Computational Statistics (2023) 38:1487–1506
https://doi.org/10.1007/s00180-022-01281-w

ORIGINAL PAPER

Clustering directional data through depth functions

Giuseppe Pandolfo[1] · **Antonio D'ambrosio**[1]

Received: 17 March 2021 / Accepted: 30 August 2022 / Published online: 19 September 2022

Abstract
A new depth-based clustering procedure for directional data is proposed. Such method is fully non-parametric and has the advantages to be flexible and applicable even in high dimensions when a suitable notion of depth is adopted. The introduced technique is evaluated through an extensive simulation study. In addition, a real data example in text mining is given to explain its effectiveness in comparison with other existing directional clustering algorithms.

Keywords Spherical random variables · Spherical distance · Textual data · von Mises-Fisher

1 Introduction

Clustering is a fundamental subject in statistics which has been widely investigated in classical multivariate analysis over the past decades. It aims at organizing a set of objects into homogeneous groups, such that objects in the same cluster (or group) are more "similar" to each other than to objects in different clusters. Because of its practical importance, clustering has received a lot of attention in various branches of statistics. However, less attention has been paid to the cluster analysis of directional data even though it is becoming more and more important in modern applications where observations are represented by angles or unit vectors. A few examples are the wind direction (meteorology), the orientation of magnetic fields in rocks (geology) and the movement of animals (biology). In the recent years, thanks to new visualization techniques (see, Buttarazzi et al. 2018; Pandolfo 2022), they have also seen a significant and growing interest by neuroscientists (to study the direction of neuronal axons and dendrites) and microbiologists (to analyze the angles formed by protein structures).

Directional data are constrained to lie on the surface of the unit $(d-1)$-dimensional hypersphere $\mathbb{S}^{d-1} := \left\{x \in \mathbb{R}^d : \|x\|_2 = 1\right\}$, where $\|x\|_2 := \sqrt{\sum_{i=1}^{d} x_i^2}$, with

✉ Giuseppe Pandolfo
giuseppe.pandolfo@unina.it

[1] Department of Economics and Statistics, University of Naples Federico II, Naples, Italy

$x = (x_1, \ldots, x_d)'$. Within the literature, they are presented and discussed by Mardia and Jupp (2000), which is a classical reference on the subject. More recent developments of this branch of statistics can be found in Ley and Verdebout (2017, 2018).

Analyzing and describing directional data requires tackling some interesting problems associated with the lack of a reference direction and with a sense of rotation not uniquely defined. One more important issue is the lack of a natural ordering, which generates a special interest in depth functions on the (hyper)sphere.

Such peculiar features make the use of classical statistical methods inappropriate, and often misleading. In this regard, consider the angles 0 and 2π on a unit circle $\mathbb{S}^1$. Their arithmetic mean is π, but they are actually the same angle and the "true" mean is 0. Thus, working with directional data requires specific techniques that consider the geometry of the manifold, and this obviously holds true also for clustering issues.

Despite the theory on clustering methods based on depth functions in $\mathbb{R}^d$ has been recently established and it is still relatively young and under development (see e.g. Jörnsten 2004, Ding et al. 2007 and Jeong et al. 2016), to the best of authors' knowledge, there is no work of such type to perform clustering of high dimensional directional data in the literature. However, angular depth functions were recently employed to perform classification of hyperspherical objects (see Pandolfo and D'Ambrosio 2021).

The idea of data depth, a measure of how deep or outlying a given multidimensional point is w.r.t. a distribution F or w.r.t. a data cloud $\{X_1, \ldots, X_n\}$, allows generalizing the concepts of median and rank to multivariate data. This way, a multidimensional center-outward order (similar to that of a real line) can be obtained. Obviously, this holds true also for directional data.

Depth-induced ordering enables a description of multivariate distributions and aids in using depth functions for clustering as shown in Jörnsten (2004) and Torrente and Romo (2021) who have proven power of depth function methodology. Then, even though the use of a depth notion in the directional data clustering problem is yet a new and unexplored tool, it is highly appealing to extend these ideas to such complex setting of data.

Hence, in this paper we propose a non-parametric method for performing cluster analysis of directional data based on the concept of data depth for directional data. The proposal aims to be considered as an alternative tool in the general framework of clustering methods for directional data.

The paper is organized as follows. In Sect. 2 the von Mises-Fisher distribution, the reference distribution in directional data analysis, is briefly recalled. In Sect. 3, we provide a brief review of angular data depths available within the literature. Sect. 4 reviews existing methods for clustering directional data. In Sect. 5 the depth-based clustering procedure is presented. Section 6 contains an evaluation of the proposed method through a simulated study. A real data application for textual data analysis is presented in Sect. 7. Finally, Sect. 8 contains some concluding remarks.

2 Preliminaries: the von Mises-Fisher distribution

In this section, we briefly review the von Mises-Fisher (vMF) distribution, which arises naturally for data distributed on the unit hypersphere. A $(d-1)$-dimensional unit random vector x (i.e., $x \in \mathbb{R}^d$ and $\|x\|_2 = 1$, or equivalently $x \in \mathbb{S}^{d-1}$) is said to have a von Mises-Fisher distribution if its probability density function is given by

$$f(x|\mu,\kappa) = c_d(\kappa)\exp\left\{\kappa\mu' x\right\},$$

where $\|\mu\|_2 = 1$, $\kappa \geq 0$ and $d \geq 2$. The normalizing constant $c_d(\kappa)$ is given by

$$c_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)},$$

where $I_\nu(\cdot)$ represents the modified Bessel function of the first kind and order ν. The density $f(x|\mu,\kappa)$ is parametrized by the mean direction μ, and the concentration parameter κ, so-called because it characterizes how strongly the unit vectors drawn according to $f(x|\mu,\kappa)$ are concentrated about μ. Larger values of κ imply stronger concentration about the mean direction. In particular when $\kappa = 0$, $f(x|\mu,\kappa)$ reduces to the uniform density on $\mathbb{S}^{d-1}$, and as $\kappa \to \infty$, $f(x|\mu,\kappa)$ tends to a point density. Such distribution is a reference one for directional data, and has properties analogous to those of the multivariate Normal distribution for data in $\mathbb{R}^d$. For example, the maximum entropy density on $\mathbb{S}^{d-1}$ subject to the constraint that $E[x]$ is fixed is a vMF density. Figure 1 illustrates the impact of κ, by presenting four samples of 1000 points on the unit sphere (i.e $d = 3$) according to four vMF distributions with the same mean direction, $\mu = (0, 0, 1)'$, but with different values of $\kappa \in \{0, 5, 20, 100\}$.

3 Data depth for directional data

In this section, we recall the notions of data depth for directional data available within the literature, that is the angular simplicial depth, the angular halfspace depth, the arc distance depth, the cosine distance depth, and the chord distance depth. The first three were introduced and investigated by Liu and Singh (1992), while the latter by Pandolfo et al. (2018).

Definition 1 Angular simplicial depth (ASD). The angular simplicial depth of a given point $x \in \mathbb{S}^{d-1}$ w.r.t. the distribution F on the unit hypersphere $\mathbb{S}^{d-1}$ is defined as follows:

$$ASD(x,F) := P_F\big(x \in \Delta\big(W_1,\ldots,W_d\big)\big),$$

where P_F denotes the probability content w.r.t. the distribution F. $W_1,\ldots,W_d$ are i.i.d. observations from F and $\Delta\big(W_1,\ldots,W_d\big)$ denotes the simplex with vertices $W_1,\ldots,W_d$.

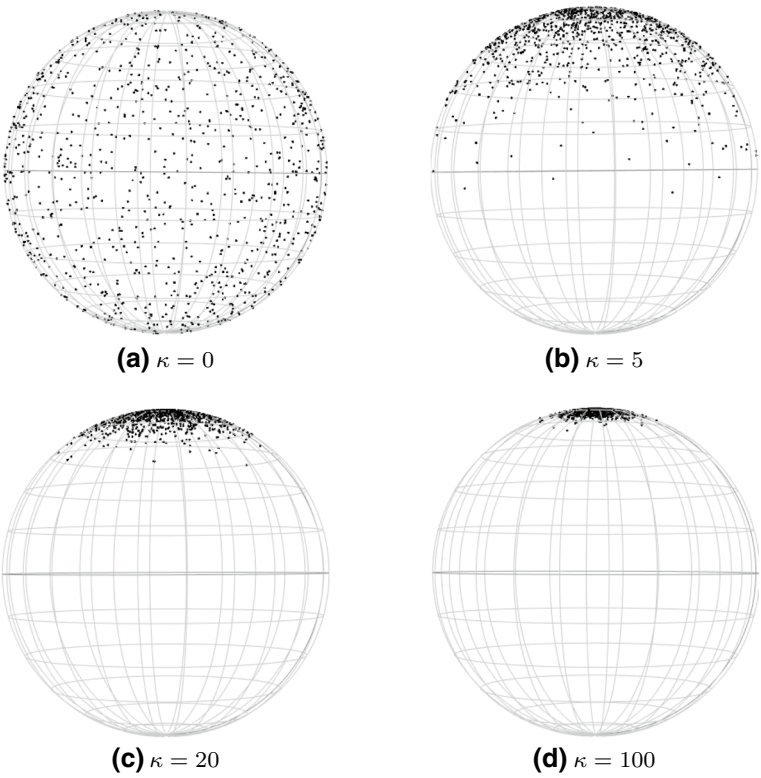

Fig. 1 Plots of four samples of 1000 data points drawn from a von Mises-Fisher distribution on $\mathbb{S}^2$ for concentration parameter κ equal to 0 (**a**), 5 (**b**), 20 (**c**) and 100 (**d**)

Definition 2 Angular Tukey's depth (ATD). The angular Tukey's depth of a given point $x \in \mathbb{S}^{d-1}$ w.r.t. the distribution F on $\mathbb{S}^{d-1}$ is defined as follows:

$$ATD(x, F) := \inf_{HS : x \in HS} P_F(HS),$$

where HS denotes the set of all closed hemispheres containing x.

Definition 3 Arc distance depth (ADD). The arc distance depth of a given point $x \in \mathbb{S}^{d-1}$ w.r.t. the distribution F on $\mathbb{S}^{d-1}$ is defined as follows:

$$ADD(x, F) := \pi - \int \ell(x, \varphi) dF(\varphi),$$

where $\ell(x, \varphi)$ is the Riemannian distance between x and φ (i.e. the length of the shortest arc joining x and φ).

Definition 4 Cosine distance depth (CDD). The cosine distance depth of a given point $x \in \mathbb{S}^{d-1}$ w.r.t. the distribution F on $\mathbb{S}^{d-1}$ is defined as follows:

$$CDD(x, F) := 2 - E_F[1 - x'W],$$

where E_F is the expectation under the assumption that W has distribution F.

Definition 5 Chord distance depth (ChDD). The chord distance depth of a given point $x \in \mathbb{S}^{d-1}$ w.r.t. the distribution F on $\mathbb{S}^{d-1}$ is defined as follows :

$$ChDD(x, F) := 2 - E_F\left[\sqrt{2(1 - x'W)}\right],$$

where E_F is the expectation under the assumption that W has distribution F.

For convenience, in the following we denote with $AD(\cdot, \cdot)$ the population version of a general angular depth function, unless a particular notion is adopted.

The sample version, henceforth denoted by $AD^n(\cdot, \cdot)$, is obtained by replacing F by its empirical counterpart $\hat{F}$ calculated from the sample $X_1, \ldots, X_n$.

All the depth functions listed above possess the following important properties:

P1. *Rotation invariance*: $AD(x, F) = AD(Ox, OF)$ for any $d \times d$ orthogonal matrix O;

P2. *Maximality at center*: $\max_{x \in \mathbb{S}^{d-1}} AD(x, F) = AD\big(x_0, F\big)$ for any F with center at x_0;

P3. *Monotonicity on rays from the deepest point*: $AD(\cdot, \cdot)$ decreases along any geodesic path $t \mapsto x_t$ from the deepest point x_0 to the antipodal point $-x_0$.

In addition, one more important property is satisfied by the distance-based depths, that is:

P4. *Minimality at the antipodal point to the center*: $AD\big(-x_0, F\big) = \inf_{x \in \mathbb{S}^{d-1}} AD(x, F)$ for any F with center at x_0.

One further available notion of depth for directional data is the angular Mahalanobis depth of Ley et al. (2014), which is based on a concept of directional quantiles. However, it requires a preliminary choice of a spherical location and for this reason it is not considered in this work. Moreover, for the purpose of this work, the ASD and ATD will not be considered. This is because they have two main drawbacks that make them not feasible for the application of the proposed method to high-dimensional data, that is: (i) They are high computationally demanding and (ii) can take zero values (see Liu and Singh 1992), while distance-based directional depths take positive values everywhere on $\mathbb{S}^{d-1}$ (but in the uninteresting case of a point mass distribution).

Note that depth functions are not to intended to be equivalent with density. However, contours of depth are often used to reveal the shape and the structure of a multivariate data set. Such contours are analogous to quantiles in the

univariate case, and they allow for the computation of L-estimators of location-scatter parameters (such as trimmed means).

Unlike the univariate case, multivariate ordering can be defined in different ways, but it is usually desirable for samples from a certain class of distributions such as the rotationally symmetric ones, the angular depth contours should track the contours of the underlying model (and thus circularly contoured). Such contours thus are formed by a $(d-1)$-dimensional sphere (a circle inside the sphere $\mathbb{S}^2$, two points on the circle $\mathbb{S}^1$) centered at x_0. Rotationally symmetric distributions are characterized by densities of the form

$$x \mapsto f_{x_0}(x) = c_{\kappa,f} f\left(x' x_0\right), \quad x \in \mathbb{S}^{d-1}, \tag{1}$$

where $f : [-1, 1] \mapsto \mathbb{R}_0^+$ is an absolutely continuous and monotone increasing function and $c_{\kappa,f}$ a normalizing constant. Such class of distributions contains the von Mises-Fisher distribution which is obtained by taking $f(u) = \exp(\kappa u)$ for some strictly positive concentration parameter κ. Note that Theorems 3 and 4 in Pandolfo et al. (2018) ensures that the maximal depth measures the concentration of F, irrespective of the chosen distance measure for distributions having density of the form given in (1).

It is usually convenient to treat depth contours in terms of their corresponding α-regions that, as usual, are defined as the collection of x values with a depth larger than or equal to α.

Definition 6 For a given angular depth function $AD(x, F)$ and for $\alpha > 0$, we call

$$AD^{\alpha}(F) \equiv \left\{x \in \mathbb{S}^{d-1} | AD(x, F) \geq \alpha\right\}$$

the corresponding α-depth region and its boundary $\partial AD^{\alpha}(x, F)$ the corresponding α -contour.

The α-depth regions for data on $\mathbb{S}^{d-1}$ usually are desired to be rotationally invariant and nested. In addition, it is worth underlying that the α-regions induced by ADD, CDD and ChDD are invariant under rotations when fixing x_0, hence they are able to reflect the symmetry of the distribution F about x_0. Roughly speaking, each depth region can be considered a sort of measure of the distance of a given point from a central point of the distribution, where the depth function takes its maximum. Hence, more dispersed data lead to larger regions.

To illustrate the angular depth ordering and its ramifications, the normalized ADD, CDD and ChDD are used to order sample points and graph their corresponding representative contours. Applying depth ordering to a sample of 1000 points drawn from a von Mises-Fisher distribution on $\mathbb{S}^2$ with concentration parameter $\kappa = 15$, the sample pth level contours in Fig. 2 were obtained. For the sake of the illustration, data are presented in spherical coordinates using angles θ and ϕ with unit radius. As one can see the contours are nested within one another.

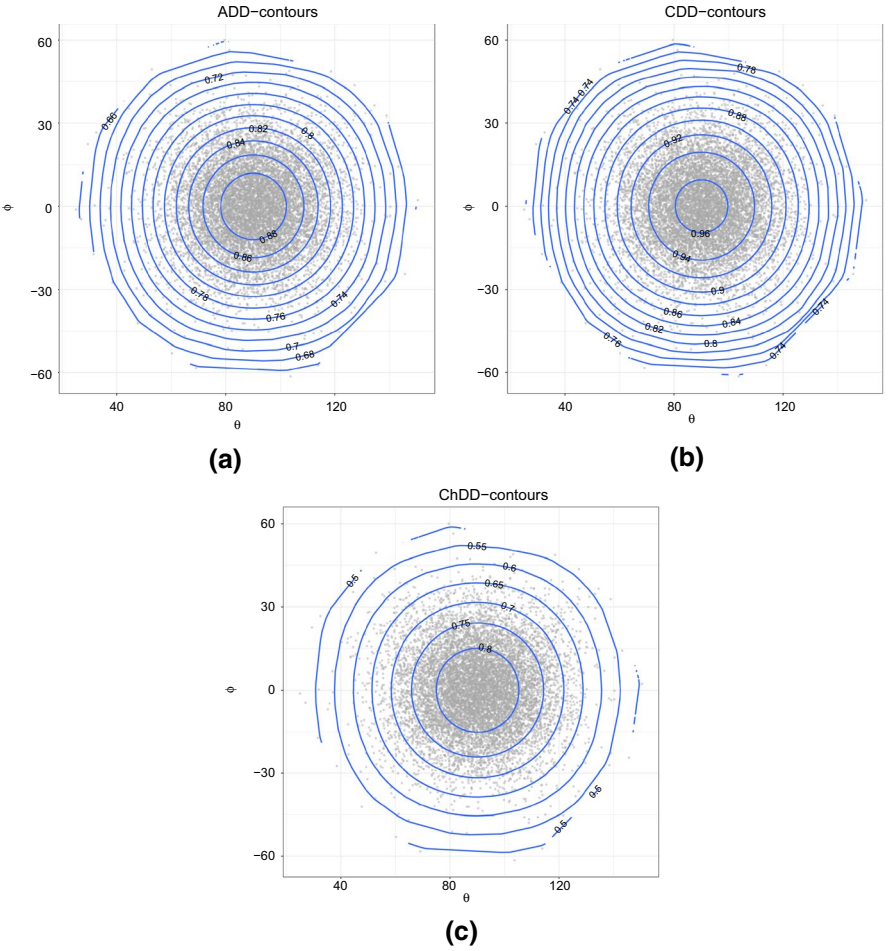

Fig. 2 Plots of the ADD (**a**), CDD (**b**) and ChDD (**c**) empirical contours of a sample of 1000 data points drawn from a von Mises-Fisher distribution on $\mathbb{S}^2$ with concentration parameter $\kappa = 15$

4 Clustering directional data: a brief review

The last few decades have witnessed an increasing interest in classification methods for directional data in a broad sense (which includes both classification and clustering techniques).

The literature on supervised classification of directional data is quite rich. SenGupta and Roy (2005) proposed a discrimination rule based on the chord distance. Ackermann (1997) adapted discriminant analysis procedures for linear data to the analysis of circular data, while a comparison of different classification rules on the unit circle was performed by Tsagris and Alenazi (2019). The Naive Bayes classifier for directional data was introduced by López-Cruz et al. (2015), and the discriminative directional logistic regression by Fernandes and Cardoso (2016). More recently,

Di Marzio et al. (2019) considered non-parametric circular classification based on Kernel density estimation. Pandolfo and D'Ambrosio (2021) and Demni and Porzio (2021) studied the depth-versus-depth (DD) classifier for directional data, while Demni et al. (2019) introduced a cosine depth based distribution method.

The development of clustering techniques for directional data has recently been an important research topic. The two most popular approaches to perform clustering of data on $\mathbb{S}^{d-1}$ are the spherical K-means (Dhillon and Modha 2001a) and the use of mixture models with vMF components. The former aims at maximizing the cosine similarity $\sum_{i=1}^{n} X_i' c_i$ between the sample $X_1, \ldots, X_n$ and K centroids $c_1, \ldots, c_\kappa \in \mathbb{S}^{d-1}$, where c_i is the centroid of the cluster containing X_i. Dhillon and Sra (2003) and Banerjee et al. (2003, 2005) gave Expectation-Maximization (EM) algorithm for fitting vMF mixtures which have spherical K-means as a particular case. Other different approaches for fitting vMF mixtures were proposed by Yang and Pan (1997), based on embedding fuzzy c-partitions in the mixtures, and Taghia et al. (2014) who addressed the Bayesian estimation of the vMF mixture model employing variational inference. Franke et al. (2016) developed an EM algorithm to fit general projected normal mixtures on $\mathbb{S}^2$.

A fuzzy variation of the K-means clustering procedure for directional data was proposed by Kesemen et al. (2016), while Benjamin et al. (2019) introduced possibilistic c-means.

5 Depth-based medoids clustering algorithm

In this section a depth-based medoids clustering algorithm (DBMCA) for directional data is introduced. Some concepts which are required for the formulation of the proposed algorithm are defined below.

Definition 7 (Depth-based partition): A depth-based partition $\mathcal{C}_k$ is a non-empty maximal and depth-based subset of a $n \times (d-1)$ data set $\mathbf{X}$ defined in $\mathbb{S}^{d-1}$ such that $\mathbf{X} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k, \ldots \mathcal{C}_K\}$, where $K \leq n$.

Definition 8 (Depth-medoid): A depth-medoid $x_{AD_k} \in \mathbb{S}^{d-1}$ is a data point belonging to the depth-based partition $\mathcal{C}_k$ having distribution $F_{\mathcal{C}_k}$ for which

$$x_{AD_k} := \underset{x \in \mathbb{S}^{d-1}}{\operatorname{argmax}} \, AD^n\left(x, \hat{F}_{\mathcal{C}_k}\right),$$

where $\hat{F}_{\mathcal{C}_k}$ is the sample distribution function of $F_{\mathcal{C}_k}$. Hence, the depth-medoid is the deepest point within the kth depth-based partition.

Then, in order to evaluate the goodness of the partition into K clusters, the proposed algorithm makes use of a depth-based cluster homogeneity measure. As highlighted in Hoberg (2000), data showing larger dispersion give rise to larger α -depth regions, and consequently, homogeneity can be measured by considering the depth values of data points within regions. More in detail, the idea is to use

a *within-class depth-based concentration measure* by following the depth-based dispersion measure proposed by Romanazzi (2009) for data in $\mathbb{R}^d$ and later used by Agostinelli and Romanazzi (2013) for the analysis of directional data. Specifically, the depth-based homogeneity within the kth cluster can be defined as the expected measure of the angular depth within each cluster $\mathcal{C}_k$ having distribution $F_{\mathcal{C}_k}$

$$DW_k := \int_{\mathbb{S}^{d-1}} AD\left(x, F_{\mathcal{C}_k}\right) dv(x), \tag{2}$$

where v denotes the Lebesgue measure on $\mathbb{S}^{d-1}$.

Denote by $\mathcal{S}_k = \{x_{k1}, \ldots, x_{kn_k}\}$ the finite set of points within the kth cluster, then (2) can be approximated for sample data as

$$\widetilde{DW}_k := \frac{1}{n_k} \sum_{i=1}^{n_k} AD^n(x_{ki}, \mathcal{S}_k). \tag{3}$$

The algorithm searches for the partition of the data points into K clusters in such a way that the expected angular depth of each cluster is maximized.

The proposed algorithm follows the well-known *partitioning around medoids* (PAM) approach (Kaufman and Rousseeuw 1987). The main difference is that we iteratively search for those points, the depth-medoids, that maximize a given depth function between themselves and other points belonging to the same partition.

More formally, let $\mathbf{X}$ be a $n \times (d-1)$ data set defined in $\mathbb{S}^{d-1}$, randomly select K observations as depth-medoids. Iteratively:

- Assign each data point to the cluster w.r.t. its angular depth is maximum (assignment step);
- Refine the K depth-medoids by choosing for each cluster those points that allow for the maximization of the depth-based within-cluster homogeneity as defined in (3) (refinement step).

Here, for evaluating the clustering results and determining the optimal number of clusters, the Silhouette index of Rousseeuw (1987) is adopted. It defines for each object in a data set, the measure of how this object is similar to other objects from the same cluster (cohesion) in comparison with objects of other clusters (separation). Specifically, in our approach, data consists of angular depth values. Hence, assume the data have been clustered into K clusters, then for each data point $x_i \in \mathcal{C}_i$, we have

$$\begin{aligned} a'(i) &= \text{angular depth value of } x_i \text{ w.r.t. } \mathcal{C}_i \\ d'(i) &= \text{angular depth value of } x_i \text{ w.r.t. the objects in any other cluster} \\ b'(i) &= \max_{j \neq i} d'(i, \mathcal{C}_j) \end{aligned}$$

then, with such measure, the silhouette coefficient is defined as:

$$s(i) = \begin{cases} 1 - b'(i)/a'(i) & \text{if } a'(i) > b'(i) \\ 0 & \text{if } a'(i) = b'(i) \\ a'(i)/b'(i) - 1 & \text{if } a'(i) < b'(i) \end{cases} \tag{4}$$

5.1 Initialization of cluster centers

Initialization of iterative algorithms can have a significant impact on the resulting performances. Several techniques can be found within the literature. Here, we propose a modified version of the well-known `K-means++` algorithm proposed by Arthur and Vassilvitskii (2007). Specifically, assuming the number of clusters is equal to K:

Step 1 Select an initial medoid m_1 uniformly at random from the data set $\mathbf{X}$.
Step 2 Then, each additional medoid m_j ($j = 2, \ldots, K$) is selected in the following way:

a. Compute the angular data depth of m_{j-1} w.r.t. $\mathbf{X}$

$$AD^n_{m_{j-1}} = AD^n(m_{j-1}, \mathbf{X}).$$

b. Compute the angular depth of each of the other non-selected data points x_i ($i = 1, \ldots, n - (j-1)$) w.r.t. $\mathbf{X}$

$$AD^n_{x_i} = AD^n(x_i, \mathbf{X}).$$

c. Then compute the difference between the angular depths of points x_i and the angular depth of the $(j-1)$th medoid as follows:

$$\Delta_{x_i} = |AD^n_{x_i} - AD^n_{m_{j-1}}|.$$

d. Choose one new data point at random as a new medoid from $\mathbf{X}$ with probability

$$\frac{\Delta_{x_i}}{\sum_{i=1}^{n-(j-1)} \Delta_{x_i}}.$$

That is, select each subsequent medoid with a probability proportional to difference between its angular depth and the angular depth of the previously chosen medoid.

Step 3 Repeat Step 2 until K medoids are chosen.

6 Simulation study

We have run a comprehensive simulation study in order to evaluate the proposed methodology. We have considered four factors, i.e. the number of clusters, the dimensionality of the problem, the level of noise, and structured vs unstructured data. We generated data with two, three, four and five theoretical partitions. We have sampled data from the von Mises-Fisher distribution in dimensions $d \in \{3, 5, 10\}$. The level of noise was set by randomly choosing the value of the concentration parameters $\kappa_{Low} \sim U[10, 12]$, $\kappa_{Medium} \sim U[6, 8]$ and $\kappa_{High} \sim U[2, 4]$ for the cases of low, medium and high noise level, respectively.

In the case of structured data, we first selected randomly a point as center of the first partition. Then,

- For the case of two centers, the second cluster center was randomly chosen with the constraint that the cosine distance from the first was between 1.7 and 2;
- In the case of three centers, we randomly added a point with the constraint that the cosine distance from both the first two data points was between 0.8 and 1.2;
- In the case of four centers, we randomly added a data point with the constraint that the cosine distance from the third center was between 1.7 and 2;
- In the case of five centers, we randomly added a data point with the constraint that the cosine distance from all other centers was between 0.5 and 0.8.

In the case of unstructured data, the theoretical centers were sampled without any constraint from a von Mises-Fisher distribution with concentration parameter $\kappa = 0$.

For any combination, we generated a sample of size 500. The proportion of cases was randomly chosen in such a way that the clusters were not extremely unbalanced. For each condition and for each level, ten data sets were generated, for a total of 720. Table 1 summarizes the factors and the levels of the factorial design.

We adopted the cosine distance depth (CDD), the arc distance depth (ADD) and the chord distance depth (ChDD). For each trial, the number of clusters ranged from 1 to 10.

As for each data set we know the theoretical partition, to determine how well the resulting partition matched the gold standard partition, we used the adjusted Rand index (ARI) of Hubert and Arabie (1985) which is a modified version of the Rand index which adjusts for agreement by chance. It is defined as follows.

Table 1 Summary of independent factors and levels of the simulation study

Factor	Level
Number of clusters	2, 3, 4, 5
Dimensions	3, 5, 10
Noise level	Low, moderate, high
Data structure	Structured, unstructured

Given an $n \times p$ data matrix X, where n is the number of objects and p the number of variables, the crisp clustering structure can be presented as a set of nonempty $K \geq 2$ subsets $\{C_1, \ldots, C_k, \ldots, C_K\}$ such that:

$$X = \bigcup_{k=1}^{K} C_k, \; C_k \bigcap C_{k'} = \emptyset, \quad \text{for } k \neq k'.$$

Two objects of X, i.e., (x, x'), are paired in C if they belong to the same cluster. Let P and Q be two partitions of X. The Rand index is calculated as follows:

$$RI = \frac{a+d}{a+b+c+d} = \frac{a+d}{\binom{n}{2}},$$

where

- a is the number of pairs $(x, x') \in X$ that are paired in P and in Q;
- b is the number of pairs $(x, x') \in X$ that are paired in P but not paired in Q;
- c is the number of pairs $(x, x') \in X$ that are not paired in P but are paired in Q and
- d is the number of pairs $(x, x') \in X$ that are not paired in either P or in Q.

The Rand index takes values in [0, 1], with 0 indicating that the two partitions do not agree for any pair of elements and 1 indicating that the two partitions are exactly the same. Moreover, the Rand index is reflexive, namely, $RI(P, P) = 1$. However, such index has some drawbacks: (i) it concentrates in a small interval close to 1, thus presenting low variability; (ii) it approaches its upper limit as the number of clusters increases; and (iii) it is extremely sensitive to the number of groups considered in each partition and their density. To overcome these problems, Hubert and Arabie (1985) proposed the adjusted Rand index (ARI), which can be defined as follows:

$$ARI = \frac{2(ad - bc)}{b^2 + c^2 + 2ad + (a+d) + (c+b)}.$$

The ARI can yield a value between -1 and $+1$. ARI equal to 1 means complete agreement between two clustering results, whereas ARI equal to -1 means no agreement between two clustering results.

6.1 Simulation results

The results of the simulation study are presented in Figs. 3 and 4 for structured and unstructured data, respectively. Boxplots of the ARI values of the proposed method for any combination of a dimension, a depth function, a noise level and a number of clusters are plotted there.

As one can see, since the performance of the depth functions under evaluation appear quite similar for both structured and unstructured data, no clear-cut indications arise on which of them is to be preferred in such cases.

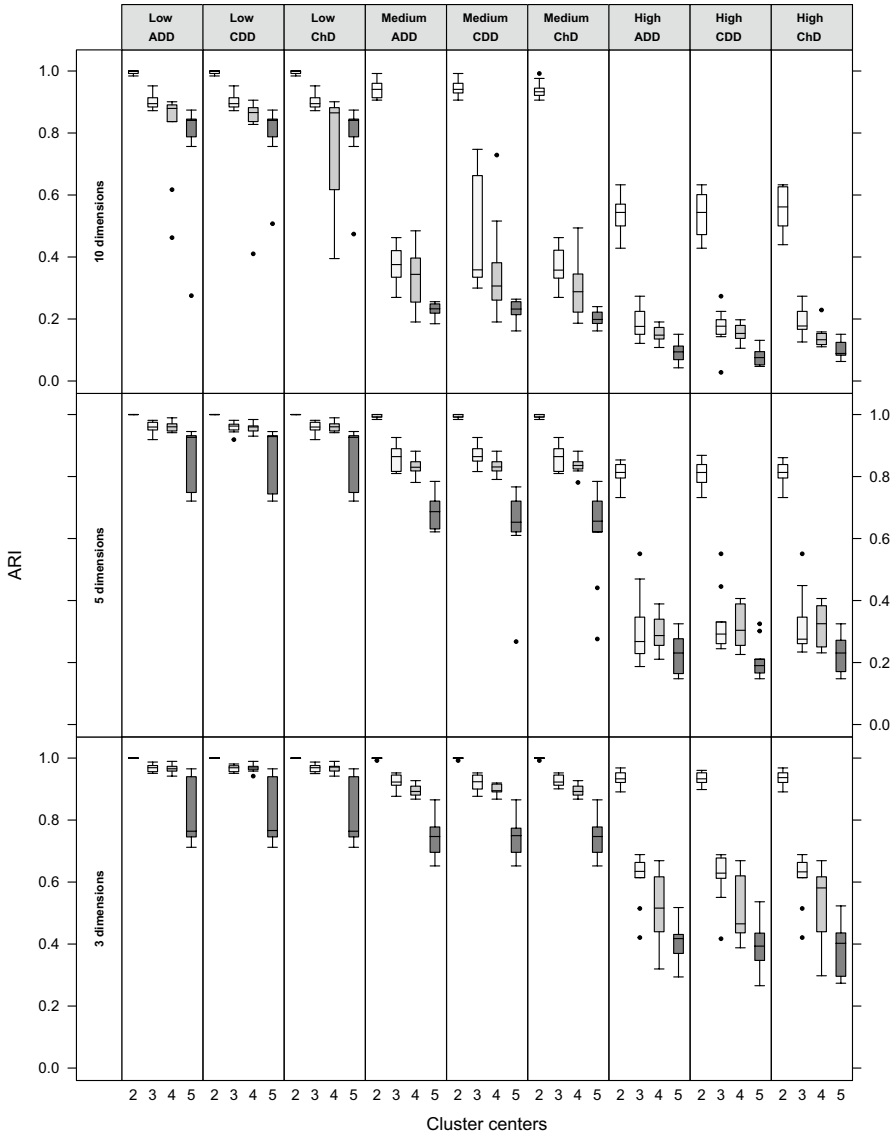

Fig. 3 Structured data. The boxplots report the adjusted Rand index (ARI) values between the true partition and the resulting partition given by DBMCA

In the case structured data and for a low noise level, the ARI values show larger variability for $K = 5$ clusters regardless of the dimension. When $K = 2$ centers are considered, the overall variability of the ARI values is really small, and in general they are approximately equal to 0.8, except for the case of high noise level in ten dimensions where they range from 0.4 to 0.6. As expected, when a high noise is

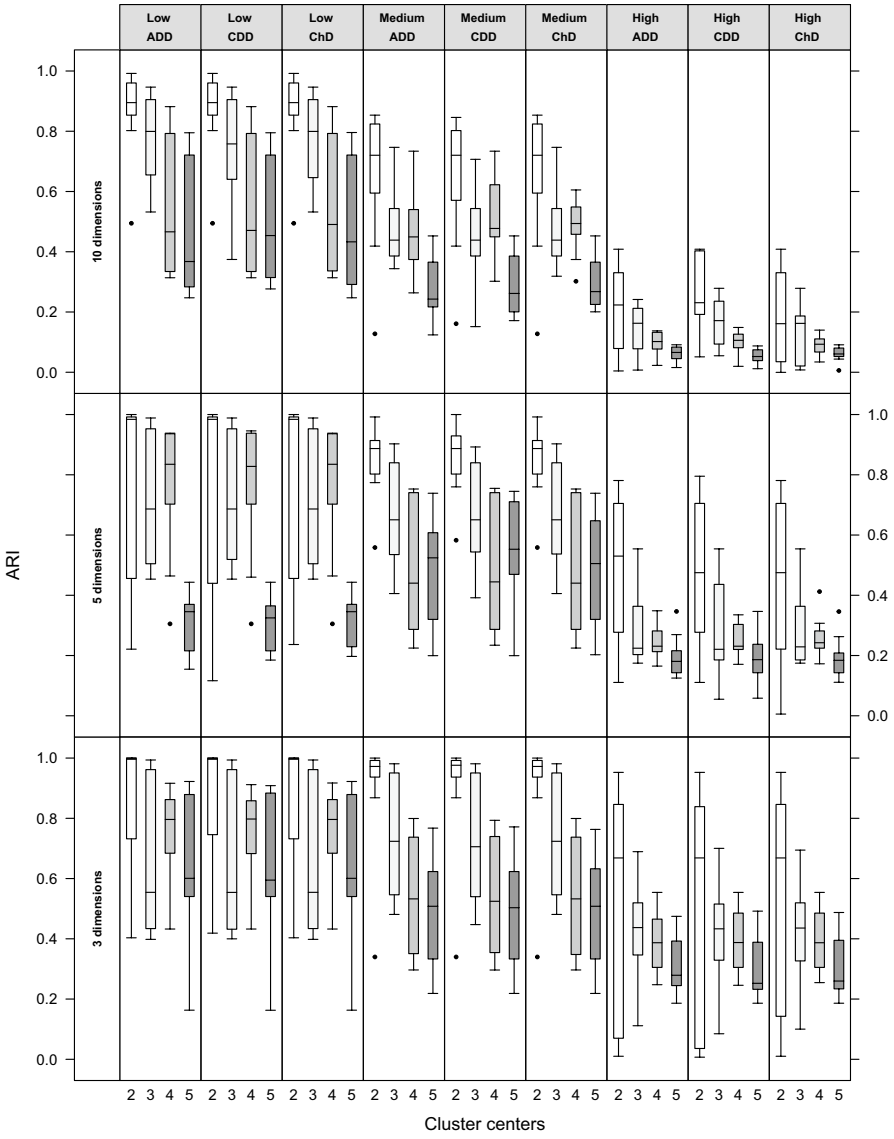

Fig. 4 Unstructured data. The boxplots report the adjusted Rand index (ARI) values between the true partition and the resulting partition given by DBMCA

introduced to obscure the underlying clustering structure to be recovered, the ARI values get generally smaller.

In the case of unstructured data, the ARI values are generally smaller and show much more variability with respect to the case of structured data. In addition, the ARI values get smaller when the number of the theoretical clusters increases. Here again, the ARI values get generally smaller when high noise occurs.

Internal validity values are always large for both structured and unstructured data sets, and thus not reported here.

7 Real data example: text clustering

Directional data can arise in textual data analysis, and text clustering in particular. It has been experimentally shown that it is often helpful to normalize the data vectors to remove the bias arising from the length of a document (Dhillon and Modha 2001b).

Here we show an example on the "res0" data set coming from the internal repository of the CLUTO software for clustering high-dimensional data sets (http://glaros.dtc.umn.edu/gkhome/views/cluto). It contains 1504 documents for which the frequency of a set of 2286 possible words have been recorded.

The idea is modeling the word frequencies as spherical data by normalizing them to 1 according to the L^2 norm. Normalization makes long or short documents comparable and projecting documents onto the unit hypersphere is especially useful for sparse data, as is typically the case with textual data. Indeed, this example shows a data matrix with more than the 89% of zero entries.

For such data set, there exists also an a-priori classification of the documents, consisting of 13 classes of documents as summarized in Table 2.

We applied the proposed procedure to the data set by adopting the cosine distance depth (CDD). We compared our proposal with the spherical K-means algorithm (SKM), and the mixture of von Mises-Fisher, by using both hard (movMFh) and soft (movMFs) partitioning. To determine the "optimal" number of clusters (ranging from 1 to 15), the silhouette approach was adopted for DBMCA and SKM algorithms, while the BIC criterion was used in the case of mixture of von Mises-Fisher distributions. A number of $K = 10$ partitions was selected for DBMCA, while $K = 9$ clusters were identified for spherical K-means. For the "hard" mixture of von Mises-Fisher distributions (movMFh) and its "soft" version (movMFs), $K = 15$ and $K = 14$ clusters were returned, respectively. The spherical K-means and the mixture of von Mises-Fisher distributions algorithms were run through the `R` packages `skmeans` (Hornik et al. 2017) and `movMF` (Hornik and Grün 2014), respectively. The depth-based medoids clustering algorithm (DBMCA) was computed by means of `R` functions written by the authors.

Since the "soft" mixture of von Mises-Fisher distributions assigns membership probabilities of each point to each of the K components (clusters), a fuzzy version

Table 2 Frequency distribution of the a-priori partitions of the Re0 data set

Class	Housing	Money	Trade	Reserves	cpi	Interest	Gnp
Proportion	0.0106	0.4043	0.2121	0.0279	0.0399	0.1456	0.0530
Class	Retail	ipi	Jobs	lei	Bop	wpi	
Proportion	0.0133	0.0246	0.0259	0.0073	0.0253	0.0100	

of the Rand and adjusted Rand index was used, that is the normalized degree of concordance (NDC) which is defined as follows.

Let $\mathbf{W}$ be a probabilistic (fuzzy) partition of the data matrix $\mathbf{X}$, and let $\mathbf{w}(\mathbf{x}_i) = (\mathbf{w}_1(\mathbf{x}_i), \mathbf{w}_2(\mathbf{x}_i), \dots, \mathbf{w}_K(\mathbf{x}_i)) \in [0,1]^K$ be the membership degree of $\mathbf{x}_i$ in the kth cluster. Given any pair $(\mathbf{x}_i, \mathbf{x}_j) \in \mathbf{X}$, Hüllermeier et al. (2012) defined a fuzzy equivalence relation on $\mathbf{X}$ in terms of similarity measure as:

$$E_{\mathbf{W}} = 1 - \|\mathbf{W}(\mathbf{x}_i) - \mathbf{W}(\mathbf{x}_j)\|,$$

where $\|\cdot\|$ is the normalized L_1-norm, yielding a value in [0, 1].

Given two fuzzy partitions, say $\mathbf{G}$ and $\mathbf{H}$, the degree of concordance between a pair of observations is defined as

$$conc(\mathbf{x}_i, \mathbf{x}_j) = 1 - \|E_{\mathbf{g}}(\mathbf{x}_i, \mathbf{x}_j) - E_{\mathbf{h}}(\mathbf{x}_i, \mathbf{x}_j)\| \in [0,1],$$

yielding to a distance measure defined by the normalized sum of *concordant pairs*:

$$d(\mathbf{G}, \mathbf{H}) = \frac{1}{n(n-1)/2} \sum_{i \neq j}^{n} \|E_{\mathbf{g}}(\mathbf{x}_i, \mathbf{x}_j) - E_{\mathbf{h}}(\mathbf{x}_i, \mathbf{x}_j)\|, \tag{5}$$

where n is the sample size. The normalized degree of concordance (NDC) between two fuzzy partitions, or between a fuzzy and a crisp partition, has been defined as (Hüllermeier et al. 2012):

$$NDC(\mathbf{G}, \mathbf{H}) = 1 - d(\mathbf{G}, \mathbf{H}). \tag{6}$$

Note that it reduces to the original Rand Index when partitions $\mathbf{P}$ and $\mathbf{Q}$ are non-fuzzy.

The adjusted version of the NDC "for the chance", called adjusted concordance index, has been defined by D'Ambrosio et al. (2021) as

$$ACI(\mathbf{G}, \mathbf{H}) = \frac{NDC(\mathbf{G}, \mathbf{H}) - \overline{NDC}(\mathbf{G}, \mathbf{H})}{1 - \overline{NDC}(\mathbf{G}, \mathbf{H})}, \tag{7}$$

where $\overline{NDC}(\mathbf{G}, \mathbf{H})$ is the mean value of the NDC$(\mathbf{G}, \mathbf{H})$, which is computed by repeatedly permute one of the two partitions, say $\tilde{\mathbf{H}}$, by keeping fixed the partition $\mathbf{G}$, and then compute NDC as in Eq. (6). For an extensive overview of the fuzzy extension of the adjusted concordance index, we refer to D'Ambrosio et al. (2021). Both NDC and ACI are implemented in the `R` package `ConsRankClass` (D'Ambrosio 2021), that has been used to produce the results in the following of the paper.

The results summarized in Table 3 give the external validation indexes for all of the clustering techniques that were applied to the data set. One can note that the most similar partitions were returned by the mixtures of von Mises-Fisher distributions (ACI = 0.5391). The partitions returned by DBMCA and SKM is quite similar as well (ARI = 0.5123). The adjusted concordance index between the partitions of DBMCA and movMFs is equal to 0.3775, which indicates a low similarity.

Table 3 External validation measures between the partitions given by the four considered clustering techniques applied to the "res0" data set

	DBMCA SKM	DBMCA movMFh	SKM movMFh		DBMCA movMFs	SKM movMFs	movMFh movMFs
ARI	0.5123	0.3918	0.4583	ACI	0.3775	0.4824	0.5391
RI	0.8981	0.8970	0.8974	NDC	0.8783	0.8894	0.9195

Table 4 contains the Rand and adjusted Rand indexes yielded by each method. Globally the employed clustering algorithms show quite similar results. A careful observation allows us to note that the highest ARI value is associated to DBMCA, while movFMh provides the largest RI value.

In addition, Table 5 reports ARI, ACI, RI and NDC between the partitions returned by each pair of algorithms. The Rand index and, consequently, the normalized degree of concordance show always higher values than ARI and ACI. Such results were expectable since RI is not able to take into consideration effects of random groupings.

According to the ARI values, we can notice a moderate "closeness" of the partitions returned by the DBMCA and SKM for $K = 3$, 4, 10, 11 and 12. The largest value of ACI is between DBMCA and movMFs when $K = 5$. On average, the partitions returned by SKM and movMF, both crisp and hard, are quite large.

8 Conclusions

We have proposed a non-parametric procedure (hence not dependent upon any distribution models) for clustering directional data. Specifically, we exploit the concept of angular data depth, which provides a measurement of the centrality of an observation within a distribution of points, to measure the similarity among spherical objects. The method is flexible and applicable even to high dimensional data sets as it is not computational intensive.

We evaluated the performances of the proposed method through an extensive simulation study. Results highlight that, when data are quite structured, it is able to recover the partitions quite well. In case of either extreme overlap (high noise) or

Table 4 External validation measures with the "true" partition for four considered clustering techniques applied to the "res0" data set

	DBMCA	SKM	movMFh		movFMs
ARI	0.2128	0.2034	0.1854	ACI	0.2094
RI	0.7642	0.7568	0.7730	NDC	0.7659

When hard and soft partitions are compared, the Rand index (RI) and the adjusted Rand index (ARI) are replaced with the normalized concordance index (NDC) and the adjusted concordance index (ACI), respectively

Table 5 Adjusted Rand index (ARI) and Rand index (RI) between DBMCA and SKM, between the DBMCA and the mixtures of von Mises-Fisher (crisp clustering, VMFc) and between SKM and VMFh partitioned by the number of clusters K

K	DBMCA-SKM		DBMCA-movMFh		DBMCA-movMFs		SKM-movMFh		SKM-movMFs	
	ARI	RI	ARI	RI	ACI	NDC	ARI	RI	ACI	NDC
1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
2	0.1985	0.5996	0.1432	0.5725	0.1505	0.5761	0.6596	0.8321	0.6753	0.8398
3	0.5437	0.7942	0.3735	0.6840	0.4998	0.7718	0.3826	0.6883	0.7336	0.8797
4	0.5652	0.8308	0.5979	0.8294	0.5077	0.7850	0.5131	0.7977	0.4519	0.7647
5	0.4625	0.8117	0.5170	0.8251	0.5865	0.8404	0.5309	0.8331	0.4344	0.7847
6	0.4673	0.8373	0.5455	0.8355	0.3356	0.7683	0.4501	0.8073	0.5309	0.8423
7	0.4915	0.8559	0.4537	0.8272	0.3936	0.8133	0.4802	0.8445	0.5811	0.8785
8	0.4977	0.8710	0.3814	0.8284	0.4649	0.8482	0.5751	0.8895	0.5537	0.8809
9	0.4645	0.8770	0.3507	0.8440	0.3386	0.8403	0.5744	0.8990	0.5676	0.8969
10	0.5181	0.9067	0.3694	0.8637	0.4116	0.8714	0.5830	0.9088	0.5685	0.9046
11	0.5201	0.9100	0.3419	0.8649	0.3966	0.8801	0.4954	0.8962	0.4842	0.8973
12	0.5304	0.9166	0.3869	0.8868	0.3546	0.8787	0.5050	0.9084	0.5030	0.9064
13	0.4365	0.9064	0.3605	0.8878	0.4030	0.8913	0.5282	0.9212	0.4263	0.9003
14	0.4425	0.9166	0.3760	0.8960	0.3480	0.8833	0.5016	0.9202	0.4593	0.9066
15	0.4448	0.9189	0.3858	0.9090	0.3724	0.9001	0.5849	0.9399	0.5076	0.9233

Normalized degree of concordance (NDC) and adjusted concordance index (ACI) between DBMCA and mixtures of von Mises-Fisher (soft clustering, VMFs) and between SKM and VMFs partitioned by number of clusters K

non-structured data (that is when theoretical partitions are not governed by a clear structure), the recovery of the partitions is not good, as expected.

In addition, we have compared our method with some other clustering techniques available in the literature by means of a real data example in textual data analysis. Here again, results provide strong empirical support for the effectiveness of our approach.

Overall, this work reveals many potential lines of work to consider in the future and more research is needed to further consolidate this interesting framework and explore its broad applications. For instance, it will be interesting to investigate its robustness properties and develop a tool to determine the number of hyperspherical clusters using the depth method.

Acknowledgements GP acknowledges the support of the National Operative Program (PON) Ricerca e Innovazione 2014-2020 (PON R &I) - Azione IV.4 - "Dottorati e contratti di ricerca su tematiche dell'innovazione".

Funding Open access funding provided by Università degli Studi di Napoli Federico II within the CRUI-CARE Agreement.

References

Ackermann H (1997) A note on circular nonparametrical classification. Biom J 39(5):577–587
Agostinelli C, Romanazzi M (2013) Nonparametric analysis of directional data based on data depth. Environ Ecol Stat 20(2):253–270
Arthur D, Vassilvitskii S (2007) *k*-means++: the advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, pp 1027–1035
Banerjee A, Dhillon I, Ghosh J, Sra S (2003) Generative model-based clustering of directional data. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 19–28
Banerjee A, Dhillon IS, Ghosh J, Sra S (2005) Clustering on the unit hypersphere using von Mises-Fisher distributions. J Mach Learn Res 6:1345–1382
Benjamin BMJ, Hussain I, Yang MS (2019) Possiblistic c-means clustering on directional data. In: 2019 12th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI), pp 1–6 https://doi.org/10.1109/CISP-BMEI48845.2019.8965703
Buttarazzi D, Pandolfo G, Porzio GC (2018) A boxplot for circular data. Biometrics 74(4):1492–1501
D'Ambrosio A (2021) ConsRankClass: classification and clustering of preference rankings. R package version 101 https://CRAN.R-project.org/package=ConsRankClass
D'Ambrosio A, Amodio S, Iorio C, Pandolfo G, Siciliano R (2021) Adjusted concordance index: an extension of the adjusted rand index to fuzzy partitions. J Classif 38(1):112–128
Demni H, Porzio GC (2021) Directional DD-classifiers under non-rotational symmetry. In: 2021 IEEE international conference on multisensor fusion and integration for intelligent systems (MFI), pp 1–6 https://doi.org/10.1109/MFI52462.2021.9591189
Demni H, Messaoud A, Porzio GC (2019) The cosine depth distribution classifier for directional data. Applications in statistical computing. Springer, Cham, pp 49–60
Dhillon IS, Modha DS (2001) Concept decompositions for large sparse text data using clustering. Mach Learn 42(1):143–175
Dhillon IS, Modha DS (2001) Concept decompositions for large sparse text data using clustering. Mach Learn 42(1–2):143–175
Dhillon IS, Sra S (2003) Modeling data using directional distributions. Tech. rep, Citeseer
Di Marzio M, Fensore S, Panzera A, Taylor CC (2019) Kernel density classification for spherical data. Stat Probab Lett 144:23–29
Ding Y, Dang X, Peng H, Wilkins D (2007) Robust clustering in high dimensional data using statistical depths. In: BMC bioinformatics, BioMed Central
Fernandes K, Cardoso JS (2016) Discriminative directional classifiers. Neurocomputing 207:141–149
Franke J, Redenbach C, Zhang N (2016) On a mixture model for directional data on the sphere. Scand J Stat 43(1):139–155
Hoberg R (2000) Cluster analysis based on data depth. Data analysis, classification, and related methods. Springer, Cham, pp 17–22
Hornik K, Grün B (2014) movmf: an r package for fitting mixtures of von mises-fisher distributions. J Stat Softw 58(10):1–31. https://doi.org/10.18637/jss.v058.i10
Hornik K, Feinerer I, Kober M, Buchta C (2017) skmeans: spherical k-means clustering. R package version 02-11 https://CRAN.R-project.org/package=skmeans
Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218
Hüllermeier E, Rifqi M, Henzgen S, Senge R (2012) Comparing fuzzy partitions: a generalization of the Rand index and related measures. Fuzzy Syst IEEE Trans 20(3):546–556

Jeong MH, Cai Y, Sullivan CJ, Wang S (2016) Data depth based clustering analysis. In: Proceedings of the 24th ACM SIGSPATIAL international conference on advances in geographic information systems. ACM

Jörnsten R (2004) Clustering and classification based on the L_1 data depth. J Multivar Anal 90(1):67–89

Kaufman L, Rousseeuw P (1987) Clustering by means of medoids. In: Dodge Y (ed) Statistical data analysis based on the L_1-norm and related methods. North-Holland Publishing Co., Amsterdam, pp 405–416

Kesemen O, Tezel Ö, Özkul E (2016) Fuzzy c-means clustering algorithm for directional data (fcm4dd). Expert Syst Appl 58:76–82

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

Ley C, Verdebout T (2018) Applied directional statistics: modern methods and case studies. CRC Press, Florida

Ley C, Sabbah C, Verdebout T et al (2014) A new concept of quantiles for directional data and the angular Mahalanobis depth. Electron J Stat 8(1):795–816

Liu R, Singh K (1992) Ordering directional data: concepts of data depth on circles and spheres. J Am Stat Assoc 20(3):1468–1484

López-Cruz PL, Bielza C, Larranaga P (2015) Directional naive Bayes classifiers. Pattern Anal Appl 18(2):225–246

Mardia KV, Jupp P (2000) Directional statistics, 2nd edn. Wiley, Chichester

Pandolfo G (2022) The GLD-plot: a depth-based graphical tool to investigate unimodality of directional data. J Stat Comput Simul 1–14

Pandolfo G, D'Ambrosio A (2021) Depth-based classification of directional data. Expert Syst Appl 169(114):433. https://doi.org/10.1016/j.eswa.2020.114433

Pandolfo G, Paindaveine D, Porzio GC (2018) Distance-based depths for directional data. Can J Stat 46(4):593–609

Romanazzi M (2009) Data depth, random simplices and multivariate dispersion. Stat Probab Lett 79(12):1473–1479

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65

SenGupta A, Roy S (2005) A simple classification rule for directional data. Advances in ranking and selection, multiple comparisons, and reliability. Springer, Cham, pp 81–90

Taghia J, Ma Z, Leijon A (2014) Bayesian estimation of the von-mises fisher mixture model with variational inference. IEEE Trans Pattern Anal Mach Intell 36(9):1701–1715

Torrente A, Romo J (2021) Initializing k-means clustering by bootstrap and data depth. J Classif 38(2):232–256

Tsagris M, Alenazi A (2019) Comparison of discriminant analysis methods on the sphere. Commun Stat Case Stud Data Anal Appl 5(4):467–491

Yang MS, Pan JA (1997) On fuzzy clustering of directional data. Fuzzy Sets Syst 91(3):319–326