Multivariate goodness-of-fit on flat and curved spaces via nearest neighbor distances

Bruno Ebner[a,*], Norbert Henze[a], Joseph E. Yukich[b,1]

[a] *Institut für Stochastik, Karlsruher Institut für Technologie, Karlsruhe, Germany*
[b] *Lehigh University, Department of Mathematics, Bethlehem, USA*

Abstract

We present a unified approach to goodness-of-fit testing in $\mathbb{R}^d$ and on lower-dimensional manifolds embedded in $\mathbb{R}^d$ based on sums of powers of weighted volumes of kth nearest neighbor spheres. We prove asymptotic normality of a class of test statistics under the null hypothesis and under fixed alternatives. Under such alternatives, scaled versions of the test statistics converge to the α-entropy between probability distributions. A simulation study shows that the procedures are serious competitors to established goodness-of-fit tests. The tests are applied to two data sets of gamma-ray bursts in astronomy.

Keywords: Multivariate goodness-of-fit test, nearest neighbors, α-entropy, manifold, test for uniformity on a circle or a sphere, Gamma-ray burst data
2000 MSC: Primary 62H15 Secondary 60F05, 60D05

1. Introduction and summary

Nearest neighbor methods have been successfully applied in a variety of fields, such as classification [15], density and regression function estimation [6, 11], and multivariate two-sample testing [18, 33, 39]. Moreover, nearest neighbor methods have also been employed in the context of testing the goodness-of-fit of given data with a distributional model; see [7, 17, 21].

This paper is devoted to a class of universally consistent goodness-of-fit tests based on nearest neighbors. These tests can be applied not only to test for uniformity on a compact domain in $\mathbb{R}^d$, but also to test for a specified density on a m-dimensional manifold embedded in $\mathbb{R}^d$, where $m \le d$. The problem of testing uniformity on manifolds has been considered in [16, 23]. Here, prominent special cases are testing for uniformity on a circle or on a sphere. For an overview of existing methods and modern techniques; see Section 6 of each of the monographs [29, 31]. Regarding related literature to statistics on manifolds, see [5, 12], as well as the references therein.

To be specific, let $\mathcal{M}$ denote a C^1 m-dimensional manifold embedded in $\mathbb{R}^d$, where $m \le d$. $\mathcal{M}$ is endowed with the subset topology and is a closed subset of $\mathbb{R}^d$. Let $\mathrm{d}x$ be the Riemannian volume element on $\mathcal{M}$. A probability density function on $\mathcal{M}$ is a measurable non-negative real-valued function f on $\mathcal{M}$ satisfying $\int_{\mathcal{M}} f(x)\,\mathrm{d}x = 1$. The support $\mathcal{K}(f)$ of f is the smallest closed set $K \subset \mathcal{M}$ such that $\int_K f(x)\,\mathrm{d}x = 1$.

Let $\mathcal{P}(\mathcal{M})$ denote the class of bounded probability density functions f on $\mathcal{M}$, and write $\mathcal{P}_b(\mathcal{M}) \subset \mathcal{P}(\mathcal{M})$ for the subset of probability density functions f such that $\mathcal{K}(f)$ is compact and either (i) $\mathcal{K}(f)$ has no boundary or (ii) $\mathcal{K}(f)$ is a C^1 submanifold-with-boundary of $\mathcal{M}$; we refer to Section 2 of [36] for details. Notice that $\mathcal{K}(f)$ could be an m-sphere (or any ellipsoid) embedded in $\mathbb{R}^d$. Let $\mathcal{P}_c(\mathcal{M})$ denote those probability density functions $f \in \mathcal{P}_b(\mathcal{M})$ which are bounded away from zero on their support.

*Corresponding author
Email addresses: bruno.ebner@kit.edu (Bruno Ebner), norbert.henze@kit.edu (Norbert Henze), joseph.yukich@lehigh.edu (Joseph E. Yukich)
[1] Research supported in part by NSF grant DMS-1406410

In what follows we let X_i, $i \geq 1$, be independent and identically distributed (iid) random variables with density f, defined on a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and we put $\mathcal{X}_n = \{X_1, \ldots, X_n\}$. Given a locally finite subset $\mathcal{X}$ of $\mathcal{M}$ and $x \in \mathcal{X}$, we write $x^{(k)}$ for the kth nearest neighbor (with respect to the Euclidean norm $|\cdot|$) of x among $\mathcal{X} \setminus \{x\}$. Let $v_m = \pi^{m/2}/\Gamma(m/2+1)$ be the volume of the unit m-sphere.

Given a fixed $\alpha \in (0, \infty)$ and a fixed positive integer J, consider the volume score function induced by the J-nearest neighbor distances

$$\xi_J^{(\alpha)}(x, \mathcal{X}) = \sum_{k=1}^{J} (v_m |x - x^{(k)}|^m)^\alpha, \tag{1}$$

i.e., sums of volumes (to power α) of the k nearest neighbor balls around x, $k \in \{1, \ldots, J\}$. When $x \notin \mathcal{X}$, we write $\xi_J^{(\alpha)}(x, \mathcal{X})$ in place of $\xi_J^{(\alpha)}(x, \mathcal{X} \cup \{x\})$. When $\mathcal{X}$ consists of $\Theta(n)$ elements in a compact subset of $\mathcal{M}$, where $an \leq \Theta(n) \leq bn$, $n \geq 1$, for some $0 < a < b < \infty$, we study the re-scaled volume scores

$$\xi_{n,J}^{(\alpha)}(x, \mathcal{X}) = \sum_{k=1}^{J} \{v_m |n^{1/m}(x - x^{(k)})|^m\}^\alpha.$$

Recalling that $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, we consider the random measure

$$\mu_{n,J}^{(\alpha)} = \sum_{X_i \in \mathcal{X}_n} \xi_{n,J}^{(\alpha)}(X_i, \mathcal{X}_n) \delta_{X_i}, \tag{2}$$

with δ_x denoting the Dirac point mass at x. If h is an arbitrary measurable bounded function on $\mathcal{M}$, we write $\langle \mu_{n,J}^{(\alpha)}, h \rangle$ for $\int_{\mathcal{M}} h(x) \, \mathrm{d}\mu_{n,J}^{(\alpha)}(x)$.

Given a fixed $f_0 \in \mathcal{P}(\mathcal{M})$, this paper considers testing goodness-of-fit of the hypothesis

$$\mathcal{H}_0 : \text{ the unknown density of } X_i \text{ is } f_0, \tag{3}$$

against general alternatives, based on the statistic

$$T_{n,J}^{(\alpha)} = \langle \mu_{n,J}^{(\alpha)}, f_0^\alpha \rangle = \sum_{X_i \in \mathcal{X}_n} \xi_{n,J}^{(\alpha)}(X_i, \mathcal{X}_n) \{f_0(X_i)\}^\alpha. \tag{4}$$

Notice that for the special case $m = d$ and $J = 1$, this type of statistic has been studied in [7, 21], but without allowing for lower-dimensional manifolds, and without considering fixed alternatives to $\mathcal{H}_0$.

In Section 2, we prove the asymptotic normality of $T_{n,J}^{(\alpha)}$ as $n \to \infty$ both under $\mathcal{H}_0$ and under fixed alternatives to $\mathcal{H}_0$, and we show that $T_{n,J}^{(\alpha)}/n$ has an almost sure limit under a fixed alternative to $\mathcal{H}_0$. When $\alpha \in (0, 1)$, this limit is, apart from a multiplicative constant, the α-entropy between f and f_0. As a consequence, the statistic $T_{n,J}^{(\alpha)}$ yields a universally goodness-of-fit test of $\mathcal{H}_0$ for each $\alpha \in (0, \infty)$, $\alpha \neq 1$, and each J. The versatility of this class of tests is demonstrated in Section 3, which presents the results of a simulation study comparing our tests with several well-known competitors. The paper concludes with some remarks and open problems.

2. Main results

The limit theory for the statistic (4) may be deduced from general theorems established in [36] and goes as follows.

Theorem 1. *If $f \in \mathcal{P}_c(\mathcal{M})$, $\alpha \in (0, \infty)$, then as $n \to \infty$ we have*

$$T_{n,J}^{(\alpha)}/n \to \sum_{k=1}^{J} \frac{\Gamma(\alpha + k)}{\Gamma(k)} \int_{\mathcal{M}} f_0(x)^\alpha f(x)^{1-\alpha} \, \mathrm{d}x \tag{5}$$

in L^2 and also $\mathbb{P}$-a.s.

Remarks. (i) Notice that the right-hand side of (1) is distribution-free if $\alpha = 1$. Thus, in view of the testing problem (3), it is indispensable to have $\alpha \neq 1$.

(ii) If $\dim \mathcal{M} = d$, if the support $\mathcal{K}(f)$ of f is a convex polyhedron, and if f_0 is the uniform density over $\mathcal{M}$, then the asserted L^2 convergence in (5) is given by Theorem 2 of [42]. That paper, which is based on [35], shows that

$$\mathrm{E}\, \xi_{n,J}^{(\alpha)}(X_1, \mathcal{X}_n) \to \int_{\mathcal{K}(f)} \mathrm{E}\, \xi_J^{(\alpha)}(\mathbf{0}, \zeta_{f(x)})\, f(x)\, \mathrm{d}x \tag{6}$$

holds in L^2 as $n \to \infty$. Here, $\mathbf{0}$ denotes a point at the origin of $\mathbb{R}^m$, and ζ_τ with $\tau \in (0, \infty)$ stands for a homogeneous Poisson process of intensity τ in $\mathbb{R}^m$, with $\mathbb{R}^m$ embedded in $\mathbb{R}^d$ so that the random variable $\xi_J^{(\alpha)}(\mathbf{0}, \zeta_\tau)$ is well-defined. As will be shown in the upcoming proof, the paper [36] upgrades (6) to give convergence of the measures at (2), it provides L^2 and a.s. convergence, and also allows $\mathcal{K}(f)$ to be replaced by a C^1 m-dimensional submanifold of $\mathbb{R}^d$.

(iii) Let $X_1, \ldots, X_n$ be iid random variables with continuous distribution function F in the unit interval $[0, 1]$. To test the hypothesis of uniformity on $[0, 1]$, [25] introduced the statistic $\sum_{i=1}^{n+1} U_i^\alpha$, where $U_i = F(X_{(i)}) - F(X_{(i-1)})$, and $0 = X_{(0)} \leq X_{(1)} \leq \cdots \leq X_{(n)} \leq X_{(n+1)} = 1$ are the order statistics of $X_1, \ldots, X_n$. If F has a continuous density f, we have $U_i \approx f(Y_i)(X_{(i)} - X_{(i-1)})$, where $Y_i = (X_{(i)} + X_{(i-1)})/2$. Since $(X_{(i)} - X_{(i-1)})$ is the volume of a 'one-dimensional sphere' centred at Y_i, the statistic $T_{n,1}^{(\alpha)}$ may be considered a multivariate *analogue* (not *generalization*) of $\sum_{i=1}^{n+1} U_i^\alpha$, since in the univariate case some of the spacings are nearest neighbor distances. From the asymptotic distribution of $\sum_{i=1}^{n+1} U_i^\alpha$ (see, e.g., [44]), it follows that

$$\frac{1}{n} \sum_{i=1}^{n+1} (nU_i)^\alpha \to \Gamma(\alpha + 1) \int_0^1 f(x)^{1-\alpha}\, \mathrm{d}x$$

in probability as $n \to \infty$. This result obviously corresponds to (5) for $J = 1$ and f_0 being the uniform density over $\mathcal{M}$, where $m = d$ and $\mathcal{M}$ has Lebesgue measure one.

(iv) If $\alpha \in (0, 1)$, the integral

$$r_\alpha(f_0, f) = \int_{\mathcal{M}} f_0(x)^\alpha f(x)^{1-\alpha}\, \mathrm{d}x$$

figuring on the right-hand side of (5) is known as the α-entropy between (the distributions associated with) f_0 and f, see [41]. Notice that $1 - r_{1/2}(f_0, f) = H^2(f_0, f)$, where $H(f_0, f)$ is the Hellinger distance between f_0 and f. By Hölder's inequality, $r_\alpha(f_0, f) \leq 1$, with equality if and only if the distributions pertaining to f_0 and f coincide. If $\alpha \in (1, \infty)$, put $W = f_0(X_1)/f(X_1)$, and recall that X_1 has density f. Then $r_\alpha(f_0, f) = \mathrm{E}\, (W^\alpha)$, and, by Jensen's inequality, $r_\alpha(f_0, f) \geq (\mathrm{E}\, W)^\alpha = 1$. As above, equality holds if the distributions associated with f_0 and f are the same.

(v) It follows from (iii) and Theorem 1 that, for fixed $\alpha \in (0, 1)$, a test of fit that rejects the hypothesis $\mathcal{H}_0$ figuring in (3) for *small* values of $T_{n,J}^{(\alpha)}$ is consistent against each fixed alternative density f. If $\alpha \in (1, \infty)$, rejection of $\mathcal{H}_0$ is for *large* values of $T_{n,J}^{(\alpha)}$, and the resulting test is universally consistent.

Before stating variance asymptotics and a central limit theorem we introduce more notation from [36], especially (3.8) and (3.9) of that paper. Given $u \in \mathbb{R}^m$, let $\zeta_\tau^u = \zeta_\tau \cup \{u\}$ be the Poisson process ζ_τ together with a point added at u. We consider an integrated 'covariance' of scores

$$V(\tau) = V^{\xi_J^{(\alpha)}}(\tau) = \mathrm{E}\, \xi_J^{(\alpha)}(\mathbf{0}, \zeta_\tau)^2 + \tau \int_{\mathbb{R}^m} \left[\mathrm{E}\, \xi_J^{(\alpha)}(\mathbf{0}, \zeta_\tau^u) \xi_J^{(\alpha)}(u, \zeta_\tau^{\mathbf{0}}) - \{\mathrm{E}\, \xi_J^{(\alpha)}(\mathbf{0}, \zeta_\tau)\}^2 \right] \mathrm{d}u$$

and an integrated 'add-one cost'

$$\delta(\tau) = \delta^{\xi_J^{(\alpha)}}(\tau) = \mathrm{E}\, \xi_J^{(\alpha)}(\mathbf{0}, \zeta_\tau) + \tau \int_{\mathbb{R}^m} \mathrm{E}\, \{\xi_J^{(\alpha)}(\mathbf{0}, \zeta_\tau^u) - \xi_J^{(\alpha)}(\mathbf{0}, \zeta_\tau)\}\, \mathrm{d}u.$$

As shown in Theorem 3.2 of [36], these integrals are finite. Let $\mathcal{N}(0, \sigma^2)$ denote a mean zero normal random variable with variance σ^2.

Theorem 2. *If $f \in \mathcal{P}_c(\mathcal{M})$ is a.e. continuous and $\alpha \in (0,\infty)$, then*

$$\lim_{n\to\infty} n^{-1}\mathrm{var}(T_{n,J}^{(\alpha)}) = \sigma^2(f_0,f) = \int_{\mathcal{M}} f_0(x)^{2\alpha} V\{f(x)\} f(x)\,\mathrm{d}x - \left[\int_{\mathcal{M}} \delta\{f(x)\} f_0(x)^{\alpha} f(x)\,\mathrm{d}x\right]^2 \in (0,\infty).$$

Moreover, as $n \to \infty$, $\{T_{n,J}^{(\alpha)} - \mathrm{E}\, T_{n,J}^{(\alpha)}\}/\sqrt{n} \rightsquigarrow \mathcal{N}[0,\sigma^2(f_0,f)]$.

Remark. Theorem 2.1 of [1] provides variance asymptotics and a central limit theorem for sums of functions of kth nearest neighbor distances in the special case $m = d$.

We next extend the asymptotic normality results to situations where the Euclidean distance on a manifold is replaced by the geodesic distance. We may do this in a general setting which goes as follows. Let $(\mathbb{X},\mathcal{F})$ be a measurable space equipped with a σ-finite measure Q and a measurable metric $\mathrm{d} : \mathbb{X}\times\mathbb{X} \to [0,\infty)$. Assume that Q has density f and that there is a $\gamma \in (0,\infty)$ such that $\inf_{x\in\mathbb{X}} Q\{B(x,r)\} \geq cr^{\gamma}, r \in [0,\mathrm{diam}\mathbb{X}]$, where $B(x,r)$ is the closed ball centered at $x \in \mathbb{X}$ and having radius r.

Given a fixed $\alpha \in (0,\infty)$ and a fixed positive integer J, consider the volume score function induced by the J-nearest neighbor distances

$$\tilde{\xi}_J^{(\alpha)}(x,\mathcal{X}) = \sum_{k=1}^{J} \{\mathrm{d}(x,x^{(k)})^m\}^{\alpha}$$

as well as the random measures $\tilde{\mu}_{n,J}^{(\alpha)} = \sum_{X_i\in\mathcal{X}_n} \tilde{\xi}_{n,J}^{(\alpha)}(X_i,\mathcal{X}_n)\delta_{X_i}$.

Put

$$\tilde{T}_{n,J}^{(\alpha)} = \langle \tilde{\mu}_{n,J}^{(\alpha)}, f_0^{\alpha}\rangle = \sum_{X_i\in\mathcal{X}_n} \tilde{\xi}_{n,J}^{(\alpha)}(X_i,\mathcal{X}_n)\,\{f_0(X_i)\}^{\alpha}\,.$$

The next theorem follows directly from Theorem 5.1 of [28] as well as Remark (iii) in Section 2 of that paper. It shows that, at least in principle, one can also use intrinsic nearest neighbor distances for testing goodness-of-fit on lower-dimensional manifolds embedded in $\mathbb{R}^d$. We assume that $f \in \mathcal{P}_c(\mathbb{X})$, that is f is bounded away from zero and infinity.

Theorem 3. *If $f_0 \in \mathcal{P}(\mathbb{X})$, $\alpha \in (0,\infty)$, and $\mathrm{var}\tilde{T}_{n,J}^{(\alpha)} \geq Cn^{1-2\alpha/\gamma}$, then as $n\to\infty$, $\{\tilde{T}_{n,J}^{(\alpha)} - \mathrm{E}\,\tilde{T}_{n,J}^{(\alpha)}\}/\sqrt{\mathrm{var}\tilde{T}_{n,J}^{(\alpha)}} \rightsquigarrow \mathcal{N}(0,1)$.*

Remarks. (i) Showing the variance lower bound $\mathrm{var}\,\tilde{T}_{n,J}^{(\alpha)} \geq Cn^{1-2\alpha/\gamma}$ is a separate problem. Notice however that when $\mathbb{X}$ is $\mathbb{R}^d$, then we may put γ to be d, and we may set the metric d to be the Euclidean metric. Then the variance lower bound holds (at least for the case $J = 1$), as shown in Theorem 2.1 and Lemma 6.3 of [34].

(ii) Theorem 5.1 of [28] shows a rate of normal convergence in the Kolmogorov distance equal to $Cn^{-1/2}$, C a generic constant.

Proof of Theorem 1. We deduce this from Theorem 3.1 of [36] with $\rho = \infty$, especially display (3.16) of [36], with the f in (3.16) of [36] set to f_0^{α} and with the κ in (3.16) of [36] set to f. Observe that $\xi_J^{(\alpha)}$ belongs to the class $\Sigma(k,r)$ of that paper, and notice that

$$\sup_n \mathrm{E}\,\{\xi_{n,J}^{(\alpha)}(X_1,\mathcal{X}_n)\}^p < \infty$$

holds for all $p \in [1,\infty)$, i.e., the moment condition (3.4) of [36] holds for all p.

The limit (3.16) of [36] tells us that as $n\to\infty$ we have convergence in L^2

$$T_{n,J}^{(\alpha)}/n \to \int_{\mathcal{M}} f_0(x)^{\alpha}\mathrm{E}\,\xi_J^{(\alpha)}(\mathbf{0},\zeta_{f(x)})f(x)\,\mathrm{d}x, \tag{7}$$

where $\xi_J^{(\alpha)}(\mathbf{0},\zeta_{f(x)})$ is defined at (1). The last assertion in Theorem 3.1 of [36] also gives a.s. convergence in (7).

Given $\tau \in (0,\infty)$ and ζ_τ, we let $X_\tau^{(k)} \in \zeta_\tau$ be the kth nearest neighbor to the origin. We compute

$$\begin{aligned}\mathrm{E}\,\xi_J^{(\alpha)}(\mathbf{0},\zeta_\tau) = \sum_{k=1}^{J} \mathrm{E}\,(v_m|X_\tau^{(k)}|^m)^\alpha &= \sum_{k=1}^{J} v_m^\alpha(\tau^{-1/m})^{\alpha m}\mathrm{E}\,(|X_1^{(k)}|)^{\alpha m} \\ &= \left(\frac{v_m}{\tau}\right)^\alpha \sum_{k=1}^{J} v_m^{-(\alpha m)/m}\frac{\Gamma(k+\alpha)}{\Gamma(k)} = \tau^{-\alpha}\sum_{k=1}^{J}\frac{\Gamma(k+\alpha)}{\Gamma(k)},\end{aligned}$$

where the penultimate equality follows by display (15) of [42] (with α replaced by αm, d replaced by m). We have thus shown

$$\mathrm{E}\,\xi_J^{(\alpha)}(\mathbf{0},\zeta_\tau) = \tau^{-\alpha}\sum_{k=1}^{J}\frac{\Gamma(k+\alpha)}{\Gamma(k)}. \tag{8}$$

Letting τ equal $f(x)$ in (7) and applying (8) gives the claimed limit (5). □

Proof of Theorem 2. This is an immediate consequence of Theorem 3.2 of [36] as well as remark (iv) on p. 2174 of [36]. In that remark we may set the function f there to f_0^α, we set $\rho = \infty$, and we put $\mu_{n,k,\rho}^{\xi}$ equal to $\mu_{n,J}^{(\alpha)}$. Keeping ρ set to infinity, it is straightforward to show that $\mu_{n,J}^{(\alpha)}$ satisfies the moment conditions (3.5) and (3.6) of [36]. Since $\mu_{n,J}^{(\alpha)}$ satisfies all the conditions of remark (iv) on p. 2174 of [36], Theorem 2 follows as desired. □

3. Simulations

By means of a simulation study, this section compares the finite-sample power performance of the test based on $T_{n,J}^{(\alpha)}$ with that of several competitors. All simulations are performed using the statistical computing environment R, see [38]. We consider testing for uniformity on the unit square $[0,1]^2$, on the unit circle $\mathcal{S}^1 = \{x \in \mathbb{R}^2 : |x| = 1\}$, and on the unit sphere $\mathcal{S}^2 = \{x \in \mathbb{R}^3 : |x| = 1\}$. Since, strictly speaking, there is not only one new test, but a whole family of tests that depend on the choice of the power α and the number J of neighbors taken into account, the impact on finite-sample power of α and J will be of particular interest. In each scenario, we consider the sample sizes $n = 50$, $n = 100$ and $n = 200$, and the nominal level of significance is set to 0.05. Throughout, critical values for $T_{n,J}^{(\alpha)}$ under $\mathcal{H}_0$ have been simulated with 100,000 replications (see Tables 7–9 in [13]), and each entry in a table referring to the power of the test is based on 10,000 replications.

At least in principle, the result of Theorem 2 can be used to construct a region of rejection by means of the normal distribution, or to specify approximate p-values. To this end, we first have to compute $\mathrm{E}\,T_{n,J}^{(\alpha)}$, which depends on the parameters as well as on the underlying manifold. As an example we consider the uniform density $f_0 = f$ for the three mentioned cases. Straightforward calculations give

$$\mathrm{E}\,T_{n,J}^{(\alpha)} = n^{1+\alpha}\sum_{k=1}^{J}\sum_{j=0}^{k-1}\binom{n-1}{j}\int_0^1 t^{j/\alpha}(1-t^{1/\alpha})^{n-1-j}\,\mathrm{d}t = \frac{n^\alpha}{\alpha+1}(n+1)(J+1+\alpha)\,\frac{\mathcal{B}(n,J+\alpha)}{\mathcal{B}(n+\alpha,J)},$$

where $\mathcal{B}(\cdot,\cdot)$ denotes the Beta function. If α is an integer this formula reduces to

$$\mathrm{E}\,T_{n,J}^{(\alpha)} = \alpha\frac{n^\alpha}{\prod_{\ell=1}^{\alpha-1}(n+\ell)}\sum_{k=1}^{J}\sum_{j=0}^{k-1}\prod_{i=1}^{\alpha-1}(j+i).$$

Since the limit variance $\sigma^2(f_0,f_0)$ does not seem to be available in a closed form that would be amenable to computations, we decided to estimate $\sigma^2(f_0,f_0)$ by means of simulations. Table 1 shows the resulting values for the torus as well as the sphere based on 100,000 replications. With these estimated values we are able to approximately standardize $T_{n,J}^{(\alpha)}$.

Table 1: Estimated $\sigma^2(f_0, f_0)$ under $\mathcal{H}_0$ (torus, sphere)

α	0.5					2				
$n\backslash J$	1	2	3	4	5	1	2	3	4	5
50	0.22	0.75	1.60	2.77	4.25	13.8	93.7	355	995	2326
100	0.22	0.76	1.61	2.77	4.26	14.7	101	384	1085	2545
200	0.22	0.75	1.60	2.75	4.19	15.2	105	394	1099	2586
500	0.22	0.74	1.59	2.76	4.27	15.4	106	404	1138	2657
1000	0.22	0.76	1.62	2.81	4.33	16.0	109	412	1158	2736
1500	0.22	0.76	1.65	2.84	4.33	15.8	109	418	1177	2765

3.1. Unit square $[0,1]^2$

For testing the hypothesis $\mathcal{H}_0$ that the distribution of X_1 is uniform over the unit square $[0,1]^2$, we considered the following competitors to the new test statistic.

(i) The Distance to Boundary Test DB (see [4]), which is based on the distance of $X_1, \ldots, X_n$ to the boundary ∂W of $W = [0,1]^2$. Writing $D_B(y, \partial W) = \min\{|x - y| : x \in \partial W\}$ for the distance of $y \in W$ to ∂W and $R = \max\{D_B(x, \partial W) : x \in W\}$ for the largest of such distances (which equals 0.5 in our case), the test statistic computes, for each $j \in \{1, \ldots, n\}$, $Y_j = D_B(X_j, \partial W)/R$. Under $\mathcal{H}_0$ the random variables $Y_1, \ldots, Y_n$ have a $\mathcal{B}(1,2)$ distribution. The test employs the Kolmogorov–Smirnov type statistic

$$DB_n = \sqrt{n} \sup_{y \in [0,1]} |G_n(y) - G_0(y)|.$$

Here, G_n is the empirical distribution function of $Y_1, \ldots, Y_n$, and G_0 is the distribution function of the $\mathcal{B}(1,2)$ distribution. Rejection of $\mathcal{H}_0$ is for large values of DB_n, and critical values can be taken from the Kolmogorov distribution. Note that this test is not consistent against some easily computable alternatives, e.g., the uniform distribution on the subset $[0.5, 1]^2$ of W.

(ii) The Maximal Spacing Test MS, see [2]. Writing $B(x, r)$ for an open ball centered at x with radius r, this test considers the maximum radius

$$\Delta_n = \sup\{r > 0 : \text{ there is some } x \text{ with } B(x, r) \subset [0,1]^2 \setminus \mathcal{X}_n\}$$

of a ball that does not contain any of $X_1, \ldots, X_n$ as an inner point. Rejection of $\mathcal{H}_0$ is for large values of the test statistic $V_n = \pi\Delta_n^2$. The limit distribution of V_n under $\mathcal{H}_0$ follows from (2.5) of [22], which states that, as $n \to \infty$,

$$nV_n - \ln n - \ln\ln n \rightsquigarrow G,$$

where the random variable G follows a Gumbel distribution with distribution function $\exp\{-\exp(-x)\}$, $x \in \mathbb{R}$. Letting u_α denote the $(1-\alpha)$-quantile of this distribution, the test rejects $\mathcal{H}_0$ at asymptotic level α if

$$V_n > n^{-1}(u_\alpha + \ln n + \ln\ln n).$$

A conceptual proof of the consistency of this test against general alternatives is given in [19].

Since dealing with nearest neighbors in the square involves boundary effects (see, e.g., [10]), we initially employed both the Euclidean metric and the torus metric, i.e., the Euclidean metric on the 3d-torus, obained as the quotient of the unit square by pasting opposite edges together via the identifications $(x, y) \sim (x+1, y) \sim (x, y+1)$. Because the power of the tests was in general somewhat higher for the torus metric than for the Euclidean metric, we decided to use the torus metric. It should be stressed that this choice conforms to the general set-up adopted in [36] so that Theorem 1 and Theorem 2 remain valid.

An empirical study of uniformity tests in several settings including the hypercube can be found in [37]. Guided by the simulation study in [3], we used a contamination and a clustering model as alternatives to the uniform distribution. The contamination model, denoted by CON, for the distribution of X_1 is the mixture

$$(1 - \varepsilon_1 - \varepsilon_2)\mathcal{U}[0,1]^2 + \varepsilon_1 \mathcal{N}_2(c_1, \sigma_1^2) + \varepsilon_2 \mathcal{N}_2(c_2, \sigma_2^2),$$

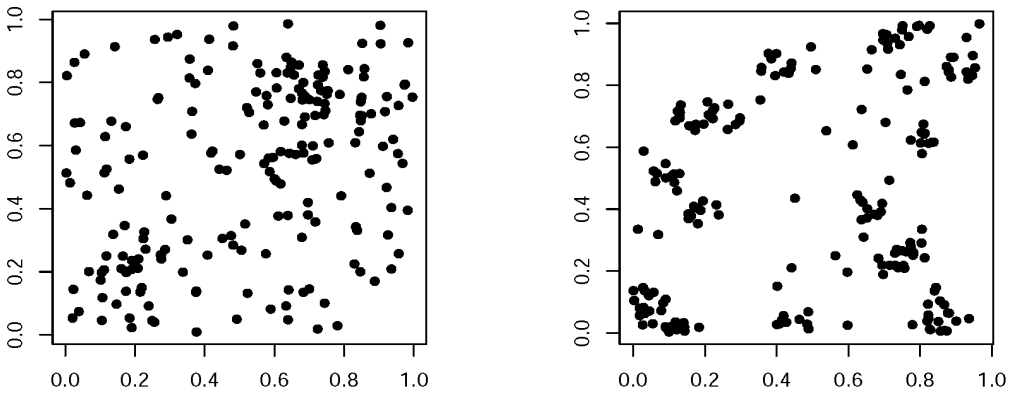

Figure 1: Realization of the CON model (left) and the CLU model (right), $n = 200$

conditionally on $X_1 \in [0,1]^2$. Here, $\varepsilon_1 = 0.135$, $\varepsilon_2 = 0.24$, $\sigma_1 = 0.09$, $\sigma_2 = 0.12$, $c_1 = (0.25, 0.25)$, $c_2 = (0.7, 0.7)$, $\mathcal{U}[0,1]^2$ is the uniform distribution over $[0,1]^2$, and $\mathcal{N}_2(c_j, \sigma_j^2)$ stands for the bivariate normal distribution with expectation vector c_j and covariance matrix $\sigma_j^2 I_2$, where I_2 is the identity matrix of order 2. In other words, this model produces a uniform background noise and two radially symmetric point sources of data, centered at the points c_1 and c_2. The additional specification "conditionally on $X_1 \in [0,1]^2$" means that a realization was discarded whenever the generated point did not fall into the unit square.

The clustering alternative CLU (say) considers an alternative to $\mathcal{H}_0$ in the non iid case, using a two step-technique. In a first step, one simulates $n_1 = 10$ iid random points with the uniform distribution $\mathcal{U}[0,1]^2$, which are then discarded but play the role of centers of clusters. In a second step, one generates, independently of each other, for each of those n_1 centers $n_2 = n/n_1$ points that are uniformly distributed in a disc of radius 0.05, the midpoint being the center. Similar to the CON alternative, each point was discarded if it fell outside $[0,1]^2$, and the point was simulated according to $\mathrm{U}[0,1]^2$ to describe a small uniform noise effect. Figure 1 shows a realization of the CON (left) and the CLU (right) model.

Table 2 shows the percentages (out of 10,000 replications) of rejections of $\mathcal{H}_0$ of the distance to boundary test and the maximal spacing test, rounded to the nearest integer. Obviously, the latter test is sensitive to a cluster alternative, but much inferior to the distance to boundary test against the contamination alternative.

Table 3 exhibits the corresponding percentages of the test based on $T_{n,J}^{(\alpha)}$. An asterisk denotes power 100%. As was to be expected, rejection rates depend crucially on the power α and the number of neighbors J taken into account. In each row, the maximum rejection rates have been highlighted using boldface ciphers. The beginning of a sequence of asterisks has also been emphasized, thus indicating the smallest value of J for which the maximum power is attained. A comparison with Table 2 shows the choice $\alpha = 0.5$ yields a very strong test against cluster alternatives, even for $J = 1$. Likewise, taking $\alpha = 0.5$ and any $J \leq 25$, the test based on $T_{n,J}^{(\alpha)}$ outperforms both DB and MS.

Table 2: Empirical rejection rates of DB and MS, unit square

Alt.	n	DB	MS	Alt.	n	DB	MS
CON	50	31	6	CLU	50	44	67
	100	58	14		100	44	85
	200	89	24		200	44	94

Table 3: Empirical rejection rates of the test based on $T_{n,J}^{(\alpha)}$, unit square

Alt.	α	$n\backslash J$	1	2	3	4	5	6	7	8	9	10	15	20	25
CON	0.5	50	14	22	29	36	43	48	53	56	59	61	**66**	65	60
		100	19	27	36	45	53	60	66	71	76	79	90	93	**94**
		200	25	38	50	60	68	75	81	86	89	91	98	99	∗
CLU		50	∗	*	*	*	*	*	*	*	*	*	*	*	*
		100	∗	*	*	*	*	*	*	*	*	*	*	*	*
		200	∗	*	*	*	*	*	*	*	*	*	*	*	*
CON	2	50	**11**	**11**	9	6	4	3	2	2	1	1	0	0	1
		100	21	27	**29**	**29**	26	23	19	14	10	7	1	0	0
		200	33	50	59	65	68	**70**	**70**	69	68	66	44	14	1
CLU		50	39	**40**	36	29	22	16	12	8	6	5	9	12	13
		100	**54**	52	43	33	24	17	12	8	5	4	6	10	13
		200	**64**	59	46	32	22	14	9	5	3	2	3	6	8
CON	5	50	13	18	19	**20**	**20**	19	17	15	14	12	10	11	13
		100	23	34	44	51	56	59	61	**63**	**63**	**63**	53	35	20
		200	36	58	74	83	89	92	94	96	97	97	**98**	98	98
CLU		50	54	63	**65**	**65**	63	61	59	57	55	61	64	61	56
		100	78	87	**88**	86	85	83	80	77	74	76	83	81	81
		200	93	**98**	97	97	96	95	93	91	89	89	94	92	92

3.2. The circle $\mathcal{S}^1$

A good overview of tests for uniformity on the circle is presented in the monograph [20]. We considered the following classical procedures.

(i) The modified Rayleigh test, suggested in [24] and denoted by *Ra* in what follows, is based on the statistic

$$Ra_n = \left(1 - \frac{1}{2n} + \frac{T_n}{8n}\right) T_n.$$

Here, $T_n = 2n\,|\overline{X}_n|^2$, and $\overline{X}_n = n^{-1}\sum_{j=1}^n X_j$ is the sample mean vector. Under $\mathcal{H}_0$, the limit distribution of Ra_n as $n \to \infty$ is the χ^2_3 distribution.

(ii) Kuiper's test (see [26]), denoted by *Ku*, uses a transformation of $X_1, \ldots, X_n$ to normed radial data $U_1, \ldots, U_n$, as described in [20], p. 153. Writing $0 \le U_{(1)} \le \cdots \le U_{(n)} \le 1$ for the order statistics of $U_1, \ldots, U_n$, Kuiper's test is a Kolmogorov–Smirnov type test using the statistic

$$Ku_n = \sqrt{n}\left\{\max_{1\le j\le n}\left(U_{(j)} - \frac{j-1}{n}\right) + \max_{1\le j\le n}\left(\frac{j}{n} - U_{(j)}\right)\right\};$$

see [20], p. 153.

(iii) Using the same radial data transformation as in (ii), Watson's test (see [43]), denoted by *Wa*, employs the statistic (see [20], p. 156)

$$Wa_n = \sum_{j=1}^n \left(U_{(j)} - \frac{2j-1}{2n} - \frac{1}{n}\sum_{\ell=1}^n U_\ell + \frac{1}{2}\right)^2 + \frac{1}{12n}.$$

The implementation and critical values of the Kuiper (ii) and the Watson (iii) test were taken from the R-package Directional, as provided by [40]. As alternative distributions on the circle we considered the von Mises–Fisher (MF) and the Bimodal von Mises–Fisher (BMF) distributions, see [20], Section 2.3 and [31], Section 9.3. Note that a unit random vector has the $(d-1)$-dimensional von Mises–Fisher distribution if its probability density function with respect to the uniform distribution is given, for all $|x| = 1$, by

$$f_{\mu,\kappa}(x) = (\kappa/2)^{d/2-1} \frac{1}{\Gamma(d/2)\, I_{d/2-1}(\kappa)} \exp(\kappa \mu^\top x). \tag{1}$$

Here, $\kappa > 0$ is a concentration parameter, the unit vector μ denotes the mean direction, I_ν is the modified Bessel function of the first kind and order ν, and the prime stands for tranpose. For the simulations in Tables 4 and 5 we chose $\mu = (1,0)^\top$ and $\kappa = 0.5$. The Bimodal von Mises–Fisher distribution is a mixture of a von Mises–Fisher distribution with $\mu = (1,0)^\top$ and $\mu = (-1,0)^\top$ with the same concentration parameter $\kappa = 1$.

A comparison of Table 4 and Table 5 shows that, among the values of α taken into account, the choice $\alpha = 5$ and $J = 10$ for $n = 50$, while $J = 25$ for $n \in \{100, 200\}$, yields the highest power of the new tests against the von Mises–Fisher distribution. This power is comparable with that of Ra, Ku and Wa. Against the bimodal von Mises–Fisher distribution, the choice $\alpha = 0.5$ and $J = 20$ results in a test that outperforms Ku and Wa for $n = 50$ and is at least as powerful as these tests if $n = 100$ or $n = 200$. Against this alternative, the Rayleigh test is not competitive.

3.3. Sphere $\mathcal{S}^2$

We now treat the case of testing for uniformity on a sphere in $\mathbb{R}^3$, for which many tests have been proposed. A good overview, also for the corresponding testing problems in higher dimensions, is given in [14, 31]. We considered the following procedures.

(i) The Rayleigh test (see [24]), denoted by $\widetilde{Ra}$, rejects the hypothesis of uniformity for large values of

$$\widetilde{Ra}_n = \left(1 - \frac{1}{2n} + \frac{T_n}{16n}\right) T_n,$$

where $T_n = 2n\,|\overline{X}_n|^2$. Under $\mathcal{H}_0$, the limit distribution of $\widetilde{Ra}_n$ as $n \to \infty$ is χ^2_3.

(ii) The data-driven Sobolev test for uniformity applied to the sphere, here called the Jupp test and denoted by JT (see [23]), computes

$$B_n(k) = S_n(k) - k(k+2)\,\ln n,$$

where

$$S_n(k) = \frac{2k+1}{n} \sum_{j,\ell=1}^{n} P_k(X_j' X_\ell),$$

and P_k is the Legendre polynomial of order k. The test statistic is then $JT_n = S_n(\widehat{k})$, where

$$\widehat{k} = \widehat{k}(n) = \inf\left\{k \in \mathbb{N} : B_n(k) = \sup_{m \in \mathbb{N}} B_n(m)\right\}. \tag{2}$$

As suggested in [23], p. 1250, a suitable approximation of the supremum in (2) can be done by considering $\sup_{1 \le m \le 5} B_n(m)$ instead. Critical values may be obtained from the χ^2_3 distribution, since $JT_n \rightsquigarrow \chi^2_3$ as $n \to \infty$ under the hypothesis $\mathcal{H}_0$ of uniformity.

Table 4: Empirical rejection rates of the tests based on Ra, Ku and Wa, circle

Alt.	n	Ra	Ku	Wa	Alt.	n	Ra	Ku	Wa
	50	58	53	58		50	6	63	61
MF	100	88	84	88	BMF	100	6	97	99
	200	*	99	*		200	6	*	*

Table 5: Empirical rejection rates of the test based on $T_{n,J}^{(\alpha)}$, circle

Alt.	α	$n\backslash J$	1	2	3	4	5	6	7	8	9	10	15	20	25
		50	8	9	10	12	13	15	16	18	19	21	32	42	**50**
MF		100	10	12	13	15	17	19	20	21	23	25	34	43	**53**
	0.5	200	12	15	18	20	23	25	27	30	31	33	43	54	**63**
		50	28	44	56	67	76	82	87	90	93	95	**98**	**98**	97
BMF		100	37	56	71	80	87	92	95	97	98	99	✱	*	*
		200	55	77	88	94	97	98	99	✱	*	*	*	*	*
		50	17	24	29	32	34	**36**	**36**	35	33	31	11	1	0
MF		100	22	33	42	48	53	57	60	63	64	66	**67**	64	50
	2	200	30	46	58	66	72	77	80	83	85	87	92	94	**95**
		50	36	**41**	37	28	16	8	3	1	0	0	0	0	0
BMF		100	64	79	**83**	**83**	81	77	71	63	53	42	3	0	0
		200	86	96	**99**	**99**	**99**	**99**	**99**	**99**	**99**	**99**	94	73	28
		50	19	27	33	37	41	44	46	48	50	**51**	50	39	15
MF		100	24	36	46	52	59	63	67	70	72	75	82	85	**86**
	5	200	32	50	62	71	78	83	86	89	91	92	96	97	**98**
		50	48	61	**66**	**66**	61	52	41	28	16	8	2	6	18
BMF.		100	75	91	96	98	**99**	**99**	**99**	**99**	**99**	98	92	51	4
		200	92	99	✱	*	*	*	*	*	*	*	*	*	*

Alt.	n	$\widetilde{Ra}$	JT	Alt.	n	$\widetilde{Ra}$	JT
	50	36	36		50	10	99
MF	100	66	66	Kent	100	13	*
	200	94	94		200	22	*

Table 6: Empirical rejection rates of $\widetilde{Ra}$ and JT, sphere.

As alternatives we considered the von Mises–Fisher distribution as in (1) with concentration parameter $\kappa = 0.5$ and mean direction to $\mu = (1, 0, 0)^\top$. A second alternative is the Kent distribution, see [31], p. 176, with density given, for $|x| = 1$, by

$$f_{\mu,\kappa,\beta}(x) = \frac{1}{c(\beta,\kappa)} \exp\{\kappa\mu^\top x + \beta x^\top(\tau_1\tau_1^\top - \tau_2\tau_2^\top)x\}.$$

Here, $c(\beta, \kappa)$ is a normalizing constant, and τ_1, τ_2 and μ are mutually orthogonal vectors. The references to the Kent distribution in Tables 6 and 7 use $\kappa = 0.25, \mu = (1, 0, 0)^\top$ and $\beta = 2$.

The results of the simulation study are given in Tables 6 and 7. The presented procedure is outperformed by both procedures for the von Mises–Fisher distribution, although Table 7 indicates that for $J > 25$ and $\alpha = 0.5$ (or $\alpha = 5$) better performances are possible. On the other hand, the new procedure is competitive to the other tests and outperforms the modified Rayleigh test for the Kent distribution if $\alpha = 0.5$ and $J \geq 7$. Notice that the Rayleigh and Jupp test have (nearly) the same power against the von Mises–Fisher distribution.

4. Gamma-ray burst data analysis

Gamma-ray bursts (GRB) are the brightest and most violent known events in the universe. They originate from extremely energetic explosions in distant galaxies. In August 2017, a GRB and a gravitational-wave (GW) event, both originating from the same collision of two neutron stars, were for the first time detected simultaneously; see [8]. The gravitational-wave (GW) event GW 170817 was observed by the Advanced LIGO and Virgo detectors, and the GRB 170817A was observed independently by the Fermi Gamma-ray Burst Monitor, and the Anti-Coincidence Shield for the Spectrometer for the International Gamma-Ray Astrophysics Laboratory. The probability of the near-simultaneous temporal and spatial observation of GRB 170817A and GW 170817 occurring by chance is 5.0×10^{-8}; see [30]. We investigate two data sets of $n = 44$ and $n = 1163$ GRB events given in the galactic coordinate system, which is a

Table 7: Empirical rejection rates of the test based on $T^{(\alpha)}_{n,J}$, sphere

Alt.	α	$n\backslash J$	1	2	3	4	5	6	7	8	9	10	15	20	25
MF.	0.5	50	7	8	9	10	11	13	14	15	16	17	23	27	**31**
		100	8	9	10	11	12	13	15	16	17	18	25	32	**39**
		200	9	11	13	14	16	17	19	20	22	24	31	39	**47**
Kent		50	66	87	95	98	99	99	✱	*	*	*	*	*	*
		100	83	97	99	✱	*	*	*	*	*	*	*	*	*
		200	96	✱	*	*	*	*	*	*	*	*	*	*	*
MF.	2	50	10	12	**13**	**13**	**13**	**13**	12	11	10	8	3	2	1
		100	13	17	19	22	24	25	**26**	**26**	25	25	20	14	6
		200	17	24	30	36	40	43	46	48	50	51	**55**	54	50
Kent		50	**14**	9	4	1	0	0	0	0	0	0	0	0	0
		100	41	**42**	36	27	18	10	5	2	0	0	0	0	0
		200	77	87	**88**	86	83	78	71	62	51	41	4	0	0
MF.	5	50	12	15	18	21	23	25	26	27	**27**	27	26	20	11
		100	15	22	28	33	38	42	44	48	50	51	57	**59**	**59**
		200	19	31	41	50	57	63	67	71	74	76	84	87	**90**
Kent		50	29	**35**	34	30	25	18	12	7	4	2	2	8	29
		100	63	79	86	**88**	**88**	87	86	83	80	76	40	1	0
		200	91	99	✱	*	*	*	*	*	*	*	*	97	96

celestial coordinate system giving points on the sphere $\mathcal{S}^2$; see Figure 2. The first data set ($n = 44$) are GRB observed by the Fermi telescope, the second ($n = 1163$) by the Swift telescope; see [9, 27]. Both data sets were obtained from the NASA website

`https://swift.gsfc.nasa.gov/archive/grb_table/`.

The question of interest is to test the null-hypothesis of intrinsic randomness, which is in astronomy typically considered to be distributed according to a uniform distribution; see [32].

As noticed in the beginning of Section 3, we need an estimated value of $\sigma^2(f_0, f_0)$ to apply the results of Theorem 2 in order to obtain critical values as well as approximate p-values. In the spirit of Monte Carlo Tests, we simulated $\sigma^2(f_0, f_0)$ under $\mathcal{H}_0$ for the specific sample sizes with 10,000 replications for $\alpha \in \{0.5, 2, 5\}$ and $J \in \{1, \ldots, 5\}$, see Table 8. Table 9 shows the approximate p-values of the tests. Obviously the Rayleigh test $\widetilde{Ra}$ does not reject the null hypothesis at any significance level, while most of the other tests do at a level of 0.05 (except $T^{(\alpha)}_{n,1}$ for $\alpha \in \{2, 5\}$ and $T^{(2)}_{n,5}$ for the Fermi data set). An explanation for the rejection of the hypothesis might be the accuracy errors in the detection of the point events (which are also listed in the Swift data set).

5. Conclusions and open problems

We have introduced a new, flexible class of universally consistent goodness-of-fit tests based on sums of powers of volumes of weighted kth nearest neighbor balls. Under fixed alternatives, scaled versions of the test statistics converge

Table 8: Estimated $\sigma^2(f_0, f_0)$ under $\mathcal{H}_0$

n	$\alpha\backslash J$	1	2	3	4	5
44	0.5	0.217	0.735	1.563	2.683	4.088
	2	13.2	90.0	343	969	2253
	5	142e04	226e05	173e06	930e06	416e07
1163	0.5	0.214	0.734	1.552	2.686	4.130
	2	15.5	107.7	410	1148	2698
	5	303e04	534e05	444e06	255e07	117e08

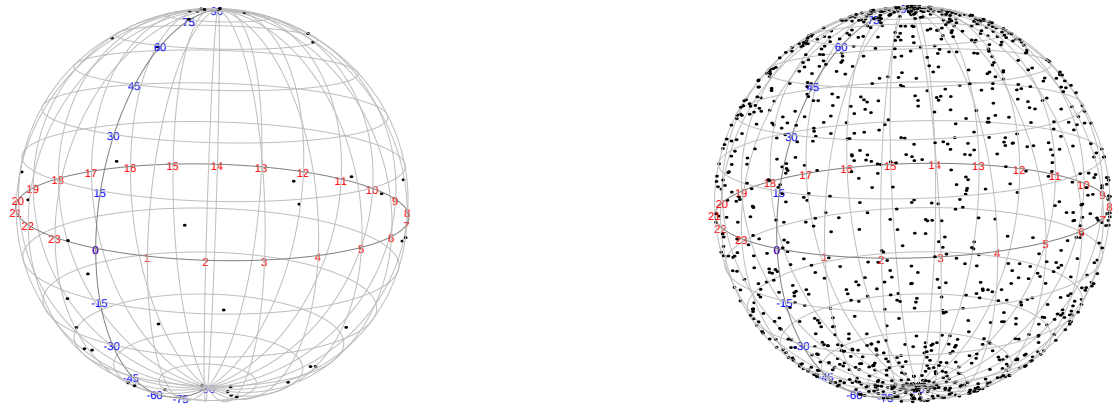

Figure 2: Gamma-ray burst observations in the galactic coordinate system, Fermi (left, $n = 44$) and Swift (right, $n = 1163$)

n	α	$T_{n,1}^{(\alpha)}$	$T_{n,2}^{(\alpha)}$	$T_{n,3}^{(\alpha)}$	$T_{n,4}^{(\alpha)}$	$T_{n,5}^{(\alpha)}$	$\widetilde{Ra}$	JT
44	0.5	8.56e-03	3.00e-02	1.66e-02	3.71e-03	5.72e-04	0.75	4.40e-05
	2	3.14e-01	1.30e-02	2.31e-03	8.67e-03	5.11e-02		
	5	1.71e-01	4.01e-02	5.38e-04	5.95e-04	1.70e-03		
1163	0.5	3.19e-06	9.94e-07	3.38e-07	7.75e-09	2.64e-11	0.27	0
	2	4.37e-02	7.24e-05	1.34e-10	9.27e-13	1.26e-13		
	5	5.10e-04	9.08e-08	0	0	0		

Table 9: Approximate p-values for the data sets

to the α-entropy between probability distributions. The approach is fairly general, since it covers both goodness-of-fit testing for distributions with a compact, 'full-dimensional' support in $\mathbb{R}^d$, but also on lower-dimensional manifolds embedded in $\mathbb{R}^d$. Our approach requires J, the maximum number of neighbors taken into account, to remain fixed as $n \to \infty$. It would be desirable to obtain limit theorems also for the case that $J = J(n)$ tends to infinity with the sample size n. Another problem is to generalize the theory to cover testing for a parametric family $\{f(\cdot;\vartheta) : \vartheta \in \Theta\}$ of densities. This could be done by substituting $f(X_i;\widehat{\vartheta}_n)$ for the weight $f_0(X_i)$, where $\widehat{\vartheta}_n$ is a suitable estimator of ϑ, based on $X_1, \ldots, X_n$. In view of the fluctuating performance in the simulation study, it is desirable to find an optimal (data dependent) choice of the parameters J and α, which remains an open problem.

A further challenge is the behavior of $T_{n,J}^{(\alpha)}$ with respect to contiguous alternatives. In this respect, Jammalamadaka and Zhou [21] considered the statistic

$$T_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[h(n f_0(X_i) v_d |X_i - X^{(1)}|^d) - \mathrm{E}_0 \left\{ h(n f_0(X_i) v_d |X_i - X^{(1)}|^d) \right\} \right], \tag{1}$$

where $\mathbb{E}_0$ denotes expectation under $\mathcal{H}_0$ and $h : [0,\infty) \to \mathbb{R}$ is some measurable function. Suppose $\{f > 0\} = \{x \in \mathbb{R}^d : f(x) > 0\}$ is open in $\mathbb{R}^d$ and f is uniformly bounded and twice continuously differentiable on $\{f > 0\}$. Jammalamadaka and Zhou [21] considered the limiting distribution of T_n^* under the sequence of alternatives

$$\mathcal{H}_{1,n} : X_i \sim f_0(x) + n^{-1/4} \ell(x),$$

where $\int_{\mathbb{R}^d} \ell(x)\,\mathrm{d}x = 0$. If ℓ is supported in a compact subset of $\{f > 0\}$ and is twice continuously differentiable on

$\{f > 0\}$, h is of bounded variation on $[0, \infty)$ and if $d < 8$, then Theorem 5.5 of [21] says that, under the sequence $H_{1,n}$,

$$T_n^* \rightsquigarrow \mathcal{N}[\mu(h), \sigma^2(h)],$$

where $\mu(h)$ and $\sigma^2(h)$ are given in Theorem 4 of [21]. Notice that the sum figuring in (1), apart from the centering, yields the statistic $T_{n,1}^{(\alpha)}$ if we put $h(t) = t^\alpha$. This function, however, is not of bounded variation on $[0, \infty)$ which shows that the result above is not applicable. Nevertheless, one may conjecture that the statistic $T_{n,J}^{(\alpha)}$ has positive asymptotic power against alternatives that approach f_0 at the rate $n^{-1/4}$.

Acknowledgements

The authors would like to thank Michael A. Klatt for his expertise in gamma-ray astronomy and for indicating the two data sets as well as an Associate Editor and an anonymous referee for their careful reading of the manuscript and for many helpful suggestions.

References

[1] Yu. Baryshnikov, M. Penrose, J.E. Yukich, Gaussian limits for generalized spacings, Ann. Appl. Probab. 19 (2009) 158–185.

[2] J.R. Berrendero, A. Cuevas, B. Pateiro-López, A multivariate uniformity test for the case of unknown support, Stat. Comput. 22 (2012) 259–271.

[3] J.R. Berrendero, A. Cuevas, B. Pateiro-López, Testing uniformity for the case of a planar unknown support, Canad. J. Statist. 40 (2012) 378–395.

[4] J.R. Berrendero, A. Cuevas, F. Vázquez-Grande, Testing multivariate uniformity: The distance-to-boundary method, Canad. J. Statist. 34 (2006) 693–707.

[5] R. Bhattacharya, V. Patrangenaru, Statistics on manifolds and landmarks based image analysis: A nonparametric theory with applications, J. Stat. Plann. Inf. 145 (2014) 1–22.

[6] G. Biau, L. Devroye, V. Dujmovic, A. Krzyżak, An affine invariant k-nearest neighbor regression estimate, J. Multivariate Anal. 112 (2012) 24–34.

[7] P.J. Bickel, L. Breiman, Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test, Ann. Probab. 11 (1983) 185–214.

[8] M. Coleman Miller, Gravitational waves: A golden binary, Nature 551 (2017) 36–37.

[9] V. Connaughton, M.S. Briggs, A. Goldstein, C.A. Meegan, W.S. Paciesas, R.D. Preece, C.A. Wilson-Hodge, M.H. Gibby, J. Greiner, D. Gruber, P. Jenke, R.M. Kippen, V. Pelassa, S. Xiong, H.-F. Yu, P.N. Bhat, J.M. Burgess, D. Byrne, G. Fitzpatrick, S. Foley, M.M. Giles, S. Guiriec, A.J. van der Horst, A. von Kienlin, S. McBreen, S. McGlynn, D. Tierney, B.-B. Zhang, Localization of gamma-ray bursts using the Fermi gamma-ray burst monitor, The Astrophysical Journal Supplement Series 216 (2015) 1–32.

[10] H. Dette, N. Henze, Some peculiar boundary phenomena for extremes of rth nearest neighbor links, Statist. Probab. Lett. 10 (1990) 381–390.

[11] L. Devroye, T. Wagner, The strong uniform consistency of nearest neighbor density estimates, Ann. Statist. 5 (1977) 536–540.

[12] P. Diaconis, S. Holmes, M. Shahshahani, Sampling from a manifold, In: Advances in Modern Statistical Theory and Applications. A Festschrift in Honor of Morris L. Eaton, IMS, Beachwood, OH, 2013, pp. 102–125.

[13] B. Ebner, N. Henze, J.E. Yukich, Multivariate goodness-of-fit on flat and curved spaces via nearest neighbor distances, ArXiv e-prints, 1612.06601.

[14] N.I. Fisher, T. Lewis, B.J.J. Embleton, Statistical Analysis of Spherical Data, Cambridge University Press, 1987.

[15] S. Gadat, T. Klein, C. Marteau, Classification in general finite-dimensional spaces with the k-nearest neighbor rule, Ann. Statist. 44 (2016) 982–1009.

[16] E.M. Giné, Invariant tests for uniformity on compact Riemannian manifolds based on Sobolev norms, Ann. Statist. 3 (1975) 1243–1266.

[17] N. Henze, The limit distribution of maxima of "weighted" rth nearest neighbour distances, J. Appl. Probab. 19 (1982) 344–354.

[18] N. Henze, A multivariate two-sample test based on the number of nearest neighbor type coincidences, Ann. Statist. 16 (1988) 772–783.

[19] N. Henze, On the consistency of the spacings test for multivariate uniformity, arXiv:1708.09211, 2017.

[20] S.R. Jammalamadaka, A. SenGupta, Topics in Circular Statistics, World Scientific Publishing, 2001.

[21] S.R. Jammalamadaka, S. Zhou, Goodness of fit in multidimensions based on nearest neighbor distances, J. Nonparam. Statist. 2 (1993) 271–284.

[22] S. Janson, Maximal spacings in several dimensions, Ann. Probab. 15 (1987) 274–280.

[23] P.E. Jupp, Data-driven Sobolev tests of uniformity on compact Riemannian manifolds, Ann. Statist. 36 (2008) 1246–1260.

[24] P.E. Jupp, Modifications of the Rayleigh and Bingham tests for uniformity of directions, J. Multivariate Anal. 77 (2001) 1–20.

[25] B.F. Kimball, Some basic theorems for developing tests of fit for the case of the non-parametric distribution function, I, Ann. Math. Statist. 18 (1947) 540–548.

[26] N.H. Kuiper, Tests concerning random points on a circle, Ned. Akad. Wet. Proc. 63 (1960) 38–47.

[27] L. Lenkic, P. Tzanavaris, S.C. Gallagher, T.D. Desjardins, L.M. Walker, K.E. Johnson, K. Fedotov, J. Charlton, A.E. Hornschemeier, P.R. Durrell, C. Gronwall, VizieR Online Data Catalog: Compact Group Galaxies UV and IR SFR (Lenkic+, 2016), VizieR Online Data Catalog, 745, 2017.

[28] R. Lachièze-Rey, M. Schulte, J.E. Yukich, Normal approximation for sums of stabilizing functionals, arXiv: 1702.00726, 2017.
[29] C. Ley, T. Verdebout, Modern Directional Statistics, CRC Press, London, 2017.
[30] LIGO Scientific Collaboration and Virgo Collaboration, Fermi Gamma-ray Burst Monitor, INTEGRAL, Gravitational Waves and Gamma-Rays from a Binary Neutron Star Merger: GW170817 and GRB 170817A, The Astrophysical Journal Letters 848 (2017) L13, 1–27.
[31] K.V. Mardia, P.E. Jupp, Directional Statistics, Wiley, New York, 2000.
[32] A. Mészáros , Z. Bagoly, L.G. Balázs, I. Horváth, R. Vavrek, Probing the isotropy in the sky distribution of the gamma-ray bursts, In: E. Costa, F. Frontera, J. Hjorth (Eds.) Gamma-Ray Bursts in the Afterglow Era. ESO ASTROPHYSICS SYMPOSIA (European Southern Observatory). Springer, New York, 2003.
[33] P.K. Mondal, M. Biswas, A.K. Ghosh, On high dimensional two-sample tests based on nearest neighbors, J. Multivariate Anal. 141 (2015) 168–178.
[34] M.D. Penrose, J.E. Yukich, Central limit theorems for some graphs in computational geometry, Ann. Appl. Probab. 11 (2001) 1005–1041.
[35] M.D. Penrose, J.E. Yukich, Weak laws of large numbers in geometric probability, Ann. Appl. Probab. 13 (2003) 277–303.
[36] M.D. Penrose, J.E. Yukich, Limit theory for point processes in manifolds, Ann. Appl. Probab. 23 (2013) 2161–2211.
[37] A. Petrie, Th.R. Willemain, An empirical study of tests for uniformity in multidimensional data, Comput. Statist. Data Anal. 64 (2013) 253–268.
[38] R Core Team, R: A language and environment for statistical computing, Statistical Computing, Vienna, Austria, 2016.
[39] M.F. Schilling, Multivariate two-sample tests based on nearest neighbors, J. Amer. Statist. Assoc. 81 (1986) 799–806.
[40] M. Tsagris, G. Athineou, A. Sajib, Directional: Directional Statistics, R package version 2.1, 2016.
[41] I. Vajda, On the amount of information contained in a sequence of independent observations, Kybernetika 6 (1970) 306–323.
[42] A.R. Wade, Explicit laws of large numbers for random nearest neighbor type graphs, Adv. Appl. Probab. 39 (2007) 326–342.
[43] G.S. Watson, Goodness-of-fit tests on the circle, Biometrika 48 (1961) 109–114.
[44] L. Weiss, The asymptotic power of certain tests of fit based on sample spacings, Ann. Math. Statist. 28 (1957) 783–786.