Algorithms for Clustering on the Sphere: Advances & Applications

Mojgan Golzy[1], Marianthi Markatou[2], and Arti Shivram[3]

***Abstract*—Model-based clustering of directional data has been proposed as a basis for clustering by many authors, using mixtures of different distributions that are natural for directional data such as von Mises-Fisher (vMF) distribution, and Watson distribution. However, when vMF and Watson distributions are used as component densities, an approximation of the concentration parameter is used to estimate κ in both cases. We present a clustering method based on mixtures of Poisson kernels on the sphere. The Poisson kernel offers a natural way of clustering data on the surface of a sphere as well as in the ball and half-sphere. We derive estimates of the parameters and describe the corresponding clustering algorithm. We compare the performance of this model with existing methods.**

***Index Terms*—Clustering, kernel method, algorithms, probability models.**

I. Introduction

MANY applications of interest involve data that can be analyzed as unit vectors on a d-dimensional sphere, or equivalently are directional in nature. Directional data arise frequently in many natural and physical sciences such as Biology, Medicine, Ecology, Geology, Material Sciences and others. Specific applications in Medicine include assessment of recovery of orthopaedic patients where the angle of knee flexion was measured [18], the study of hospital emergency room entrance times [7], and the study of circadian rhythms (i.e. study of body temperature fluctuations, sleep-wakefulness cycle etc.). In medical imaging, diffusion magnetic resonance imaging (dMRI) allows one to examine the microscopic diffusion of water molecules in biological tissue in vivo. Water molecules are in constant thermal motion, but this motion is constrained by surrounding structures such as nerves, cells & tissue. Measurements of this diffusion are useful in the study of anisotropic structures like white matter fibers in the central nervous system, and reveal microstructural properties of the underlying tissue. Gene expression, cancer cell data and word counts in a corpus of documents are additional examples of directional data, once normalized to have unit norm.

Manuscript received June 29, 2016; revised July 20, 2016.

[1]M. Golzy is a postdoctoral researcher with the Department of Biostatistics, School of Public Health and Health Professions, University at Buffalo, NY, e-mail:mojgango@buffalo.edu

[2]M. Markatou is the Professor and Associate Chair of Research and Healthcare Informatics, Department of Biostatistics, School of Public Health and Health Professions, University at Buffalo, NY, e-mail: markatou@buffalo.edu

[3]A. Shivram is a postdoctoral researcher with the Department of Biostatistics and the Department of Computer Science & Engineering, University at Buffalo, NY, e-mail: ashivram@buffalo.edu

The authors would like to acknowledge the Department of Biostatistics, SPHHP, for financial support (in the form of start-up package to the second author of the paper).

In ecology, the prevailing wind direction is considered as an important factor in many studies including those that involve pollutant transport, whereas geologists study paleocurrents to infer the direction of the flow of rivers. Currently, new materials, such as polymer, metal foams or fibre-reinforced materials, have found many applications in industry. But extensive use of these may be limited because of the difficulty to quantify their performance (i.e. durability under stress or permeability). Microstructural models based on distributions on the sphere are used to simulate macroscopic properties in model materials to understand the microstructucture relations [15].

Conventional methods suitable for the analysis of linear data cannot be applied for directional data due to its circular nature. The statistical methods that are used to handle such data are given in several references such as [29], [14], [25], and [20]. Due to the non-linearity of the hyper-sphere, clustering on the spherical manifold is often treated in an ad-hoc manner by either ignoring the geometry of the sphere or using overly-restricted models.

Computational methods have been developed and used for clustering directional data. Some commonly used non-parametric approaches are K-means clustering [11], [27], [24], spherical K-means [9], and online spherical K-means [32]. Reference [31] evaluates the performance of different criterion functions in the context of partitional clustering algorithms for document datasets.

Generative (parametric) approaches such as multivariate mixture models provide methods that have distinct advantages over competing non-probabilistic approaches for certain problems. Generative approaches allow uncertainty in cluster membership, and direct control over the variability allowed within each cluster (as captured by the variance characteristics of each component model). A list of references on generative approaches to text clustering can be found in [33]. Reference [3] considered a finite mixture of von Mises-Fisher distributions to cluster text and genomic data. The spherical k-means algorithm, has been shown to be a special case of a generative model based on a mixture of von Mises-Fisher (vMF) distributions with equal priors for the components and equal concentration parameters [4], [2]. A comparative study of some generative models based on the multivariate Bernoulli, multinomial distributions, and the generative model based on a mixture of von Mises-Fisher (vMF) distributions is presented in [34].

References [5] and [28] discuss a generative model of mixtures of Watson distributions on a hypersphere and derive numerical approximations of the parameters in an Expectation Maximization (EM) setting. Each paper presents a different approximation for the estimation of the concentration parameter (κ) in both cases, when vMF and Watson

ISBN: 978-988-14047-1-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

distributions are used.

Reference [10] proposed to use the inverse stereographic projections of multivariate normal distributions. This distribution allows a clustering with various shapes (not just spherical) and orientations, but the Maximum Likelihood (ML) estimate of the mean direction does not have a closed-form expression. The Kent distributions [19] have also this feature and allow different shapes and orientations. Nevertheless, the estimation of their parameters is problematic [10].

We present a clustering method based on mixtures of Poisson kernels on the sphere, with important mathematical and physical interpretations. The Poisson kernel is a density with respect to uniform measure and offers a natural way of assessing goodness of fit, a strength that is not present in existing algorithms. We derive estimates of the parameters of the mixture of Poisson kernels in an Expectation Maximization setting, and describe the corresponding clustering algorithm. We compare the performance of this model with existing methods.

II. Existing Literature

Clustering approaches can be categorize as either generative (also known as parametric or probabilistic) or discriminative (non-parametric). The performance of an approach, and of a specific method within that approach, is quite data dependent; there is no clustering method that works "best" across all types of data. Generative models, however, often provide greater insight into the anatomy of the clusters.

A. Discriminative (non-parametric) Approaches

K-means clustering [11] is an iterative algorithm. Given a set of N observations $\mathcal{X}$, where each observation is a d-dimensional real vector, k-means clustering partitions the N observations into $M(\leq N)$ sets $\mathcal{X}_1, \cdots, \mathcal{X}_M$ by minimizing the within-cluster sum of squares.

Spherical K-means [9], uses cosine similarity instead of Euclidean distance, that measures the cosine of the angle formed by two vectors. In other words, the objective function is

$$Q(\{\mathcal{X}_k^{(t)}\}_{j=1}^{M}) = \sum_{k=1}^{M} \sum_{\mathbf{x} \in \mathcal{X}_k} \mathbf{x}^T \boldsymbol{\mu}_k, \tag{1}$$

where $\boldsymbol{\mu}_k$ is the concept vector (the normalized mean vector) corresponding to the partition $\mathcal{X}_k$. We seek a solution that maximizes the objective function in (1). Spherical K-means algorithm is preferred to standard K-means for clustering of the document vectors or any type of high-dimensional observations on the unit sphere. However, it is sensitive to initialization and outliers.

Online Spherical K-means [32] is an extension of skmeans which has a competitive learning nature; as a data point is processed, centroids are updated correspondingly by

$$\boldsymbol{\mu}_{k(x)}^{(new)} = \frac{\boldsymbol{\mu}_{k(x)} + \eta \mathbf{x}}{\left\| \boldsymbol{\mu}_{k(x)} + \eta \mathbf{x} \right\|}, \tag{2}$$

where η is the learning rate. It is less sensitive to initialization. Spherical k-means clustering can be performed, using the package skmeans in R software, by using the function skmeans [16].

B. Probabilistic (parametric) Approaches

The parametric mixture model approach to clustering, assumes each cluster is generated by its own density function that is unknown. The overall data is modeled as a mixture of individual cluster density functions. In reality, the unknown densities may not be from the same family of distributions. In this section we consider mixture models in which the densities are from the same family of distributions.

For $d \geq 2$, let S^{d-1} be the unit sphere. The probability density function of a mixture with M components on the hypersphere, S^{d-1}, is given by

$$f(\mathbf{x}|\Theta) = \sum_{j=1}^{M} \alpha_j f_j(\mathbf{x}|\boldsymbol{\theta}_j), \tag{3}$$

where M is the number of clusters, α_j's are the mixture proportions that are non-negative and sum to one and $\Theta = (\alpha_1, \cdots, \alpha_M, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_M)$.

Some commonly used mixture models are:

1) Mixtures of Von-Mises-Fisher Distributions [3]: Given $\boldsymbol{\mu} \in S^{d-1}$, and $\kappa \geq 0$, the probability distribution function of von Mises-Fisher Distribution (vMF) is defined by

$$f(\mathbf{x}|\boldsymbol{\mu}, \kappa) = c_d(\kappa) e^{\kappa \mu^T \mathbf{x}}, \tag{4}$$

where $\boldsymbol{\mu}$ is a vector orienting the center of the distribution, κ is a parameter to control the concentration of the distribution around the vector $\boldsymbol{\mu}$. The normalizing constant $c_d(\kappa)$ is given by

$$c_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}, \tag{5}$$

where $I_r(.)$ represents the modified Bessel function of the first kind of order r. The von Mises distribution is unimodal and symmetric about μ. Banerjee et al. performed Expectation Maximization (EM) [8], [6] for a finite vMF mixture model to cluster text and genomic data.

The advantage of using the class of vMF distributions in the mixture model is that it incorporates many density shapes. The disadvantage is that numerical estimation of the concentration parameter involves functional inversion of the ratios of Bessel functions. Thus, it is not possible to directly estimate the κ values in high dimensional data and an asymptotic approximation of κ is used for estimating κ. The package movMF in R software can be used for fitting a mixture of vMF distribution [17].

2) Mixture of Watson distributions [5], [28]: Given $\boldsymbol{\mu} \in S^{d-1}$, and κ, the probability function of the Watson distribution is defined by:

$$f(\mathbf{x}|\boldsymbol{\mu}, \kappa) = M(1/2, d/2, \kappa)^{-1} e^{\kappa(\boldsymbol{\mu}^T \mathbf{x})^2}, \tag{6}$$

where $M(1/2, d/2, \kappa)$ is the confluent hyper-geometric function also known as Kummer function.

The advantage of using the class of Watson distributions in the mixture model is that it shows superior performance for noisy, thinly spread clusters over the von Mises-Fisher distributions [5]. The disadvantage

ISBN: 978-988-14047-1-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

is that in high-dimensions, maximum likelihood equations pose severe numerical challenges. Also, numerical estimation of the concentration parameter involves a ratio of Kummer functions. Thus, similar to vMF, it is not possible to directly estimate the κ values and in both papers an asymptotic approximation of κ is used for estimating κ.

3) Mixture of inverse stereographic projection of multivariate normal distribution [10]. The density function, which is denoted by $\mathcal{L}_{\mu,\Sigma}$, is the stereographic projection of $N_{d-1}(\mathbf{0},\Sigma)$ on the plane of dimension $d-1$ perpendicular to $\boldsymbol{\mu}$. It allows clustering with various shapes and orientations.
The advantage of using the class of inverse stereographic projection of multivariate normal distribution in the mixture model is that it allows a clustering with various shapes and orientations. The disadvantage is that, there is no closed expression for $\boldsymbol{\mu}_{MLE}$. In practice, it is obtained via a heuristic search algorithm.

III. Proposed clustering Method

A. Introduction to Poisson Kernel

For $d \geq 2$, let S^{d-1} be the unit sphere and $\mathbb{B}^d$ be the unit Ball in $\mathbb{R}^d$. The Poisson kernel is defined by

$$P_d(\mathbf{x},\boldsymbol{\zeta}) = \frac{1-\|\mathbf{x}\|^2}{\omega_d\|\mathbf{x}-\boldsymbol{\zeta}\|^d}, \tag{7}$$

for $(\mathbf{x},\boldsymbol{\zeta}) \in \mathbb{B}^d \times S^{d-1}$ $\omega_d = 2\pi^{d/2}\{\Gamma(d/2)\}^{-1}$ is the surface area of the unit sphere in $\mathbb{R}^d$.

Some properties of Poisson Kernel are [1]:

- $P_d(\mathbf{x},\boldsymbol{\zeta}) > 0$ for all $\mathbf{x} \in \mathbb{B}^d$ and all $\boldsymbol{\zeta} \in S^{d-1}$;
- let σ be the normalized surface-area measure on S^{d-1} (so that $\sigma(S^{d-1}) = 1$) then, for all $\mathbf{x} \in \mathbb{B}^d$

$$\int_{S^{d-1}} P_d(\mathbf{x},\boldsymbol{\zeta}) d\sigma(\boldsymbol{\zeta}) = 1; \tag{8}$$

- for every $\boldsymbol{\eta} \in S^{d-1}$ and every $\delta > 0$

$$\int_{|\boldsymbol{\zeta}-\boldsymbol{\eta}|>\delta} P_d(\mathbf{x},\boldsymbol{\zeta}) d\sigma(\boldsymbol{\zeta}) \to 0 \text{ as } \mathbf{x} \to \boldsymbol{\eta}. \tag{9}$$

Definition: For $d \geq 2$, a d-dimensional unit random vector $\mathbf{x}$ is said to have a *d-variate Poisson kernel distribution* on S^{d-1} if its density is given by

$$f(\mathbf{x}|\rho,\boldsymbol{\mu}) = \frac{1-\rho^2}{\omega_d\|\mathbf{x}-\rho\boldsymbol{\mu}\|^d}, \tag{10}$$

where $\|\boldsymbol{\mu}\| = 1$, $0 < \rho < 1$ and $\omega_d = 2\pi^{d/2}\{\Gamma(d/2)\}^{-1}$.

It can be written as

$$f(\mathbf{x}|\rho,\boldsymbol{\mu}) = \frac{1-\rho^2}{\omega_d(1+\rho^2-2\rho\ \mathbf{x}.\boldsymbol{\mu})^{d/2}}, \tag{11}$$

where $\mathbf{x}.\mathbf{y}$ indicate the inner product of $\mathbf{x}$ and $\mathbf{y}$.

The Poisson kernel density is unimodal and symmetric around $\mathbf{x} = \boldsymbol{\mu}$. We note that $\mathbf{x}.\mathbf{y} = \cos(\alpha)$ where α is the angle between the vectors $\mathbf{x}$ and $\mathbf{y}$ and so

$$\frac{1-\rho}{\omega_d(1+\rho)^{d-1}} < f(\mathbf{x}|\rho,\boldsymbol{\mu}) < \frac{1+\rho}{\omega_d(1-\rho)^{d-1}}. \tag{12}$$

Therefore, if $\rho \to 0$ then $f(\mathbf{x}|\rho,\boldsymbol{\mu}) \to 1/\omega_d$ which is the uniform density on S^{d-1} and if $\rho \to 1$, $f(\mathbf{x}|\rho,\boldsymbol{\mu})$ converges to the point density.

B. Proposed Model

Let $\mathcal{X}$ be a set of sample unit vectors drawn independently from mixtures of Poisson kernel distributions with parameter space $\Theta = (\alpha_1,\cdots,\alpha_M,\rho_1,\cdots,\rho_M,\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_M)$ where M is the number of clusters, and α_j's are the mixture proportions that are non-negative and sum to one. We assume the following probabilistic model:

$$f(\mathbf{x}|\Theta) = \sum_{j=1}^{M} \alpha_j f_j(\mathbf{x}|\rho_j,\boldsymbol{\mu}_j), \tag{13}$$

C. Estimation of Parameters

Let $\mathcal{X}$ be a data set of N independently sampled points that follows (13) and $\mathcal{Z}$ be the corresponding set of hidden random variables that indicates the particular distribution from which these points are sampled. The expectation of the likelihood over the given posterior distribution p can be written as

$$\sum_{j=1}^{M}\sum_{i=1}^{N} ln(\alpha_j)p(j|\mathbf{x_i},\Theta) + \sum_{j=1}^{M}\sum_{i=1}^{N} ln(f_j(\mathbf{x_i}|\rho_j,\boldsymbol{\mu}_j)p(j|\mathbf{x_i},\Theta). \tag{14}$$

Following the standard EM algorithm, we obtain

$$\alpha_j = 1/N \sum_{i=1}^{N} p(j|\mathbf{x_i},\Theta). \tag{15}$$

The distribution of the hidden variables is given by

$$p(j|\mathbf{x_i},\Theta) = \frac{\alpha_j f_j(\mathbf{x_i}|\rho_j,\boldsymbol{\mu}_j)}{\sum_{l=1}^{M}\alpha_l f_l(\mathbf{x_i}|\rho_l,\boldsymbol{\mu}_l)}. \tag{16}$$

For details on EM algorithm and estimation of α_k's we refer to [8] and [6].
The Lagrangian for the second term of (14) is given by

$$\sum_{j=1}^{M}\sum_{i=1}^{N}\{ln(1-\rho_j^2) - ln(\omega_d) - d\ ln\|\mathbf{x_i}-\rho_j\boldsymbol{\mu_j}\|\}\times p(j|\mathbf{x_i},\Theta) + \sum_{j=1}^{M}\lambda_j(1-\|\boldsymbol{\mu}_j\|^2). \tag{17}$$

To obtain the estimates of the parameters we maximize the above expression subject to $0 < \rho_j < 1$ for each j.

The estimates of the parameters $\boldsymbol{\mu_k}$ and ρ_k can be obtained using the following iterative re-weighted algorithm; Let $\hat{\Theta}^{(0)} = \{\hat{\alpha}_1^{(0)},\cdots,\hat{\alpha}_M^{(0)},\hat{\rho}_1^{(0)},\cdots,\hat{\rho}_M^{(0)},\hat{\boldsymbol{\mu}}_1^{(0)},\cdots,\hat{\boldsymbol{\mu}}_M^{(0)}\}$ be the initial values of the parameters, then we define $\hat{\alpha_k}^{(t+1)}, \hat{w}_{ik}^{(t+1)}, \hat{\boldsymbol{\mu}}_k^{(t+1)}$ and $\hat{\rho}_k^{(t+1)}$ for $t = 1,\cdots$ iteratively as follows;

$$p(j|\mathbf{x_i},\hat{\Theta}^{(t)}) = \frac{\alpha_j^{(t)} f_j(\mathbf{x_i}|\hat{\rho}_j^{(t)},\hat{\boldsymbol{\mu}}_j^{(t)})}{\sum_{l=1}^{M}\alpha_l^{(t)} f_l(\mathbf{x_i}|\hat{\rho}_l^{(t)},\hat{\boldsymbol{\mu}}_l^{(t)})}, \tag{18}$$

$$\hat{\alpha_k}^{(t+1)} = 1/N \sum_{i=1}^{N} p(k|\mathbf{x_i},\hat{\Theta}^{(t)}), \tag{19}$$

$$\hat{w}_{ik}^{(t+1)} = \frac{p(k|\mathbf{x_i},\hat{\Theta}^{(t)})}{\|\mathbf{x_i}-\hat{\rho_k}^{(t)}\hat{\boldsymbol{\mu}_k}^{(t)}\|^2}, \tag{20}$$

$$\hat{\boldsymbol{\mu}}_k^{(t+1)} = s_k * \frac{\sum_{i=1}^{n}\hat{w}_{ik}^{(t)}\mathbf{x_i}}{\|\sum_{i=1}^{n}\hat{w}_{ik}^{(t)}\mathbf{x_i}\|}, \tag{21}$$

ISBN: 978-988-14047-1-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

where $s_k = \text{sign}(\hat{\boldsymbol{\mu}}_k^{(t+1)}.\hat{\boldsymbol{\mu}}_k^{(t)})$.
And $\hat{\rho}_k^{(t+1)}$ is the solution to the equation

$$\frac{-2nx\hat{\alpha_k}^{(t)}}{1-x^2} + d\|\sum_{i=1}^{n}\hat{w}_{ik}^{(t)}\mathbf{x_i}\| - dx\sum_{i=1}^{n}\hat{w}_{ik}^{(t)} = 0, \quad (22)$$

subject to the constraint $0 < \rho < 1$.

The proposed model based on the Poisson kernel (and hence our clustering method) has several advantages. Those are as follows: (a) the estimation of the parameters of the kernel does not require any approximations; (b) an expression for the Poisson kernel of an upper half-sphere (or lower half-sphere) can be obtained by certain Möbius transformations. This can be used to create a clustering method on the half-sphere. (c) The Poisson kernel itself has inferential capacity, in that it can be used to develop goodness of fit procedures for testing model appropriateness. These can be constructed following [22] and [23].

IV. Experimental Results

To obtain an understanding of the performance of the proposed clustering algorithm we conducted a small simulation study and evaluated our model on real world data. We benchmark our performance against competing state of art clustering methods and report our results.

The goal of our limited simulation is to assess the performance of the models in an environment that is favorable to the state-of-the-art ([3]). Performance is measured by macro-precision and macro-recall that are computed as functions of the overlap between the component distributions. We simulated 100 data sets as follows. For each sample, we generated a random unit vector in S^{d-1} from a mixture of three von Mises-Fisher (movMF) distributions with equal mixing proportions and random mean center. We used a common and fixed concentration parameter κ, in each case, as given in the tables. As κ increases the overlap of the points decreases, providing a better separation of the different component densities.

Figures 1-4 show 3D plots of 4 simulated data sets in S^2 based on mixtures of three component von Mises-Fisher (vMF) distributions that illustrate the overlap of the different components for different values of κ. The data were generated from the vMF distribution class to favor the state-of-the-art algorithm of [3].

For each sample, we used three clustering methods; 1) The Spherical K-means 2) a mixture of three von Mises-Fisher distribution and 3) a mixture of three Poisson kernel-based distributions (proposed model).

The approach given in [12], was used for the initialization of the parameters. Stopping rule for the iteration in our algorithm was 100 runs. We compared the performance of each algorithms using the macro-precision and macro-recall [26]. Suppose $\omega_1, \cdots, \omega_c$ are the true classification classes. For a given clustering, let a_t denote the number of data objects that are correctly assigned to the class ω_t, b_t denote the data objects that are incorrectly assigned to the class ω_t, and c_t denote the data objects that are incorrectly rejected from the class ω_t. The precision and recall are defined as $p_t = \frac{a_t}{a_t+b_t}$ and $r_t = \frac{a_t}{a_t+c_t}$ for $1 \leq t \leq c$. The macro-precision, and macro-recall, are the averages across classes of the precisions and recalls.

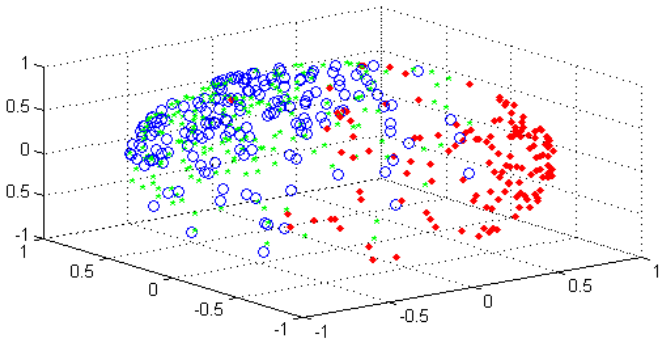

Fig. 1. Mixtures of vMF: κ=4

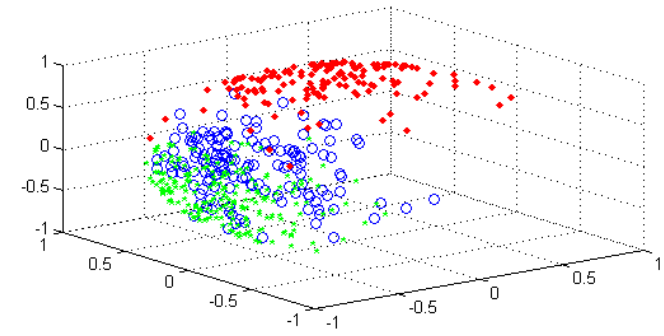

Fig. 2. Mixtures of vMF: κ=10

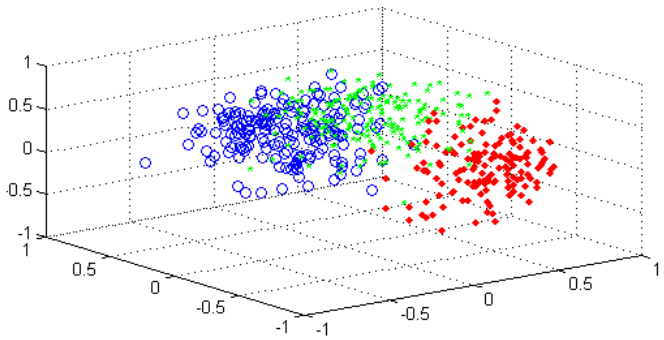

Fig. 3. Mixtures of vMF: κ=15

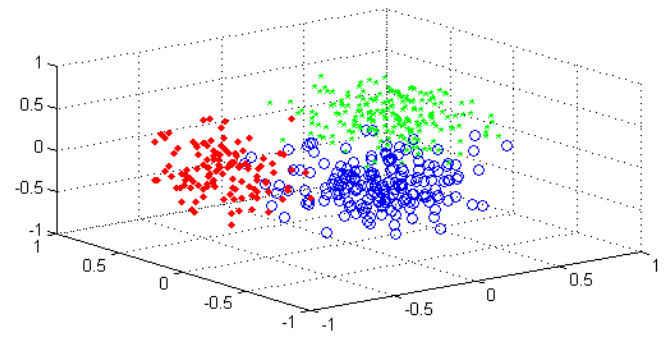

Fig. 4. Mixtures of vMF: κ=20

A. Simulations results

The statistical software R was used for all analyses. Spherical k-means clustering was performed by using the function skmeans in R software [16]. Mixture of vMF clustering was

ISBN: 978-988-14047-1-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

performed by using the function movMF in R software [17], and using the approximation given in [3] for estimation of the concentration parameters.

Tables I & II present mean macro-precision/recall and their associated standard deviations for sample sizes 500 and $5,000$ and dimensions 50 and 100 respectively. The mixture of vMF algorithm is expected to outperform the other methods. Our results indicate that the Poisson kernel based method, for large κ, provides slightly better macro-precision and macro-recall. However, when the standard deviation is accounted for the Poisson kernel based method performs equivalently with the state-of-the-art, and all methods show approximately the same performance.

Table I: Mean and standard deviation of the macro-precision & macro-recall of three clustering algorithms, when $N = 500$, $d = 50$ and number of clusters $= 3$.

κ	Eval.	spkmeans	mix-vMF	mix-PKBD
4	m-p	**0.298** (0.07)	0.295 (0.07)	0.295 (0.07)
	m-r	**0.377** (0.02)	0.372 (0.02)	0.371 (0.02)
10	m-p	0.617 (0.09)	0.630 (0.10)	**0.641** (0.08)
	m-r	**0.627** (0.064)	0.624(0.064)	0.626 (0.061)
15	m-p	**0.848** (0.024)	0.843 (0.063)	**0.848** (0.047)
	m-r	**0.848** (0.024)	0.843(0.043)	0.846 (0.035)
20	m-p	0.9369 (0.017)	0.9378 (0.017)	**0.9383** (0.017)
	m-r	0.9368 (0.017)	0.9368 (0.018)	**0.9374** (0.017)

Table II: Mean and standard deviation of the macro-precision & macro-recall of three clustering algorithms, when $N = 5,000$, $d = 100$ and number of clusters$= 3$.

κ	Eval.	spkmeans	mix-vMF	mix-PKBD
4	m-p	0.285 (0.059)	0.281 (0.064)	**0.29** (0.062)
	m-r	**0.355** (0.008)	0.354 (0.009)	0.353 (0.008)
10	m-p	0.583 (0.024)	**0.59** (0.046)	0.588 (0.053)
	m-r	0.583 (0.02)	**0.588** (0.03)	**0.588** (0.03)
20	m-p	0.8560 (0.011)	0.8563 (0.011)	**0.8564** (0.011)
	m-r	0.856 (0.011)	0.856 (0.011)	**0.8562** (0.011)

B. Illustrative examples

Here we compare our method against well-established in the literature methods, using real data sets. The data points are projected onto the sphere by normalizing them so the associated vectors have length one. In what follows we briefly describe the data sets.

1. The household data set was obtained from "HSAUR2" in R software [13], see "https://cran.r-project.org/web/packages/HSAUR2/HSAUR2.pdf". The data is part of a data set collected from a survey on household expenditures and gives the expenses of 20 single men and 20 single women on four commodity groups (housing, food, goods and services). Hornik et al. [17] focused only on three of those commodity groups (housing, food and service) to obtain 3-dimensional data for easier visualization. We will focus on all four commodity groups. The scale of measurement of the data is interval.
2. The Wisconsin Breast Cancer Database was obtained from UC Irvine Machine Learning Repository "https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)". The original breast cancer databases were obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg [30]. The objective is to identify each of a number of benign or malignant classes. There are 16 missing attribute values which we removed from the data set. The data frame has 699 observations on 11 variables, one being a character variable, 9 being categorical (1 through 10), and 1 target class. The variables are: Id, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei,Bland Chromatin, Normal Nucleoli, Mitoses and Class Source.
3. The Landsat Multi-Spectral Scanner Image Data (satellite data set) from package "mlbench" in R software [21]. The database consists of the multi-spectral values of pixels in 3x3 neighbourhoods in a satellite image, and the classification associated with the central pixel in each neighbourhood. The aim is to predict this classification, given the multi-spectral values. The database is a (tiny) sub-area of a scene, consisting of 82 x 100 pixels. Each line of data corresponds to a 3x3 square neighbourhood of pixels completely contained within the 82x100 subarea. Each line contains the pixel values in the four spectral bands (converted to ASCII) of each of the 9 pixels in the 3x3 neighbourhood and a number indicating the classification label of the central pixel. The data has 6435 rows and 37 columns (x1-x36 continuous variables and class). The classes are; red soil, cotton crop, grey soil, damp grey soil, soil with vegetation stubble, and very damp grey soil.

Table III: The macro-precision and macro-recall of three clustering algorithms for various data sets.

Data Set	Eval.	spkmeans	mix-vMF	mix-PKBD
Household	m-p	0.847	0.870	**0.926**
n =40		(97.35%)	(100%)	(106.43%)
d =4	m-r	0.825	0.825	**0.925**
M =2		(100%)	(100%)	(112.12%)
Cancer	m-p	0.694	**0.734**	0.723
n =683		(94.55%)	(100%)	(98.5%)
d =9	m-r	0.704	0.710	**0.712**
M =3		(99.29%)	(100%)	(100.42%)
Satellite	m-p	**0.609**	0.579	0.606
n =6435		(105.18%)	(100%)	(104.83%)
d =36	m-r	0.533	**0.558**	0.543
M =6		(95.51%)	(100%)	(97.31%)

The approach given in [12], was used for the initialization of the parameters. Stopping rule for the iteration in our algorithm was 100 runs. We notice that if we specify different seeds in the movMF algorithm, we obtain different results. Similar observations hold when we carried out our simulations indicating potential instability of the movMF algorithm. The best results were obtained when no seed was used. Table III present the results in terms of macro-precision and macro-recall of the different algorithms. Note that 100% performance in either macro-precision or macro-recall indicates the performance set for the state-of-the-art algorithm. The percentages, listed in Table III, are obtained by dividing the current algorithm's macro-precision/recall with the movMF macro-precision/recall and multiplying by 100. The Poisson kernel based model outperforms all other models in the case of "Household" data and either is the top performer or performs equivalently to the top performer in all other cases.

ISBN: 978-988-14047-1-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

V. Discussion & Conclusion

In this paper, we presented an algorithm for clustering data on the sphere that is based on the Poisson kernel. Our results indicate that the method performs equivalently to the state of the art and outperforms the state of the art for certain data structures. The method has advantages in that no approximation is needed to estimate the parameters; further an expression of the Poisson kernel for the upper (lower) half-sphere can be obtained and used to allow clustering of data that reside in these spaces. Future work will incorporate further characterization of the data structures that can be expected to obtain superior results when Poisson kernel based clustering is used, an extensive study of the role of initialization on the performance of the algorithm, and determination of the optimal number of clusters. Additionally, we will study the conditions for identifiability of the mixture of Poisson kernels and we will explore the inferential properties of the kernel for use in the context of clustering.

References

[1] S. Axler, P. Bourdon, and W. Ramey, *Harmonic Function Theory*, Springer-Verlag New York, Inc. 2nd edition. 2001

[2] A. Banerjee, I. S. Dhillon, J. Ghosh and S. Sra, "Clustering on hyperspheres using Expectation Maximization," *Technical Report TR -03-07. Department of Computer Sciences, University of Texas.* 2003

[3] A. Banerjee, I. S. Dhillon, J. Ghosh and S. Sra, "Clustering on the unit hypersphere using von Mises-Fisher distributions," *Journal of Machine Learning Research*, 6:13451382, 2005

[4] A. Banerjee, J. Ghosh, "Frequency sensitive competitive learning for clustering on high-dimensional hyperspheres," *Proc. IEEE Int. Joint Conf. Neural Networks* pp. 15901595, 2002

[5] A. Bijral, M. Breitenbach, G. Z. Grudic, "Mixture of Watson distributions: A generative model for hyperspherical embeddings," *In: Artificial Intelligence and Statistics* (AISTATS). pp. 3542. 2007

[6] J. Bilmes, "A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models", *Technical Report ICSI-TR-97-021, University of California, Berkeley*, 1997

[7] D. R. Cox, P. A. Lewis, *The Statistical Analysis of Series of Events*, Methuen's Monographs on Applied Probability and Statistics, John Wiley, London, 1966

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, Series B, 39:138, 1977

[9] I. S. Dhillon, and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Machine Learning*, 42(1):143175, 2001

[10] J. Dortet-Bernadet, N. Wicker, " Model-based clustering on the unit sphere with an illustration using gene expression profiles," *Biostatistics*, 9(1): 66-80, 2008

[11] R. O. Duda, and P. E. Hart, *Pattern classification and scene analysis*, Wiley, 1973

[12] R. Duwairi, M. Abu-Rahmeh, "A novel approach for initializing the spherical K-means clustering algorithm," *Simulation Modelling Practice and Theory,* 54, 4963, 2015

[13] B. S. Everitt, T. Hothorn, "HSAUR2: *A Handbook of Statistical Analyses Using R (2nd Edition)*, R package version 1.1-6,2013 URL http://CRAN.R-project.org/package=HSAUR2.

[14] I. Fisher. *Statistical Analysis of Circular Data*, Cambridge University Press, 1996

[15] J. Franke, C. Redenbach, and N. Zhang, "On a mixture model for directional data on the sphere," *Scandinavian Journal of Statistics*, 43,139-155, 2016

[16] K. Hornik, I. Feinerer, M. Kober, and C. Buchta. "Sperical k-means clustering," *Journal of Statistical Software,* 50(10):1-22,2012 http://doi.org/10.18637/jss.v050.i10.

[17] K. Hornik, B. Grün, "movMF: An R package for fitting mixtures of von Mises-Fisher distributions," *Journal of Statistical Software,* 58(10), 1-31. 2014 https://www.jstatsoft.org/article/view/v058i10

[18] S. R. Jammalamadaka and Y. R. Sarma, "A correlation coefficient for angular variables," *In Statistical Theory and Data Analysis II*, 349-364, North Holland, Amsterdam. 1988

[19] J. T. Kent, "The Fisher-Bingham distribution on the sphere," *J. Royal. Stat. Soc., Series B*, 44(1), 71-80, 1982

[20] A. Lee, "Circular data," *Wiely Interdisciplinary Review: Computational Statistics* 2 477-486, 2010

[21] F. Leisch and E. Dimitriadou, "Package 'mlbench'," R Package version 2.1-1, 2015 URL https://cran.r-project.org/web/packages/mlbench/mlbench.pdf.

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

[23] B. G. Lindsay, M. Markatou, and S. Ray, "Kernels, degrees of freedom, and power properties of quadratic distance goodness-of-fit tests," *J Am Stat Assoc.*, 109:505, 395-410, 2014

[24] R. Maitra and I. P. Ramler, "A k-mean-directions algorithm for fast clustering of data on the sphere," *Journal of Computational and Graphical Statistics*, 19(2) 377-396, 2010 DOI: 10.1198/jcgs.2009.08155

[25] K.V. Mardia, P. Jupp, *Directional Statistics*, New York, John Wiley and Sons Ltd., 2nd edition. 2000

[26] D. S. Modha, and W. S. Spangler, "Feature weighting in k-Means clustering," *Machine Learning*, 52(3),217237, 2003 DOI: 10.1023/A:1024016609528

[27] I. P. Ramler, "Improved statistical methods for k-means clustering of noisy and directional data," *Graduate Theses and Dissertations*. Paper 10949. 2008

[28] S. Sra, D. Karp, " The multivariate Watson distribution: Maximum-Likelihood estimation and other aspects," *J. of Multivariate Analysis* 114, 256269, 2013 doi:10.1016/j.jmva.2012.08.010

[29] G. S. Watson, *Statistics on sphere*, John Wiley & Sons, 1983

[30] W. H. Wolberg, O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," *In Proceedings of the National Academy of Sciences*, 87, 9193–9196, 1990

[31] Y. Zhao and G. Karypis. "Empirical and theoretical comparisons of selected criterion functions for document clustering," *Machine Learning*, 55(3):311331, June 2004

[32] S. Zhong, "Efficient online Spherical K-means Clustering," *Proceedings of International Joint Conference on Neural Networks*, Montreal, Canada, 2005

[33] S. Zhong and J. Ghosh. "A unified framework for model-based clustering," *Journal of Machine Learning Research*, 4:10011037, November 2003a

[34] S. Zhong and J. Ghosh. "A comparative study of generative models for document clustering," *In Workshop on Clustering High Dimensional Data : Third SIAM Conference on Data Mining*, April 2003b

ISBN: 978-988-14047-1-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)