Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v13n2p358

The interpoint depth for directional data
By Pandolfo

Published: 14 October 2020

Electronic Journal of Applied Statistical Analysis
Vol. 13, Issue 02, October 2020, 358-374
DOI: 10.1285/i20705948v13n2p358

The interpoint depth for directional data

Giuseppe Pandolfo*

University of Naples Federico II, Department of Industrial Engineering
P.le Tecchio, 80, 80125 Napoli (Italy)

Published: 14 October 2020

The notion of interpoint depth is applied to spherical spaces by using an appropriate angular distance function for data lying on the surface of the unit hypersphere. The traditional multivariate methods, indeed, are not suitable for the analysis of directional data and this holds true also for distance measures and related depth-based methods. The interpoint depth for directional data possesses some nice properties and can be used for high dimensional data analysis. This notion of depth is particularly useful to investigate local features of distribution, such as multimodality, and can be exploited to deal with many statistical problems. The behavior of the proposed depth is investigated by means of simulated data. In addition, three interesting applications are presented.

keywords: Data depth, Spherical distance, Spherical variables, Uniformity.

1 Introduction

Directional data arise when observations are recorded as directions, that is unit vectors, on the surface of the unit $(q-1)$-dimensional hypersphere $S^{q-1} := \{x \in \mathbb{R}^q : x'x = 1\}$, for $q \geq 2$. Such data can be found in many scientific fields such as geology, meteorology and biology, just to name a few. Most applications are for $q = 2$ (circular data) or $q = 3$ (spherical data), but applications in higher dimensions can also be found, e.g. in gene-expression analysis (see Banerjee et al., 2005), text mining (see Buchta et al., 2012), or image processing as well as pattern recognition (see Wilson et al., 2014). For a comprehensive overview on such data see Mardia and Jupp (2009) and Ley and Verdebout (2017).

*Corresponding author: giuseppe.pandolfo@unina.it

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

Statistical methods, to produce sensible results, must take into account the peculiar features of directional data, i.e. the lack of a well-defined reference direction, the arbitrariness of the sense of rotation and, as it occurs for multivariate data in $\mathbb{R}^q$, the absence of a natural ordering. In this regard, the approach proposed by Liu and Singh (1992) provides a coherent depth-based method to non-parametrically analyze directional data.

The concept of statistical data depth plays a very important role in non-parametric statistics because it leads to a natural center-outward ordering of data (in $\mathbb{R}^q$ and S^{q-1} as well) and thus opens the possibility of using non-parametric methods in high dimensions with no need of distributional assumptions.

The field of applications of depth functions is vast and still growing. Some are non-parametric location and scatter estimation, outlier detection, classification and clustering. However, data depth are well suited for unimodal distribution and consequently unable to capture any local feature such as multimodality. For this reason, Lok and Lee (2011) introduced and investigated the notion of interpoint depth for multivariate and functional data. Although this cannot be strictly considered a notion statistical depth function in $\mathbb{R}^q$ as defined by Zuo and Serfling (2000), it can be considered a useful alternative when the purpose is to get insights into the local features of a distribution (especially for multimodality). However, it is worth underlying that depth functions do not do the same job as density functions. Indeed, it is advisable to use them together, rather than to replace them by a depth function which is sensitive to multimodality. For instance, in univariate case, the density function might be bimodal and the median might have low or even zero density, while the depth function is always maximized at the symmetry center.

Despite some literature exists on the exploitation of depths based on interpoint distance (see e.g. Lok and Lee, 2011, Liu and Modarres, 2011, and Dong and Lee, 2014), to the best of the author's knowledge, the application to hyperspherical data received no attention even if it can be really meaningful in such framework as well.

Hence, the definition of interpoint depth of Lok and Lee (2011) is applied to directional data by using a suitable angular distance function in Section 2. Some properties are also established. The remainder of the paper is organized as follows. The empirical behavior of the proposed depth is investigated in Section 3. Section 4 presents some possible applications of such depth for the analysis of directional data. Section 5 reports some final remarks.

2 Spherical interpoint distance depth

In this section the concept of interpoint depth of Lok and Lee (2011) is applied to directional data. This notion of depth appears to be really useful to capture multimodality in $\mathbb{R}^q$ and, to the best of the author's knowledge, its directional version has not been investigated yet in the literature.

Let $X_1, X_2, \ldots, X_n$ be a spherical random sample independently distributed as a distribution F on S^{q-1}. The main idea is considering a set of $m < n$ data points scattered in a "neighborhood" of a given point $x \in S^{q-1}$, that is the open set of data points that

are closer to x according to an appropriate measurable, non-negative and bounded angular distance function $d_{sph}(\cdot,\cdot)$ (the triangle inequality property is not needed). Then, the distance matrix between data points can be used to infer the local features of the distribution. The mean distance from x to its neighborhood is then defined in the usual way by integration.

Definition 2.1. (Interpoint spherical distance depth) For any point $x \in S^{q-1}$, let $\eta(t|x,F) = \mathbb{P}_F(d_{sph}(x,X) \leq t)$ for a random variable $X \sim F$. The interpoint distance has distribution function $\bar{\eta}(t|F) = \mathbb{P}_F(d_{sph}(X,Y) \leq t) = E_F[\eta(t|X,F)]$ for independent X and Y distributed as F. Then the interpoint spherical distance depth $ID_{d_{sph}}$ is defined as

$$ID_{d_{sph}}(x,F,\delta) := \left[\frac{1+\eta^{-1}(\delta|x,F)}{\bar{\eta}^{-1}(1-\xi|F)}\right]^{-1},$$

where the parameters δ and $\xi \in (0,1)$ are fixed, and $\bar{\eta}^{-1}(1-\xi|F)$ is a normalization factor so that $ID_{d_{sph}} \in [0,1]$ with with η^{-1} denotes the inverse of η and $\bar{\eta}^{-1}$ the inverse of $\bar{\eta}$.

This notion of depth measures how close a point x is to the center(s) of a distribution F on S^{q-1} according to a neighborhood containing probability at least equal to δ. Decreasing δ leads to an increase of the sensitivity of the depth function to local features of the distribution. A reasonable choice is to take a value of δ that is smaller than 0.5 and ξ very close to zero. The factor $\bar{\eta}^{-1}(1-\xi|F)$ is aimed at providing a range for the interpoint depth in $[0,1]$ and at the same time makes it scale-invariant. In principle, it can be any constant value (it does not not affect the ranking of points provided by the depth function).

The empirical version is given by

$$ID_{d_{sph}}(x,F_n,\delta) = \left[\frac{1+\inf\left\{t|n^{-1}\sum_{i=1}^{n} I(d_{sph}(x,X_i) \leq t) \geq \delta\right\}}{\inf\left\{t|\binom{n}{2}^{-1}\sum_{1\leq i_1<i_2\leq n} I(d_{sph}(X_{i_1},X_{i_2}) \leq t) \geq 1-\xi\right\}}\right]^{-1}.$$

One important advantage of such depth is its computational ease since it is not based on geometrical structures such as hemispheres or spherical simplices. Obviously one important issue concerns the choice of the spherical distance to be adopted. The most appropriate and natural choice appears to be the arc distance, $d_{arc}(\cdot,\cdot)$, which is defined as

$$d_{arc}(x,y) = \arccos\left(x'y\right), \tag{1}$$

where x and y are two points on the surface of the $(q-1)$-dimensional sphere S^{q-1}, and $x'y$ is the usual inner product of two points. This is also known as great-circle distance and measures the length of the shortest arc joining two points on the surface of a sphere, and ranges between 0 and π. Using the distance measure defined in (1) makes it possible to define the spherical median $\tilde{\mu}$ (Fisher, 1985) of a random sample $X_1, X_2, \ldots, X_n$ as follows

$\tilde{\mu} = \underset{y \in S^{q-1}}{\arg\min} \sum_{i=1}^{n} \arccos\left(X_i' y\right)$.

Liu and Singh (1992) proposed the arc distance depth of a point $x \in S^{q-1}$ which is defined as π minus the expected arc distance of x to a distribution on S^{q-1}. However, this notion of depth for spherical data is not suitable for analyzing multimodal distributions and it remains constant in the case of bimodal antipodally symmetric distributions, that is when the density is such that $f(-x) = f(x)$ (see Liu and Singh, 1992 for more details). Hence, the adoption of the arc distance measure gives rise to the definition of the interpoint arc distance depth $ID_{d_{arc}}(x, F, \delta)$ which possesses some interesting and useful properties analyzed in the next section.

2.1 Properties of $ID_{d_{arc}}$

The interpoint arc distance depth is *rotational invariant*, that is for a $q \times q$ orthogonal matrix R it holds $ID_{d_{arc}}(Rx, F_R, \delta) = ID_{d_{arc}}(x, F, \delta)$, where F_R denotes the image of F by the transformation $x \mapsto Rx$. This property, which is of great importance in the spherical setting, is inherited by the arc distance function which is rotational invariant by definition. In addition, note that for $\delta = 0.5$ (and $\xi = 1$), the $ID_{d_{arc}}$ deepest point is the directional least median of squares (DLMS), i.e the midpoint of the shortest arc which contains in its interior and boundary at least half of the data.

Assume the distribution F on S^{q-1} admits density f which is bounded, positive and continuously differentiable. Let x_0 be a point on $\in S^{q-1}$ and $\mathcal{B}_{(x_0,r_0)} = \left\{x \in S^{q-1} : d_{sph}(x_0, y) \leq r_0\right\}$ be the geodesic ball centered at x_0 with radius r_0.

Proposition 2.1. *The interpoint arc distance depth has a local minimum or maximum at point x_0. Let which $x_0 \in S^{q-1}$ and $r_0 > 0$ satisfy:*

$$\int_{\mathcal{B}_{(x_0,r_0)}} f(z)\, dz = \delta \quad \text{and} \quad \int_{\mathcal{B}_{(x_0,r_0)}} \nabla f(z)\, dz = 0.$$

according to whether $\int_{\mathcal{B}_{(x_0,r_0)}} \nabla^2 f(z)\, dz$ is negative or positive definite, respectively.

Proposition 2.1 ensures that $ID_{d_{arc}}$ attains its maximum or minimum at point x_0 within the region with probability density equal to δ. The points at extremes of the region have the same density.

Proof. The unit $(q-1)$-dimensional sphere is a subset of the q-dimensional Euclidean space, hence the proof is the same of Proposition 1 in Dong and Lee (2014) and thus omitted. □

Proposition 2.2. *Let F be a (circularly contoured) non uniform rotationally symmetric distribution on S^{q-1} with unique mode at μ_0 that admits density function of the form $C_q f(\kappa x' \mu_0)$, where C_q is a normalizing constant, kappa the dispersion parameter and $f : [-1, 1] \to \mathbb{R}^+$ a monotone strictly decreasing function. Assume that d_{sph} is based on a monotone strictly decreasing function $d : [-1; 1] \to \mathbb{R}^+$. Then $ID_{d_{sph}}$ is maximized at μ_0, from which it decreases monotonically along any ray.*

Proof. From the unimodal rotationally symmetry it holds that for any geodesic path $\nu \to \mu_\nu$ from μ_0 to $\mu_1 = -\mu_0$ (i.e. the antipodal point to μ_0), keeping the parameter κ fixed, $ID_{d_{sph}}$ depends on the angle between x_i' and μ_0 for each $i = 1, \ldots, n$. The monotonicity assumption on f implies that, for any $t \in [-1; 1]$, $P\left[x_i'\mu_0 \geq t\right]$ is monotone decreasing. Then, the proof is the same of Proposition 2 in Dong and Lee (2014) and thus omitted. □

3 Empirical behavior: illustrations

The behavior of the interpoint spherical depth function based on the arc distance function is investigated by means of simulated data. For this purpose data were generated from the von Mises-Fisher (vMF) distribution, that is the directional analog of the Gaussian distribution, whose density function is given by

$$f(x, \mu, \kappa) = \frac{\kappa^{q/2-1}}{2\pi^{q/2} I_{q/2-1}(\kappa)} \exp\left\{\kappa \cos\left(x'\mu\right)\right\},$$

where $I_v(\kappa)$ is the modified Bessel function of first kind and order v. The density is parametrized by the mean direction μ and the concentration parameter κ, that measures how strongly data are concentrated around the mean direction μ (larger values indicate stronger concentration of the unit vectors around μ).

For the sake of illustrations only the circular case ($q = 2$) is considered. Specifically, data were simulated under an antipodally symmetric distribution obtained by a mixture of two von Mises-Fisher distributions $F_1 = \frac{1}{2}vMF(\pi/2, 6) + \frac{1}{2}vMF(3/2\pi, 6)$, and a bimodal directional distributions $F_2 = \frac{1}{2}vMF(\pi/4, 6) + \frac{1}{2}vMF(3/4\pi, 6)$. Finally, a distribution which consists of a mixture of two uniform distributions such that $F_3 = \frac{1}{2}U[\pi/3, 2/3\pi] + \frac{1}{2}U[5/6\pi, 7/6\pi]$ is also considered.

Plots of the $ID_{d_{arc}}$ shape under these settings are depicted in Figure 1 along with the density of the underlying distribution, for $\delta \in \{0.15, 0.30\}$ and $\xi = 0.1$. In order to make the visualizations easier to read, some noise was added to avoid depth curves' over-plotting. As one can see, the multimodality of the distributions is clearly reflected by $ID_{d_{arc}}$ for $\delta = 0.15$ in each of the considered settings. In the case of antipodally symmetric and bimodal distributions, it exhibits two centers in correspondence of the two modes and the constancy in the case of uniformity. It must be noted that when δ is equal to 0.30, $ID_{d_{arc}}$ becomes more flat, and peaks are less clearly identified. This underlines the importance of the choice of δ which should be made at hand according to the specific application situation since there is no optimal rule in general.

In addition, the generalized von Mises (GvM) family of Gatto and Jammalamadaka (2007) was also considered. Such extension offers an alternative (and flexible) means of modeling multimodal or asymmetric circular data. The GvM density is given by

$$f(x, \mu_1, \mu_2, \kappa_1, \kappa_2) = \frac{1}{2\pi G_0(\delta, \kappa_1, \kappa_2)} \exp\left\{\kappa_1 \cos(x - \mu_1) + \kappa_2 \cos 2(x - \mu_2)\right\},$$

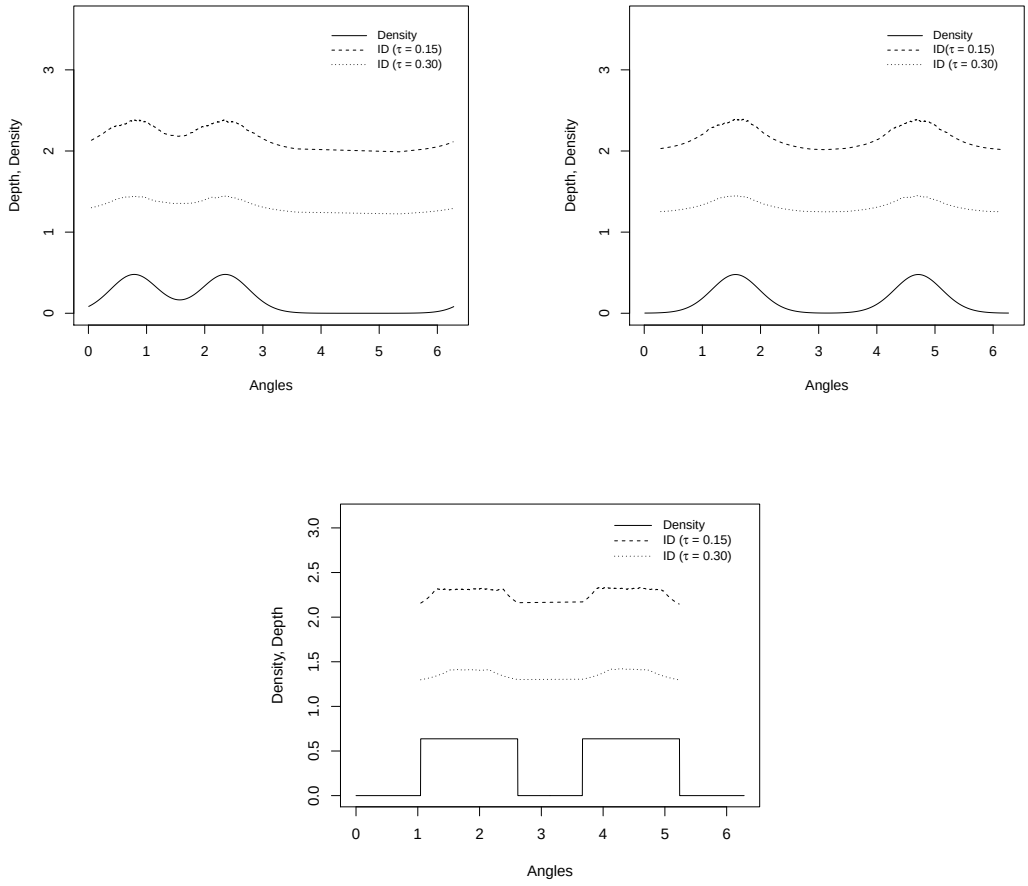

Figure 1. Plot of the interpoint spherical depth functions based on the arc length distance for a bimodal distribution (upper-left panel), an antipodally symmetric distribution (upper-right panel) and a mixture of uniform distributions (lower panel). Calculation was performed for $\delta \in \{0.15, 0.30\}$ and $\xi = 0.1$.

where G_0 denotes the normalizing constant given by

$$G_0\left(\delta, \kappa_1, \kappa_2\right) = \frac{1}{2\pi} \int_0^{2\pi} \exp\left\{\kappa_1 \cos x + \kappa_2 \cos 2\left(x + \delta\right)\right\} dx.$$

Specifically, two cases were considered by setting the parameters as follows:

1. $\mu_1 = 0$, $\mu_2 = 0.5$, $\kappa_1 = 1$, $\kappa_2 = 0.6$,
2. $\mu_1 = 0$, $\mu_2 = 1$, $\kappa_1 = 0.8$, $\kappa_2 = 3$.

The $ID_{d_{arc}}$ under such settings are depicted in Figure 2 along with the density of the underlying distribution, for $\delta \in \{0.15, 0.30\}$ and $\xi = 0.1$. In order to make the visualizations easier to read, some noise was added to avoid depth curves' over-plotting. As one can see, here again the $ID_{d_{arc}}$ is able to reflect the shape of the considered distributions, especially for $\delta = 0.15$.

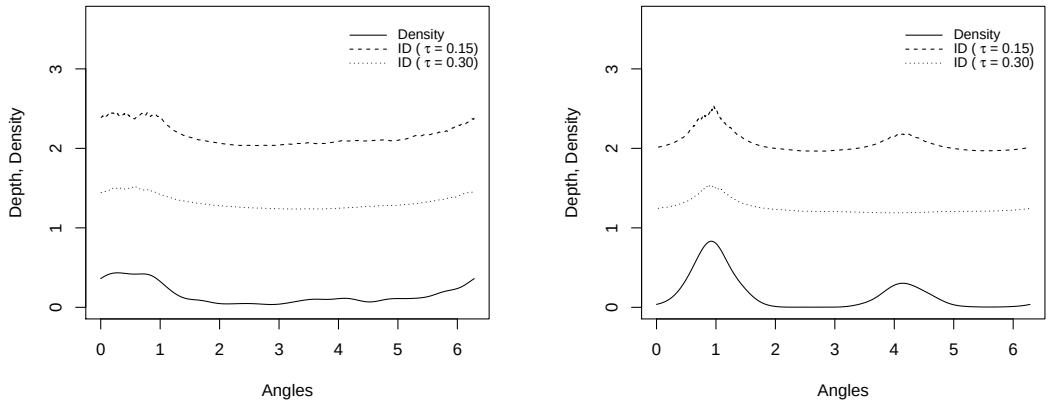

Figure 2. Plot of the interpoint spherical depth functions based on the arc length distance for a generalized von Mises with with parameters $\mu_1 = 0$, $\mu_2 = 0.5$, $\kappa_1 = 1$, $\kappa_2 = 0.6$ (left panel), and a generalized von Mises with parameters $\mu_1 = 0$, $\mu_2 = 1$, $\kappa_1 = 0.8$, $\kappa_2 = 3$ (right panel). Calculation was performed for $\delta \in \{0.15, 0.30\}$ and $\xi = 0.1$.

3.1 Real data example

The following example shows the usefulness of the interpoint depth for directional data in the analysis of real data. The considered data set concerns wind directions in the Atlantic coast of Galicia (NW Spain) already analyzed by Oliveira et al. (2014) and freely downloadable from the Spanish Portuary Authority (`http://www.puertos.es`). Data contains hourly observations of wind direction in winter season (from November to February) from 2003 until 2012. A subset of size of $n = 200$, by taking the observations

with a lag period of 95 hours, was considered here. As one can see in Figure 3, data suggest some degree of multimodality and the ID_{arc} for $\delta = 0.15$ appears to correctly identify the main peaks of the distribution around 1 and some other minor groups between 3 and 2π. Instead, when δ is set to 0.30 and $\xi = 0.1$, the ID_{arc} is more flat, and peaks are less clearly identified. .

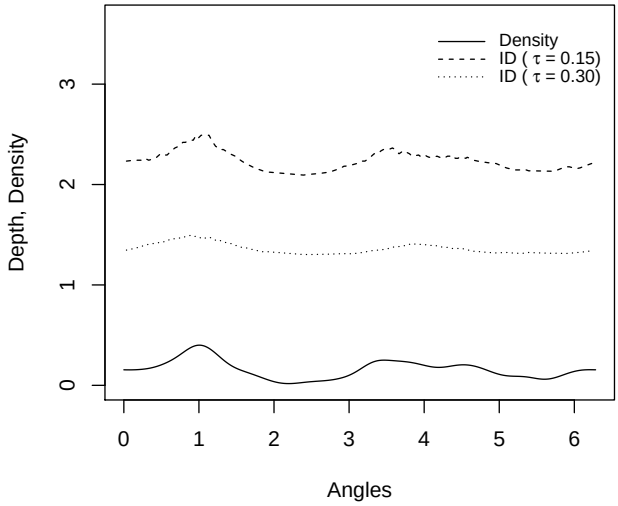

Figure 3. Plot of the interpoint spherical depth functions based on the arc length distance for the wind directions data in the Atlantic coast of Galicia (NW Spain). Calculation was performed for $\delta \in \{0.15, 0.30\}$ and $\xi = 0.1$. In order to make the visualizations easier to read, some noise was added to avoid depth curves' over-plotting.

4 Applications

In this section three applications, related to the location estimation, graph-based testing for equal distributions and spherical uniformity testing, are presented.

4.1 Location estimation

Given a notion of data depth, there is a natural choice of location parameter for the underlying distribution, namely the deepest point which is considered as a robust location estimator. A simulation study was performed to investigate the efficiency and robustness properties of the deepest points of the interpoint arc distance depth. For each combination of dimension $q \in \{3, 5, 10\}$, concentration parameter $\kappa \in \{3, 5, 10\}$ and sample size $n \in \{50, 100, 250, 500\}$, $R = 1000$ independent random samples of size n from the von Mises-Fisher distribution (vMF) with location parameter x^* were generated. For each

sample, an estimate $\hat{x_i^*}$ of the location parameter x^* is obtained. The empirical squared error for each replication is computed as follows:

$$SE\left(i\right) = \left\| \hat{x_i^*} - x^* \right\|^2 = 2\left(1 - \hat{x_i^*}' x^*\right).$$

The parameter δ and ξ were set equal to 0.5 and 0.1, respectively. Table 1 reports the resulting empirical mean squared errors which were computed as follows

$$EMSE = \frac{1}{R}\sum_{i=1}^{R} SE\left(i\right).$$

As expected, the efficiency of the $ID_{d_{arc}}$-estimator at the von Mises-Fisher distribution generally increases as the concentration of the data increases along with larger sample size n. This is true for all the considered dimensions even if it is worth noting that the efficiency is higher in the three dimensional space.

		n			
		50	***100***	***250***	***500***
q=3	$\kappa = 3$	0.866	0.688	0.348	0.524
	$\kappa = 5$	0.765	0.364	0.347	0.268
	$\kappa = 10$	0.239	0.242	0.125	0.093
q=5	$\kappa = 3$	1.107	1.00	1.084	0.989
	$\kappa = 5$	0.726	0.875	0.595	0.782
	$\kappa = 10$	0.456	0.451	0.281	0.354
q=10	$\kappa = 3$	1.339	1.333	1.082	1.324
	$\kappa = 5$	1.105	0.983	0.936	0.918
	$\kappa = 10$	0.642	0.731	0.663	0.658

Table 1: Empirical mean squared error of $ID_{d_{arc}}$ deepest point as location estimator from $R = 1000$ independent random samples for dimension $q \in \{3, 5, 10\}$, sample size $n \in \{50, 100, 250, 500\}$ and concentration parameter $\kappa \in \{3, 5, 10\}$ under von Mises-Fisher distribution.

Moving to the investigation of the robustness, the simulations were run considering the contaminated von Mises-Fisher model $vMF_\epsilon = (1-\epsilon)\, vMF(\mu,\kappa) + \epsilon\Delta_{x_c}$, where Δ_{x_c} denotes the point mass contamination at the point x_c, with $\epsilon = 10\%$ and 20% in dimension $d = 3$. Again, $R = 1000$ samples for different sample sizes $n \in \{100, 250, 500\}$ and concentration parameter $\kappa \in \{5, 10\}$ for dimension $q = 3$ were generated from vMF_ϵ. Because of the boundness of the spherical space, where contamination cannot be put at infinity, two different types of contamination were chosen. Specifically, antipodal and orthogonal to μ point mass contaminations were considered. In each sample, the $ID_{d_{arc}}$-deepest point was computed. The resulting empirical mean squared errors are provided in Table 2. The results show that the estimator associated with $ID_{d_{arc}}$ enjoys good robustness properties for both types of contamination. However, the orthogonal contamination seems to have less impact on the $ID_{d_{arc}}$-based location estimator.

	Contamination							
	Antipodal				***Orthogonal***			
	$\epsilon = 10\%$		$\epsilon = 20\%$		$\epsilon = 10\%$		$\epsilon = 20\%$	
$\boldsymbol{n}$	$\kappa = 5$	$\kappa = 10$	$\kappa = 5$	$\kappa = 10$	$\kappa = 5$	$\kappa = 10$	$\kappa = 5$	$\kappa = 10$
100	0.615	0.203	0.788	0.412	0.353	0.171	0.364	0.181
250	0.470	0.196	0.740	0.353	0.337	0.150	0.358	0.133
500	0.396	0.208	0.625	0.241	0.259	0.125	0.325	0.110

Table 2: The empirical squared errors of $ID_{d_{arc}}$-estimator of location obtained from $R = 1000$ independent random samples of size $n \in \{100, 250, 500\}$ from the contaminated distribution $vMF_\epsilon = (1-\epsilon)\, vMF(\mu,\kappa) + \epsilon\Delta_{x_c}$, where Δ_{x_c} denotes the point mass contamination (antipodal and orthogonal) at the point x_c, with $\epsilon = 10\%$ and 20%. In each case, the corresponding empirical mean square error is provided.

4.2 Depth-based graphical tool to test for equal distributions on spheres

Here, the proposal is to adopt a simple graphical tool by exploiting the properties of $ID_{d_{arc}}$, which is able to detect multimodality, and the so called depth vs. depth (DD) plot introduced by Liu et al. (1999). The DD-plot was introduced to graphically compare two multivariate distributions or samples through data depth and then further investigated by Li and Liu (2004) and also by Chavan and Shirke (2016). This is always a

two-dimensional plot regardless of the dimensions of the data. Distributional differences such as location, scale or skewness lead to different graphical patterns in the DD-space. This tool was also adopted to perform supervised depth-based classification of multivariate data in Li et al. (2012) and later extended to directional objects by Pandolfo et al. (2018) and Pandolfo and Porzio (2018). Specifically, the DD-plot displays the values of the depth function of the combined sample under the two corresponding empirical distributions. If these are identical, then the plot is a diagonal line from $(0,0)$ to $(1,1)$ in $\mathbb{R}^d$. This works also for spherical distributions if an appropriate rotational invariant data depth for directional data is adopted.

Let F and G be two distributions on S^{q-1} and $AD(\cdot,\cdot)$ be a rotational invariant depth for directional data. Let $X = \{X_1, \ldots, X_n\}$ and $Y = \{Y_1, \ldots, Y_m\}$ be two random sample drawn from F and G, respectively. Then, the empirical angular depth of each point $z \in X \cup Y$ with respect to F and G are computed

$$AD(z, F_m) \quad \text{and} \quad AD(z, G_n),$$

which are used as coordinates to plot data points on the DD-space. If $F = G$, then $AD(z, F_m) = AD(z, G_n)$ for all points.

To illustrate, samples from F and G with size $n = 1000$ for each one were generated, and $ID_{d_{arc}}$ was adopted by setting $\delta = 0.3$ and $\xi = 0.1$. The upper-left panel of Figure 4 displays the DD-plot of two samples drawn from a von Mises-Fisher distributions in dimension $q = 3$ with mean direction $(0,0,1)'$ and concentration parameter $\kappa = 3$. As one can see the data points are concentrated around the diagonal line. In the upper-right panel of Figure 4 the pattern deviates from the diagonal line because samples are drawn from two von Mises-Fisher distribution with different location and equal concentration. More in detail, the data cloud is symmetrically displayed around the diagonal line. Finally, the lower panel of Figure 4 displays the DD-plot of two samples drawn from two von Mises-Fisher distributions which differ in location and concentration (dispersion). Here, the data points' shape is not symmetric about the diagonal line and the observations seem to move towards the x-axis.

4.3 Test of uniformity on spheres

Testing uniformity of data on spheres is a fundamental hypothesis in directional statistics. The most known test of uniformity for directional data the are Rayleigh test (which rejects uniformity if the resultant of a sample is too large) and the Watson test (which is an adaptation for the circle of the Cramer-von Mises test). Numerous other uniformity tests have been proposed.

One alternative non-parametric approach is to exploit the properties of data depth for directional data when uniform distributions on S^{q-1} are considered. Specifically, in such case, the deepest point may not be unique and any rotational invariant depth function based on distances, such as $ID_{d_{arc}}$, turns out to be constant over S^{q-1}. This is a relevant difference between depths for directional data and depths for data in $\mathbb{R}^q$.

Theorem 4.1. *(Constancy of ID_{arc} on S^{q-1}) Let F be a continuous uniform distribution on S^{q-1} with density $f(\cdot)$, then $ID(x, F) = c$ with $\delta = 0.5$ and $\xi = 1$, for a positive*

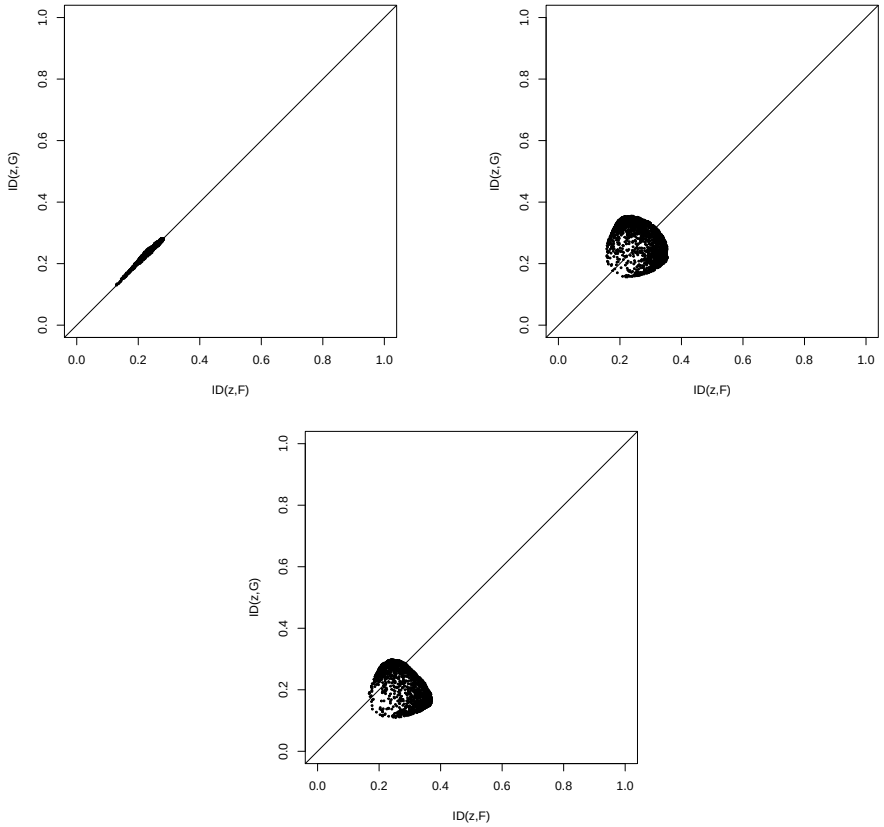

Figure 4. DD-plots based on $ID_{d_{arc}}$ of two samples drawn from identical von Mises-Fisher distributions (upper-left panel), from von Mises-Fisher distribution with different location and equal concentration (upper-right panel), and from two von Mises-Fisher distributions with different location and concentration (lower panel). In each case $n = 1000$ for each generated sample and $ID_{d_{arc}}$ was adopted with $\delta = 0.3$ and $\xi = 0.1$.

constant c and all $x \in S^{q-1}$. Moreover, the constant c is then equal to $\frac{\pi}{2+\pi}$.

Proof. The proof follows from:

i) in the case of uniform distribution F on S^{q-1} the 50^{th} quantile of the interpoint arc distance from x to the sample X_i is equal to $\frac{\pi}{2}$ irrespective of the dimension q, and

ii) the expected arc interpoint distance between uniformly distributed vectors on S^{q-1} is equal to $\frac{\pi}{2}$ irrespective of the dimension d.

The result follows from simple algebraic calculations. □

Then, a statistical depth-based test uniformity of n i.i.d. random vectors $X = \{X_1, \ldots, X_n\}$ from a distribution F on the sphere can be defined and the following test statistic can be used:

$$T_n = \sup_x \left| ID_{d_{arc}}(x, F_n) - c \right|, \tag{2}$$

where the constant c is independent of the dimension q and is equal to $\frac{2\pi}{2+\pi}$ ($c \approx 0.61$). Large values of T_n indicate that the distribution is unlikely to be uniform on S^{q-1}. The implementation of this test would require the knowledge of the exact distribution of the test statistic. Indeed, for testing the null hypothesis of spherical uniformity, the null hypothesis is rejected when $T_n > p_n$, where p_n is a percentile (according to the test level) of the distribution of T_n under the null hypothesis.

An approximation of the distribution of T_n under the null hypothesis can be obtained through the bootstrap resampling method. Hence, for R bootstrap replicates $T_n^{(i)}$, $i \in \{1, \ldots, R\}$ are obtained. Then, the critical values are obtained by computing the order-α quantile of the bootstrap distribution of T_n. A similar approach was followed by Dutta et al. (2011) and Paindaveine and Van Bever (2013) to evaluate data depth-based test of symmetry for data in $\mathbb{R}^q$.

A simulation study was conducted in order to investigate the finite-sample behavior of this test. The data were generated under the null hypothesis of uniformity on the $(q-1)$-dimensional unit sphere. Conversely, under the alternative hypothesis, data were generated from a von Mises-Fisher distribution with location $\mu = (1, 0, 0)'$ and concentration $\kappa = 1.5$. Specifically, the following setups were considered:

Setup 1: $X \sim U(\theta, \phi)$, where $\theta \sim \text{Unif}(0, 2\pi)$ and $\phi \sim \text{Unif}(0, \pi)$;

Setup 2: $X \sim vMF\left((1, 0, 0)', 1.5\right)$.

Simulations were run by generating samples of size $n \in \{100, 250, 500\}$. For each combination, the observed significance level was computed 1000 times, according to the resampling procedure above described from $R = 1000$ bootstrap samples. Table 3 reports the proportion of cases where the null hypothesis was rejected for the nominal value of $\alpha \in \{0.01, 0.05, 0.10\}$. This table clearly shows good level as well as power properties of the proposed test procedure.

	Setup 1			***Setup 2***		
Nominal level (α) $\rightarrow$	1%	5%	10%	1%	5%	10%
$n = 100$	0.095	0.055	0.014	0.997	0.996	0.989
$n = 250$	0.099	0.049	0.014	1.000	1.000	1.000
$n = 500$	0.105	0.054	0.012	1.000	1.000	1.000

Table 3: Proportion of cases where the null hypothesis of spherical uniformity was rejected for three nominal values of α, namely, 0.01, 0.05 and 0.10 under unimodal symmetric alternative.

Some tests of spherical uniformity fail when the alternative hypothesis concerns two antipodal mean directions (i.e. for antipodally symmetric distributions). This is, for instance, the case of the well known Rayleigh test of (hyper)spherical uniformity. The ID_{arc} can be really useful and satisfactorily responding when such case needs to be considered. In order to adopt the test statistic defined in (2) and to allow ID_{arc} to better capture the antipodally symmetric of the distribution, it is advisable to choose a value of the tuning parameter $\delta \approx 0.25$. However, in this case the value of the constant c varies according to the dimension of the data. Indeed, as shown in Cai et al. (2013), the distribution of the density of the arc distance between pairs of vectors uniformly distributed on S^{q-1} is unimodal with mode around $\pi/2$, and as the dimension increases the concentration of angles around $\pi/2$ gets stronger and stronger. Hence, the density varies according to the dimension q:

$$f_q(\theta) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{q}{2}\right)}{\Gamma\left(\frac{q-1}{2}\right)} \left(\cos \frac{\theta}{q-2}\right)^{q-2}, \quad \theta \in [0, \pi], \tag{3}$$

where Γ denotes the gamma function.

Now, it is simple to compute the inverse function of (3) and the get any th-order quantile of the arc distance distribution of uniformly distributed vectors on S^{q-1} for a fixed dimension q. Below some numerical examples are presented to illustrate its applications in the case of antipodal symmetry as alternative hypothesis in dimension $q = 3$.

Hence, if $X = \{X_1, \ldots, X_n\}$ is a random sample of vectors uniformly distributed on S^2, then $ID(x, F) = c$ with $\delta = 0.25$ and $\xi = 1$, for a positive constant c and all $x \in S^2$. Moreover, $c = \frac{3\pi}{6+2\pi}$ ($c \approx 0.76$).

To evaluate how the proposed test performs the antipodally symmetric Bingham distribution (Bing(A)) introduced by Bingham (1974) is considered. Its density function is

given by

$$f(x, A) = \frac{1}{c_{Bing}(A)} \exp\left\{-x'Ax\right\}, \tag{4}$$

where $A = MZM'$ denotes a $q \times q$ matrix with $M \in \mathbb{R}^{q \times q}$ an orthogonal matrix ($MM' = M'M = I^{q \times q}$) and Z is a diagonal concentration matrix. The normalizing constant $c_{Bing}(A)$ is expressed as a hyper-geometric function of A. The density is symmetric ($f(x) = f(-x)$) and the exponent is quadratic in x.

The following setups with increasing concentration (actually a low concentration value leads to a uniform distribution) around the two modes were considered:

Setup 3: $X \sim$Bing(A), with $A = \text{diag}(1, 0, -1)$

Setup 4: $X \sim$Bing(A), with $A = \text{diag}(2, 0, -2)$.

Again, $R = 1000$ bootstrap samples of size $n \in \{100, 250, 500\}$ were generated, and the null hypothesis was rejected for three nominal values of α, namely, 0.01, 0.05 and 0.10 (with δ set equal to 0.25). The rejection frequencies are reported in Table 4. In the first case, with data less concentration around the two modes, the proposed test performs well when for $n = 250$ and 500 (rejection frequencies are slightly different), while for $n = 100$ it shows worse results and thus it should be used with more caution.

	Setup 3			***Setup 4***		
Nominal level (α) $\rightarrow$	1%	5%	10%	1%	5%	10%
$n = 100$	0.849	0.772	0.550	0.997	0.997	0.997
$n = 250$	0.994	0.993	0.976	1.000	1.000	1.000
$n = 500$	0.996	0.996	0.996	1.000	1.000	1.000

Table 4: Proportion of cases, out of 1000, where the null hypothesis of spherical uniformity was rejected for three nominal values of α, namely, 0.01, 0.05 and 0.10 under antipodally symmetric alternative.

5 Concluding remarks

In this article, a notion interpoint data depth function for directional data has been proposed and investigated. This is based on the arc distance, and hence called "interpoint arc distance depth" (ID_{arc}), and turns out to be useful in case of multimodality on

spherical spaces and thus really useful for analyzing local features of data on the hyperspheres, where standard notions of depth function are not informative. Some theoretical properties are established and several interesting applications in location estimation, graphical test of equal distributions and test of spherical uniformity are shown through several simulated examples. Results show the efficiency and robustness of ID_{arc}, and the powerful of the statistics based on it to test spherical uniformity. In addition, this notion of data depth has the important advantage to be not computational expensive. Further research could concern the use of other alternative distance measure for directional data and additional applications where depths based on interpoint distance for directional data may be of some interest.

References

Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S. (2005). Clustering on the unit hypersphere using von Mises-Fisher distributions. *J. Mach. Learn. Res.*, 6:1345–1382.

Bingham, C. (1974). An antipodally symmetric distribution on the sphere. *Ann. Statist.*, 2(6):1201–1225.

Buchta, C., Kober, M., Feinerer, I., and Hornik, K. (2012). Spherical k-means clustering. *Journal of Statistical Software*, 50(10):1–22.

Cai, T., Fan, J., and Jiang, T. (2013). Distributions of angles in random packing on spheres. *The Journal of Machine Learning Research*, 14(1):1837–1864.

Chavan, A. R. and Shirke, D. T. (2016). Nonparametric tests for testing equality of location parameters of two multivariate distributions. *Electronic Journal of Applied Statistical Analysis*, 9(2):417–432.

Dong, Y. and Lee, S. M. (2014). Depth functions as measures of representativeness. *Statistical Papers*, 55(4):1079–1105.

Dutta, S., Ghosh, A. K., Chaudhuri, P., et al. (2011). Some intriguing properties of tukey's half-space depth. *Bernoulli*, 17(4):1420–1434.

Fisher, N. (1985). Spherical medians. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 342–348.

Gatto, R. and Jammalamadaka, S. R. (2007). The generalized von mises distribution. *Statistical Methodology*, 4(3):341–353.

Ley, C. and Verdebout, T. (2017). *Modern directional statistics*. Chapman and Hall/CRC.

Li, J., Cuesta-Albertos, J. A., and Liu, R. Y. (2012). DD-classifier: Nonparametric classification procedure based on dd-plot. *Journal of the American Statistical Association*, 107(498):737–753.

Li, J. and Liu, R. Y. (2004). New nonparametric tests of multivariate locations and scales using data depth. *Statistical Science*, pages 686–696.

Liu, R. Y., Parelius, J. M., Singh, K., et al. (1999). Multivariate analysis by data depth:

descriptive statistics, graphics and inference,(with discussion and a rejoinder by liu and singh). *The annals of statistics*, 27(3):783–858.

Liu, R. Y. and Singh, K. (1992). Ordering directional data: concepts of data depth on circles and spheres. *The Annals of Statistics*, pages 1468–1484.

Liu, Z. and Modarres, R. (2011). Lens data depth and median. *Journal of Nonparametric Statistics*, 23(4):1063–1074.

Lok, W. and Lee, S. M. (2011). A new statistical depth function with applications to multimodal data. *Journal of Nonparametric Statistics*, 23(3):617–631.

Mardia, K. V. and Jupp, P. E. (2009). *Directional statistics*, volume 494. John Wiley & Sons.

Oliveira, M., Crujeiras, R. M., and Rodríguez-Casal, A. (2014). Circsizer: an exploratory tool for circular data. *Environmental and ecological statistics*, 21(1):143–159.

Paindaveine, D. and Van Bever, G. (2013). From depth to local depth: a focus on centrality. *Journal of the American Statistical Association*, 108(503):1105–1119.

Pandolfo, G., D'Ambrosio, A., and Porzio, G. C. (2018). A note on depth-based classification of circular data. *Electronic Journal of Applied Statistical Analysis*, 11(2):447–462.

Pandolfo, G. and Porzio, G. (2018). DD-classifier for angular data with an application to protein structures. In *Book of Abstracts*, page 66.

Wilson, R. C., Hancock, E. R., Pekalska, E., and Duin, R. P. (2014). Spherical and hyperbolic embeddings of data. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2255–2269.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Annals of statistics*, pages 461–482.