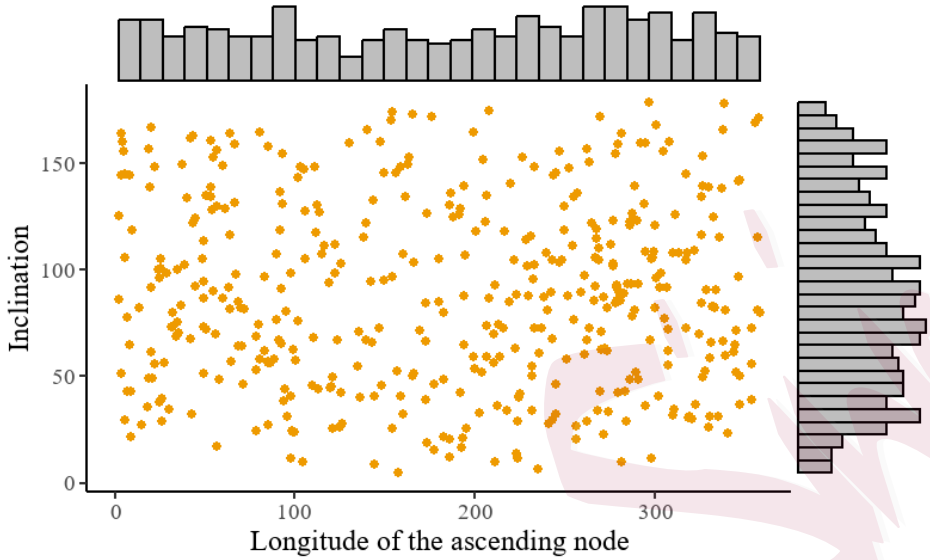

Figure 6: The scatter plot and histograms of the longitude and inclination of the sample comet orbits data. The sample size of the data is 439.

Golzy and Markatou (2020) proposed an EM-type algorithm to estimate the parameters of a PKBD model. We use their method to fit a PKBD model to the comet orbits data. The estimated mean vector is $(-0.70, -0.50, 0.51)$ and the estimated concentrated parameter $\rho = 0.083$, which indicates a very small violation of uniformity on the sphere.

6. Discussion and conclusions

In this paper we introduce a class of kernels, called diffusion kernels, and study in detail one of its members, the Poisson kernel. We then construct tests of uniformity and study the performance of these tests in terms of level and power. Our results

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

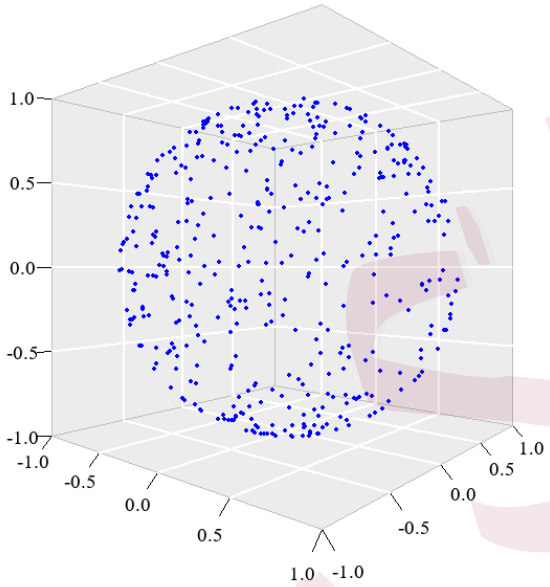

Figure 7: The distribution of the directed unit normals of the sample comet orbits in a 3-dimensional Euclidean space. The size of the data equals 439.

indicate consistency of the tests as the sample size $n \to \infty$. We further propose an algorithm to obtain the tuning parameter that provides the test with maximum power and relate the result to the degrees of freedom of the tests.

The proposed tests for uniformity are powerful and easy to compute. Simulation results show that the proposed tests are competitive with Rayleigh, Ajne, Giné and Bingham's tests and outperform these statistics in the case of multi-modal alternatives. Their performance on a range of alternatives indicates that they can be used in practice in all cases.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

We note that the newly proposed kernel tests, not only can test uniformity, but they can also test hypotheses such as $H_0 : F = PKBD(\rho_0)$, where ρ_0 is a specified parameter. In this case, the test statistics can be obtained from the distance $D_K(F, G)$, where $G = PKBD(\rho_0)$, that is a Poisson kernel based distribution with specified parameter $\rho = \rho_0$ and specified mean direction $\boldsymbol{\mu}$. Therefore, we obtain similar V- and U-statistics as before, with the difference now that our Poisson kernel is centered, not with respect to the uniform distribution, but with respect to the $PKBD(\rho_0)$.

The aforementioned situation falls under the case of a simple null hypothesis because all parameters of the PKBD density are specified. To treat the case of a composite null hypothesis, stated as $H_0 : F = PKBD(\rho, \boldsymbol{\mu})$ is a more complicated, but very practical, task. In this case the parameters of the density $\rho, \boldsymbol{\mu}$ need to be estimated from the data. This action introduces complications due to the dependence of the distance $D_K(\hat{F}, PKBD(\hat{\rho}, \hat{\boldsymbol{\mu}}))$ on the estimated parameters $\hat{\rho}, \hat{\boldsymbol{\mu}}$. In other words, $D_K(\hat{F}, PKBD(\hat{\rho}, \hat{\boldsymbol{\mu}}))$ is no longer a simple quadratic function of $\hat{F}$. Lindsay, Markatou and Ray (2014) discuss a possible way to handle this difficulty, which consists of approximating the distance $D(\hat{F}, PKBD(\hat{\rho}, \hat{\boldsymbol{\mu}}))$ with a quadratic distance that is based on a modified kernel being centered with respect to the estimated density $PKBD(\hat{\rho}, \hat{\boldsymbol{\mu}})$. The details of this approach, as well as a different approach based on a second order von-Mises approximation of the distance, constitute topics for future work.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

Supplementary Material

The supplementary material presents a brief review of the tests of uniformity that exist in the literature. Furthermore, we briefly discuss Poisson kernel-based densities and provide the technical details and proofs that establish the results presented in the main paper. Finally, we provide additional empirical results.

Acknowledgements

This work was funded by a KALEIDA Foundation grant awarded to the second author.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

Table 2: Power of various tests of uniformity when the alternative is a mixture of several von Mises-Fisher distributions of dimension 2 or 3. S_n, T_n are computed using the tuning parameter that produces the maximum power.

Dimension	Number of the modes	κ	Sample size	Bingham	Rayleigh	Ajne	Giné	S_n	T_n
2	4	10	100	0.001	0	0	0.810	**0.980**	**0.980**
2	4	20	100	0	0	0	1	**1**	**1**
2	4	10	500	0.001	0	0	1	**1**	**1**
2	4	20	500	0	0	0	1	**1**	**1**
2	8	20	100	0	0.003	0.002	0	**0.140**	**0.100**
2	8	45	100	0	0	0	0	**0.990**	**0.980**
2	8	20	500	0	1	1	0.001	**1**	**1**
2	8	45	500	0	1	1	1	**1**	**1**
3	6	10	100	0.010	0	0	0.750	**0.990**	**0.980**
3	6	20	100	0	0	0	1	**1**	**1**
3	6	10	500	0.900	0	0	1	**1**	**1**
3	6	20	500	1	0	0	1	**1**	**1**
3	14	20	100	0	0	0	0	**0.970**	**0.950**
3	14	30	100	0	0	0	0.020	**1**	**1**
3	14	20	500	0	0	0	0.250	**1**	**1**
3	14	30	500	0	0	0	1	**1**	**1**

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

Table 3: Evaluation of all tests in terms of power when the alternative distribution is a mixture of 4 or 8 PKBDs. The dimension of the data is 2 and the concentration parameter $\rho_i (i = 1, ..., 4 \; or \; 1, ..., 8)$ of each PKBD is 0.8. S_n, T_n are computed using the tuning parameter that produces the maximum power.

Dimension	Number of the modes	ρ_i	Sample size	Bingham	Rayleigh	Ajne	Giné	S_n	T_n
2	4	0.8	100	0.007	0	0	0.739	**0.976**	**0.966**
2	4	0.8	500	0.009	0	0	1	**1**	**1**
2	8	0.8	200	0.004	0.001	0.002	0.009	**0.211**	**0.155**
2	8	0.8	500	0.007	0.937	0.937	0.022	**0.982**	**0.975**

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

References

Ajne, B. (1968). A simple test for uniformity of a circular distribution. *Biometrika 55*(2), 343–354.

Axler, S., P. Bourdon, and R. Wade (2001). *Harmonic Function Theory*, Volume 137. New York: Springer-Verlag.

Bernoulli, D. (1735). Quelle est la cause physique de l'inclinaison des plans des orbites des planètes? In 'Recueil des pieces qui ont remporté le prix de l'Académie Royale des Sciences de Paris 1734, 93–122', Académie Royale des Sciences de Paris, Paris. *Reprinted in Daniel Bernoulli, Werke 3*, 226–303.

Bingham, C. (1974). An Antipodally Symmetric Distribution on the Sphere. *The Annals of Statistics 2*(6), 1201 – 1225.

Cuesta-Albertos, J. A., A. Cuevas, and R. Fraiman (2009). On projection-based tests for directional and compositional data. *Statistics and Computing 19*(4), 367–380.

Dai, F. and Y. Xu (2013). *Approximation theory and harmonic analysis on spheres and balls*. Springer Monographs in Mathematics, Springer.

Dortet-Bernadet, J.-L. and N. Wicker (2008). Model-based clustering on the unit sphere with an illustration using gene expression profiles. *Biostatistics 9*(1), 66–80.

García-Portugués, E., D. Paindaveine, and T. Verdebout (2020). On optimal tests for rotational symmetry against new classes of hyperspherical distributions. *Journal of the American Statistical Association 115*(532), 1873–1887.

García-Portugués, E. and T. Verdebout (2018). An overview of uniformity tests on the hypersphere. *arXiv preprint arXiv:1804.00286*.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

Giné, E. (1975). Invariant tests for uniformity on compact riemannian manifolds based on sobolev norms. *The Annals of Statistics 3*(6), 1243–1266.

Golzy, M. and M. Markatou (2020). Poisson kernel-based clustering on the sphere: Convergence properties, identifiability, and a method of sampling. *Journal of Computational and Graphical Statistics 29*(4), 758–770.

Hornik, K. and B. Grün (2014). movmf: an r package for fitting mixtures of von mises-fisher distributions. *Journal of Statistical Software 58*(10), 1–31.

Jupp, P. E. (2008). Data-driven sobolev tests of uniformity on compact riemannian manifolds. *The Annals of Statistics 36*(3), 1246–1260.

Lafferty, J. and G. Lebanon (2005). Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research 6*(1), 129–163.

Lindsay, B. G., M. Markatou, and S. Ray (2014). Kernels, degrees of freedom, and power properties of quadratic distance goodness-of-fit tests. *Journal of the American Statistical Association 109*(505), 395–410.

Lindsay, B. G., M. Markatou, S. Ray, K. Yang, and S.-C. Chen (2008). Quadratic distances on probabilities: A unified foundation. *The Annals of Statistics 36*(2), 983–1006.

Lindsay, B. G. and A. Qu (2003). Inference functions and quadratic score tests. *Statistical Science 18*(3), 394–410.

Markatou, M., Y. Chen, G. Afendras, and B. G. Lindsay (2017). Statistical distances and their role in robustness. In *New Advances in Statistics and Data Science*, pp. 3–26. Springer.

Pycke, J.-R. (2010). Some tests for uniformity of circular distributions powerful against multimodal alterna-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0347

tives. *Canadian Journal of Statistics 38*(1), 80–96.

Rayleigh, L. (1919). Xxxi. on the problem of random vibrations, and of random flights in one, two, or three dimensions. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 37*(220), 321–347.

Tenreiro, C. (2019). On the automatic selection of the tuning parameter appearing in certain families of goodness-of-fit tests. *Journal of Statistical Computation and Simulation 89*(10), 1780–1797.

Department of Biostatistics, University at Buffalo, State University of New York, Buffalo, New York. Currently Employed at Eli Lilly and Company, Indianapolis, Indiana.

E-mail: yuxindin@buffalo.edu

Department of Biostatistics, University at Buffalo, State University of New York, Buffalo, New York.

E-mail: markatou@buffalo.edu

Department of Biostatistics, University at Buffalo, State University of New York, Buffalo, New York.

E-mail: gsaracen@buffalo.edu