Clustering directions based on the estimation of a mixture of von Mises-Fisher distributions

Adelaide Figueiredo

Faculty of Economics of University of Porto and LIAAD-INESC TEC

Abstract

Background:

In the statistical analysis of directional data, the von Mises-Fisher distribution plays an important role to model unit vectors. The estimation of the parameters of a mixture of von Mises-Fisher distributions can be done through the Estimation-Maximization algorithm.

Objective:

In this paper we propose a dynamic clusters type algorithm based on the estimation of the parameters of a mixture of von Mises-Fisher distributions for clustering directions, and we compare this algorithm with the Estimation-Maximization algorithm. We also define the between-groups and within-groups variability measures to compare the solutions obtained with the algorithms through these measures.

Results:

The comparison of the clusters obtained with both algorithms is provided for a simulation study based on samples generated from a mixture of two Fisher distributions and for an illustrative example with spherical data.

Keywords: *Directional data, Dynamic Clusters algorithm, EM algorithm, von Mises--Fisher distribution.*

AMS 2000 subject classification: 62H11, 62H30.

1 Introduction

Clustering data in the unit sphere is an important task in modern data analysis, for example, in clustering text documents when analysing textual data.

One approach to address such issue is the spherical k-means clustering. This technique was proposed by Dhillon and Modha (2001) and implemented in a R package, called skmeans by Hornik *et al.* (2012), and it is based on the cosine similarity to obtain a partition of term weight representation of the documents.

Other works that have appeared in the literature for clustering directional data are based on model-based clustering methods. For instance, Peel *et al.* (2001) used the Kent distribution (Kent, 1982) to form groups of fracture data through a model-based clustering and Dortet-Bernadet and Wicker (2008) supposed a model-based clustering of data that lies on a unit sphere and applied this clustering method to gene expression profiles. Banerjee *et al.* (2005) applied a model-based clustering of directional data to text analysis. These authors considered the estimation of a mixture of von Mises-Fisher distributions using two variants of the Estimation-Maximisation EM algorithm, denoted by soft-movMF and hard-movMF algorithms. Another variant of the EM algorithm, denoted by stochastic EM was given by Celeux and Govaert (1992). Banerjee *et al.* (2005) showed that the spherical k-means algorithm may be obtained as a variant of the EM algorithm for the maximum likelihood estimation of the mean direction parameters of a mixture of von Mises-Fisher distributions with common concentration parameter κ, using hard-max classification E-step.

Figueiredo and Gomes (2015) proposed an algorithm based on the dynamic clusters algorithm proposed by Diday and Schroeder (1976) for the estimation of the parameters of a mixture of Watson distributions defined on the hypersphere and compared it with the EM algorithm, proposed by Dempster *et al.* (1977) for problems of incomplete data. Similarly in this paper, to estimate the parameters of a mixture of k von Mises-Fisher distributions and obtain a partition of the sample into clusters, we propose a dynamic clusters type algorithm and we compare it with the EM algorithm. This proposed algorithm has the advantage of converging quickly to a local optimum, while the EM algorithm may converge slowly to the local optimum. On the other hand, the EM algorithm provides strongly consistent estimators with asymptotic normal distribution (Redner and Walker, 1984). For comparing the solutions obtained in both algorithms, we define between-groups and within-groups variability measures. Then, for several generated samples and a real data set we compare the solutions obtained with these algorithms.

In Section 2 we recall the von Mises-Fisher distribution and the maximum likelihood estimators of the parameters of this distribution. In Section 3 we describe the EM algorithm and we propose the dynamic clusters type algorithm for the estimation of a mixture of k von Mises-Fisher distributions. In Section 4 we define the variability measures and we compare the algorithms through these measures, using simulated data from von Mises--Fisher populations and a real data set. In Section 5 we present some concluding remarks.

2 von Mises-Fisher distribution

The von Mises-Fisher distribution is one of the most used distributions in the statistical analysis of directional data. It is usually denoted by $M_p\left(\boldsymbol{\mu},\kappa\right)$ and has probability density function defined by

$$f(\mathbf{x}|\boldsymbol{\mu},\kappa)=c_p(\kappa)\,\exp(\kappa\boldsymbol{\mu}^T\mathbf{x})\quad \mathbf{x}\in S_{p-1},\quad \boldsymbol{\mu}\in S_{p-1},\ \kappa>0\ , \tag{1}$$

where the normalising constant is given by $c_p(\kappa)=\kappa^{\frac{p}{2}-1}\big/[(2\pi)^{p/2}\,I_{p/2-1}(\kappa)]$, and $I_\nu(.)$ denotes the modified Bessel function of the first kind and order ν and S_{p-1} denotes the unit sphere in $\mathbb{R}^p$. This distribution is called von Mises distribution for circular data and Fisher distribution for spherical data. The parameter $\boldsymbol{\mu}$ is the vector of the mean direction and κ is the concentration parameter around $\boldsymbol{\mu}$. This distribution is rotationally symmetric about $\boldsymbol{\mu}$.

Let $(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n)$ be a random sample of size n from the von Mises-Fisher distribution, $M_p\left(\boldsymbol{\mu},\kappa\right)$. Let $\overline{R}$ be the resultant length mean of the sample defined by $\overline{R}=\left(\overline{\mathbf{x}}^T\overline{\mathbf{x}}\right)^{1/2}=\|\overline{\mathbf{x}}\|$, where $\overline{\mathbf{x}}$ is the sample mean vector of $(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n)$ defined by $\overline{\mathbf{x}}=\sum_{i=1}^n\mathbf{x}_i/n$. The maximum likelihood estimator of $\boldsymbol{\mu}$ is the sample mean direction, i.e.,

$$\widehat{\boldsymbol{\mu}}=\overline{\mathbf{x}}_0=\frac{\overline{\mathbf{x}}}{\left(\overline{\mathbf{x}}^T\overline{\mathbf{x}}\right)^{1/2}}=\frac{\overline{\mathbf{x}}}{\|\overline{\mathbf{x}}\|}$$

and the maximum likelihood estimator of κ is the solution of the equation

$$A_p(\kappa)=\|\overline{\mathbf{x}}\|,$$

where the function $A_p(\kappa)$ is defined by $A_p(\kappa)=c_p'(\kappa)/c_p(\kappa)=I_{p/2}(\kappa)/I_{p/2-1}(\kappa)$.

For more details about this distribution, see for instance, Mardia and Jupp (2000, p. 198).

3 Estimation of a mixture of k von Mises-Fisher distributions

A mixture of k von Mises-Fisher components C_1,...,C_k has probability density function given by

$$\psi(\mathbf{x}|Q)=\sum_{j=1}^{k}\pi_j f(\mathbf{x}|\boldsymbol{\theta}_j)\quad \mathbf{x}\in S_{p-1}, \tag{2}$$

where $\boldsymbol{\theta}_j=(\boldsymbol{\mu}_j,\kappa_j)$, $\boldsymbol{\mu}_j\in S_{p-1},\kappa_j>0$, and $f(\mathbf{x}|\boldsymbol{\theta}_j)$ is the density function of C_j component, i.e., the density of $M_p(\boldsymbol{\mu}_j,\kappa_j)$ distribution. The parameters $\pi_j, j=1,...,k$ with $0<\pi_j<1$ and $\sum_{j=1}^k\pi_j=1$ are the proportions of the mixture and $Q=(\boldsymbol{\nu},\boldsymbol{\theta})$, with $\boldsymbol{\nu}=(\pi_1,....,\pi_k)$

and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k)$ is the vector of unknown parameters of the mixture.
For the estimation of the parameters of the mixture, we review the *EM* algorithm and its variants (soft-movMF, hard-movMF and stochastic *EM*) in Subsection 3.1 and we propose a dynamic clusters type algorithm in Subsection 3.2.

3.1 *E*M algorithm

The Estimation-Maximization (*EM*) algorithm is used to obtain the maximum likelihood estimates of the parameters of the mixture and can be briefly described as follows.
Let $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ be a random sample from the mixture and let $Z = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n)$ be the missing data, where the indicator vector $\mathbf{z}_i = (Z_{i1}, Z_{i2}, ..., Z_{ik})^T$, with $Z_{ij} = \begin{cases} 1 & if\ \mathbf{x}_i \in C_j \\ 0 & if\ \mathbf{x}_i \notin C_j \end{cases}$, $\sum_{j=1}^{k} Z_{ij} = 1$ indicates the component of the mixture for $\mathbf{x}_i$. The expected log-likelihood associated with the complete sample $(\mathbf{x}_1, ..., \mathbf{x}_n, Z)$, derived in Appendix A, is given by

$$L(Q|\mathbf{x}_1, .., \mathbf{x}_n, Z) = \sum_{i=1}^{n}\sum_{j=1}^{k} t_j(\mathbf{x}_i) \ln[\pi_j f(\mathbf{x}_i|\boldsymbol{\mu}_j, \kappa_j)], \tag{3}$$

where $t_j(\mathbf{x}_i)$ is the a-posteriori probability of $\mathbf{x}_i$ belonging to C_j defined by $t_j(\mathbf{x}_i) = \pi_j f(\mathbf{x}_i|\boldsymbol{\theta}_j) \diagup [\sum_{h=1}^{k} \pi_h f(x_i|\boldsymbol{\theta}_h)],\ j = 1, ..., k.$ So (3) may be written as

$$L(Q|\mathbf{x}_1, .., \mathbf{x}_n, Z) = \sum_{i=1}^{n}\sum_{j=1}^{k} t_j(\mathbf{x}_i) \ln \pi_j + \sum_{i=1}^{n}\sum_{j=1}^{k} t_j(\mathbf{x}_i)[\ln c_p(\kappa_j) + \kappa_j \boldsymbol{\mu}_j^T \mathbf{x}_i]. \tag{4}$$

Let

$$L_1(Q|\mathbf{x}_1, .., \mathbf{x}_n, Z) = \sum_{i=1}^{n}\sum_{j=1}^{k} t_j(\mathbf{x}_i)[\ln c_p(\kappa_j) + \kappa_j \boldsymbol{\mu}_j^T \mathbf{x}_i] \tag{5}$$

and let

$$L_2(Q|\mathbf{x}_1, .., \mathbf{x}_n, Z) = \sum_{i=1}^{n}\sum_{j=1}^{k} t_j(\mathbf{x}_i) \ln \pi_j. \tag{6}$$

To estimate the vector of unknown parameters Q, the *EM* algorithm uses iteratively the two steps: Estimation (E) and Maximization (M).
The algorithm starts with an initial solution: $Q^0 = (\pi_1^0, ..., \pi_k^0, \boldsymbol{\mu}_1^0, \kappa_1^0, ..., \boldsymbol{\mu}_k^0, \kappa_k^0)$ or with an initial partition into k groups, and then determine the estimates Q^0 based on the partition.
In the mth iteration ($m \geq 1$) the steps are:

E-Step

For *j=1,...,k, i=1,...,n,* calculate the a-posteriori probability of $\mathbf{x}_i$ belonging to the jth component of the mixture

$$t_j^{(m)}(\mathbf{x}_i) = \frac{\pi_j^{(m)} f(\mathbf{x}_i|\boldsymbol{\mu}_j^{(m)}, \kappa_j^{(m)})}{\sum\limits_{h=1}^{k} \pi_h^{(m)} f(\mathbf{x}_i|\boldsymbol{\mu}_h^{(m)}, \kappa_h^{(m)})}. \tag{7}$$

M-Step

Use estimates $t_j^{(m)}(\mathbf{x}_i)$ to maximize $L_1(Q|\mathbf{x}_1,...,\mathbf{x}_n,Z)$ subject to the constraint $\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j = 1$ and $L_2(Q|\mathbf{x}_1,..,\mathbf{x}_n,Z)$ subject to the constraint $\sum_{j=1}^k \pi_j = 1$. The estimators obtained, derived in Appendix B, are the following:

- The maximum likelihood estimator of $\boldsymbol{\mu}_j$ in the $(m+1)$th iteration, $\widehat{\boldsymbol{\mu}}_j^{(m+1)}$ is given by

$$\widehat{\boldsymbol{\mu}}_j^{(m+1)} = \frac{\sum\limits_{i=1}^n t_j^{(m)}(\mathbf{x}_i)\mathbf{x}_i}{\left\| \sum\limits_{i=1}^n t_j^{(m)}(\mathbf{x}_i)\mathbf{x}_i \right\|}, \qquad j=1,...,k. \tag{8}$$

- The maximum likelihood estimator of κ_j in the $(m+1)$th iteration, $\widehat{\kappa}_j^{(m+1)}$ is the solution of the equation

$$A\left(\widehat{\kappa}_j^{(m+1)}\right) = \frac{R_j}{\sum\limits_{i=1}^n t_j^{(m)}(\mathbf{x}_i)}, \qquad j=1,...,k, \tag{9}$$

  where R_j is the length of the vector $\sum_{i=1}^n t_j^{(m)}(\mathbf{x}_i)\mathbf{x}_i$, that is $R_j = \| \sum_{i=1}^n t_j^{(m)}(\mathbf{x}_i)\mathbf{x}_i\|$.

- The maximum likelihood estimator of π_j in the $(m+1)$th iteration, $\widehat{\pi}_j^{(m+1)}$ is given by

$$\widehat{\pi}_j^{(m+1)} = \frac{1}{n}\sum_{i=1}^n t_j^{(m)}(\mathbf{x}_i), \qquad j=1,...,k. \tag{10}$$

In the particular case of components with the same concentration parameter κ, the estimates of $\boldsymbol{\mu}_j$ and π_j, $j = 1,...,k$, are given by the expressions (8) and (10) and the estimate of the common concentration parameter κ, derived in Appendix C, is the solution of the equation

$$A(\kappa) = \frac{\sum\limits_{j=1}^k R_j}{n}, \tag{11}$$

where R_j is defined as before.

The EM algorithm is assumed to have converged if the relative change in the log--likelihood values is smaller than a threshold or if the relative absolute change in the parameters is smaller than a threshold. A partition $(P_1,...,P_k)$ of the sample is obtained assigning $\mathbf{x}_i$ to the component for which the a-posteriori probability is the largest, that is,

$$P_j = \left\{ \mathbf{x}_i : t_j(\mathbf{x}_i) = \max_h t_h(\mathbf{x}_i), \quad h = 1,...,k \right\} \tag{12}$$

and when $t_j(\mathbf{x}_i) = t_h(\mathbf{x}_i)$ consider $\mathbf{x}_i \in P_j$, if $j < h$.
This algorithm is denoted by soft-movMF algorithm by Banerjee *et al.* (2005, p. 1357).

These authors also proposed the hard-movMF algorithm (p. 1358), which is a modification of the soft-movMF by adding a hardening step (H-step) between E-step and M-step. This step is:

H-Step

Replace the a-posteriori probabilities by assigning each observation with probability 1 to the component for which its a-posteriori probability is maximum.

Celeux and Govaert (1992) denoted the previous algorithm by Classification EM algorithm and proposed another variant of the EM algorithm, the stochastic EM, where instead of the hardening step, a stochastic step (S-step) is added between E-step and M-step. This step is:

S-Step

Assign at random each observation to one component with probability equal to its a--posteriori probability.

These three variants of the EM algorithm are implemented in a R package called movMF (see Hornik and Grun, 2014).

3.2 Dynamic Clusters type algorithm

Let E be a finite sample. The aim is to determine a partition $P = (P_1, P_2, ..., P_k)$ of E into k classes, so that for every j $(1 \leq j \leq k)$, P_j may be considered as a sample from a population with density $f_{\boldsymbol{\theta}}$.

Let $f_{\boldsymbol{\theta}}$ $(\boldsymbol{\theta} \in \mathbb{L})$ be the family of probability densities, from which the distributions of the different components belong: $\boldsymbol{\theta}$ is a vectorial parameter and $\mathbb{L}$ its definition space:

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = c_p(\kappa) \exp(\kappa \boldsymbol{\mu}^T \mathbf{x}), \quad \mathbf{x} \in S_{p-1}, \quad \boldsymbol{\theta} = (\boldsymbol{\mu}, \kappa). \tag{13}$$

Let $\mathbb{P}_k$ be the set of partitions of E into k classes and let $\mathbb{L}_k$ be the set of vectors of dimension k of $\mathbb{L}$. The method starts with an initial partition $(P_1^0, P_2^0, ..., P_k^0)$ of E or starts with a vector of dimension k of values of the unknown parameter $\left(\boldsymbol{\theta}_1^0, ..., \boldsymbol{\theta}_k^0\right)$.

The two following functions f and g are successively applied until obtaining stable elements of L and P:

$$f : \mathbb{L}_k \rightarrow \mathbb{P}_k$$

$$L \rightarrow P$$

where $L = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k)$ and $P = (P_1, ..., P_k)$, so that for $\forall_i$ $1 \leq i \leq k$, P_i is the set of observations, which are less distant from the distribution $f_{\boldsymbol{\theta}_i}$ than from others. Then, it is important to define a function D, which measures the distance from an observation $\mathbf{x} \in E$ to a distribution $f_{\boldsymbol{\theta}}$:

$$D : E \times L \rightarrow \mathbb{R}^+$$

$$(\mathbf{x}, \boldsymbol{\theta}) \to D(\mathbf{x}, \boldsymbol{\theta}).$$

The distance is defined by

$$D(\mathbf{x}, \boldsymbol{\theta}) = \ln\left[\frac{C}{f_\theta(\mathbf{x})}\right], \tag{14}$$

where C is a constant defined by $C \geq \max\{f_{\boldsymbol{\theta}}(\mathbf{x}) |\ \boldsymbol{\theta} \in \mathbb{L},\ \mathbf{x} \in E\}$. Then

$$D(\mathbf{x}, \boldsymbol{\theta}) = C - \ln c_p(\kappa) - \kappa \boldsymbol{\mu}^T \mathbf{x} \tag{15}$$

and each group P_i is defined by

$$\begin{aligned} P_i &= \{\mathbf{x} \in E | D(\mathbf{x}, \boldsymbol{\theta}_i) \leq D(\mathbf{x}, \boldsymbol{\theta}_j),\ \forall_{j \neq i} \text{ with } i < j \text{ if } D(\mathbf{x}, \boldsymbol{\theta}_i) = D(\mathbf{x}, \boldsymbol{\theta}_j)\} \\ &= \left\{\mathbf{x} \in E | \ln c_p(\kappa_i) - \kappa_i \boldsymbol{\mu}_i^T \mathbf{x} \geq \ln c_p(\kappa_j) - \kappa_j \boldsymbol{\mu}_j^T \mathbf{x},\ \forall_{j \neq i}\right\}. \end{aligned}$$

The function

$$g : \mathbb{P}_k \to \mathbb{L}_k$$

$$P \to L$$

is such that for $\forall_i$ $(1 \leq i \leq k)$, $\boldsymbol{\theta}_i$ satisfies the condition

$$\sum_{\mathbf{x} \in P_i} D(\mathbf{x}, \boldsymbol{\theta}_i) = \inf_{\boldsymbol{\theta} \in L} \sum_{\mathbf{x} \in P_i} D(\mathbf{x}, \boldsymbol{\theta}). \tag{16}$$

The optimum value of $\boldsymbol{\theta}_i$ is the maximum likelihood estimator of $\boldsymbol{\theta}_i$ associated with P_i and the optimum criterion is function of the partition P^* and $L^* \in \mathbb{L}$ obtained in convergence:

$$\begin{aligned} W(L^*, P^*) &= \sum_{1 \leq i \leq k} D(P_i^*, \boldsymbol{\theta}_i^*) = \sum_{1 \leq i \leq k} \sum_{\mathbf{x} \in P_i^*} D(\mathbf{x}, \boldsymbol{\theta}_i^*) \\ &= C + \sum_{1 \leq i \leq k} \sum_{\mathbf{x} \in P_i} \left[-\ln c_p(\widehat{\kappa}_i) - \widehat{\kappa}_i \widehat{\boldsymbol{\mu}}_i^T \mathbf{x}\right], \end{aligned}$$

where C is the constant previously defined and $\widehat{\boldsymbol{\mu}}_i$ and $\widehat{\kappa}_i$ are the maximum likelihood estimators of $\boldsymbol{\mu}_i$ and κ_i respectively, based on the sample P_i. Then,

$$W(L^*, P^*) = C - \sum_{1 \leq i \leq k} \left[card(P_i) \ln c_p(\widehat{\kappa}_i) + \sum_{\mathbf{x} \in P_i} \widehat{\kappa}_i \widehat{\boldsymbol{\mu}}_i^T \mathbf{x}\right], \tag{17}$$

where $card(P_i)$ is the number of observations of P_i. So the parameters $\boldsymbol{\theta}_i$ are estimated based on the k classes P_i: the function g defines the estimation by the maximum likelihood method and the function f enables us to define again k new classes P_i and then, evaluate again the value of the criterion W.

4 Comparison of the algorithms

For comparing the solutions obtained with the algorithms, we define next the between--groups and within-groups variability measures, in the decomposition of the total variability used to test the null hypothesis of a common mean vector across k von Mises-Fisher populations with concentration parameters not necessarily equal. This test was considered in the literature for the particular case of equal concentration parameters, for the circle or the sphere, see for instance, Mardia and Jupp (2000, pp. 222-226), Watson (1956), Watson and Williams (1956) and Harrison *et al.* (1986).

Let $\mathbf{x}_{i1}, ..., \mathbf{x}_{in_i}$ $(i = 1, ..., k)$ be k independent random samples of sizes $n_1, ..., n_k$ from populations $M_p\left(\boldsymbol{\mu}_i, \kappa_i\right)$, with mean vector $\boldsymbol{\mu}_i$ and concentration parameter κ_i. Let $n = n_1 + ... + n_k$ be the global sample size. The null hypothesis of interest is

$$H_0 : \boldsymbol{\mu}_1 = ... = \boldsymbol{\mu}_k = \boldsymbol{\mu},$$

against the alternative hypothesis that at least one of the equalities is not satisfied.
Next we consider the concentration parameters κ_i unknown, but if these parameters are unknown, we have to estimate them through their maximum likelihood estimates for instance. Let's consider the following identity

$$2\kappa_i(1 - \boldsymbol{\mu}^T\mathbf{x}_{ij}) = 2\kappa_i(1 - \boldsymbol{\mu}_i^T\mathbf{x}_{ij}) + 2\kappa_i(\boldsymbol{\mu}_i^T\mathbf{x}_{ij} - \boldsymbol{\mu}^T\mathbf{x}_{ij}).$$

Summing from $i = 1$ to k, $j = 1$ to n_i and replacing $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_i$ by their maximum likelihood estimates, the following identity is obtained

$$2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\kappa_i(1 - \widehat{\boldsymbol{\mu}}^T\mathbf{x}_{ij}) = 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\kappa_i(1 - \widehat{\boldsymbol{\mu}}_i^T\mathbf{x}_{ij}) + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\kappa_i(\widehat{\boldsymbol{\mu}}_i^T\mathbf{x}_{ij} - \widehat{\boldsymbol{\mu}}^T\mathbf{x}_{ij}), \quad (18)$$

where $\widehat{\boldsymbol{\mu}}$=$(\sum_i \kappa_i \sum_j \mathbf{x}_{ij})\big/\| \sum_i \kappa_i \sum_j \mathbf{x}_{ij}\|$ and $\widehat{\boldsymbol{\mu}}_i$=$\sum_j \mathbf{x}_{ij}\big/\| \sum_j \mathbf{x}_{ij}\|$=$\sum_j \mathbf{x}_{ij}\big/R_i$, $i = 1, ..., k$. The previous identity can be written as

$$\begin{aligned} 2\left(\sum_{i=1}^{k}\kappa_i n_i - \left\|\sum_{i=1}^{k}\kappa_i\sum_{j=1}^{n_i}\mathbf{x}_{ij}\right\|\right) &= 2\left(\sum_{i=1}^{k}\kappa_i n_i - \sum_{i=1}^{k}\kappa_i\left\|\sum_{j=1}^{n_i}\mathbf{x}_{ij}\right\|\right) + \\ &+2\left(\sum_{i=1}^{k}\kappa_i\left\|\sum_{j}\mathbf{x}_{ij}\right\| - \left\|\sum_{i=1}^{k}\kappa_i\sum_{j=1}^{n_i}\mathbf{x}_{ij}\right\|\right) \end{aligned}$$

or equivalently,

$$2\left(\sum_{i=1}^{k}\kappa_i n_i - R\right) = 2\sum_{i=1}^{k}\left(\kappa_i n_i - \kappa_i R_i\right) + 2\left(\sum_{i=1}^{k}\kappa_i R_i - R\right), \quad (19)$$

where $R = \| \sum_{i=1}^{k} \kappa_i \sum_{j=1}^{n_i} \mathbf{x}_{ij}\|$ and R_i is the resultant length of the ith sample. This identity is the decomposition of total variability $\sum_{i=1}^{k} \kappa_i n_i - R$ into within-groups variability

$\sum_{i=1}^{k}(\kappa_i n_i - \kappa_i R_i)$ and between-groups variability $\sum_{i=1}^{k}\kappa_i R_i - R$. The test statistic is defined by

$$F = \frac{\left(\sum\limits_{i=1}^{k}\kappa_i R_i - R\right)\big/(k-1)(p-1)}{\sum\limits_{i=1}^{k}(\kappa_i n_i - \kappa_i R_i)\big/(n-k)(p-1)} \tag{20}$$

and the hypothesis H_0 is rejected for large values of F. When all concentration parameters κ_i are equal to κ, the statistic (20) reduces to the following statistic given in Mardia and Jupp (2000, pp. 222-223):

$$F = \frac{\left(\sum\limits_{i=1}^{k} R_i - R\right)\big/(k-1)(p-1)}{\left(n - \sum\limits_{i=1}^{k} R_i\right)\big/(n-k)(p-1)}, \tag{21}$$

where R_i is resultant length of the ith sample and R is the resultant length of the global sample. The F-statistic has under H_0 approximately the $F_{(k-1)(p-1),(n-k)(p-1)}$ distribution for large κ.

4.1 Simulation study

We generated samples of size n from a mixture of equal proportions of two Fisher distributions $F(\mathbf{e}_3, \kappa)$ and $F(\boldsymbol{\mu}, \kappa)$, with a common concentration parameter κ. We considered without loss of generality, $\mathbf{e}_3 = (0,0,1)^T$ and $\boldsymbol{\mu} = (0, (1-\cos\theta)^{1/2}, \cos\theta)^T$, where θ is the angle between $\boldsymbol{\mu}$ and $\mathbf{e}_3$. Should other mean directions have been used, which form an angle θ, the same results would have been obtained. We considered two sample sizes $n = 20, 40$, several angles of separation between the two components, $\theta = 30^o, 90^o, 150^o$ and two values of the common concentration parameter $\kappa = 5, 10$.

For generating observations from the Fisher distribution, we used the method given in Wood (1994). We supposed that the parameters of the mixture are unknown and we estimated these parameters based on each generated sample, using the three variants of the EM algorithm (soft-movMF, hard-movMF and stochastic EM) described in Subsection 3.1 and the dynamic clusters type algorithm described in Subsection 3.2. We obtained the estimates of the concentration parameters and the angle between the estimated mean directions, indicated in the Table 1 for the sample size of 20 and concentration parameters of the components equal to 5 or 10 and in Table 2 for the sample size of 40 and concentration parameters of the components equal to 5. In these tables, we also present for each sample, the classification results (confusion matrix), the sizes of the groups and within-groups and between-groups variability measures for the final solution, which were obtained by the expressions given in the previous subsection, where the concentration parameters were replaced by their maximum likelihood estimates. We note that when the angle $\theta = 30^o$, i.e,

Table 1: Confusion matrices, size groups, estimates of the parameters, variability measures (between-groups and within-groups) and F-statistic for the EM algorithm (soft-movMF, hard-movMF, stochastic EM) and dynamic clusters algorithm (DC), for the sample size of 20 and concentrations equal to 5 or 10 (*: the results for the other three methods are equal)

κ	θ (^{o})	Algorithm	Group	Conf. matrix		n_i	$\widehat{\kappa}_i$	$\widehat{\theta}$	Bet./	F
				1	2				With.	
5	30	Soft-movMF	1	7	3	14	15.5	46.4	23.8/	23.3
			2	7	3	6	18.1		18.5	
		Hard-movMF	1	7	3	14	16.6	48.1	26.2/	17.1
			2	7	3	6	20.3		20.0	
		DC	1	6	4	13	17.3	46.4	22.2/	17.1
			2	7	3	7	18.8		23.3	
	90	Soft-movMF	1	6	4	6	14.4	77.5	32.0/	30.6
			2	0	10	14	5.4		18.8	
		Hard-movMF	1	8	2	8	9.2	77.9	31.1/	28.0
			2	0	10	12	7.3		20.0	
		Stochastic EM	1	6	4	6	16.2	78.9	34.3/	30.9
			2	0	10	14	5.6		20.0	
		DC	1	9	1	12	6.1	75.4	27.6/	24.9
			2	3	7	8	10.0		20.0	
	150	Soft-movMF*	1	10	0	10	5.9	127.3	70.3/	63.3
			2	0	10	10	12.5		20.0	
10	30	Soft-movMF	1	4	6	4	77.5	42.3	34.3/	31.8
			2	0	10	16	14.2		19.4	
		Hard-movMF	1	8	2	13	12.8	33.9	15.6/	14.1
			2	5	5	7	32.6		20.0	
		DC	1	7	3	11	14.4	33.4	13.5/	12.2
			2	4	6	9	20.9		20.0	
	90	Soft-movMF*	1	10	0	10	10.9	90.0	72.9/	65.6
			2	0	10	10	18.2		20.0	
	150	Soft-movMF*	1	10	0	10	10.9	142.4	163.3/	147.0
			2	0	10	10	20.4		20.0	

the components are poorly separated, the stochastic EM algorithm did not converge for any run. When the components are reasonably or well-separated, i.e., $\theta = 90^o$ or $\theta = 150^o$, all algorithms lead to the same solution, in general.

Table 2: Confusion matrices, size groups, estimates of the parameters, variability measures (between-groups and within-groups) and F-statistic for the EM algorithm (soft-movMF, hard-movMF, stochastic EM) and dynamic clusters algorithm (DC), for the sample size of 40 and concentrations equal to 5 (*: the results for the other three methods are equal)

θ (^{o})	Algorithm	Group	Conf. matrix		n_i	$\widehat{\kappa}_i$	$\widehat{\theta}$	Bet./	F
			1	2				With.	
30	Soft-movMF	1	17	3	28	4.4	43.1	20.4/	23.3
		2	11	9	12	9.4		33.2	
	Hard-movMF	1	11	9	20	5.8	56.4	28.8/	27.4
		2	9	11	20	9.4		40.0	
	DC	1	13	7	20	9.0	53.9	24.2/	23.0
		2	7	13	20	5.3		40.0	
90	Soft-movMF*	1	20	0	20	6.1	110.4	89.0/	84.6
		2	0	20	20	6.3		40.0	
150	Soft-movMF*	1	20	0	20	5.9	77.9	126.7/	121.7
		2	0	20	20	4.5		39.6	

From the results indicated in Tables 1-2, we conclude the following:

- The algorithms gave the same solution when the two components are well separated, that is, when θ =90^o for κ=5 and when θ =90^o or 150^o for κ=10. Therefore, in these cases, the confusion matrix is the same and the variability measures coincide for all the algorithms, as well as the estimates of the concentration parameters and the estimate of the angle between the mean directions.

- When the concentration of the components increases, the rate of misclassified observations decreases (or remains equal) and the between-groups variability increases, as well the F-statistic for components with moderate or large separation.

- For each algorithm, the rate of misclassified observations decreases as the separation between the components increases, and for well-separated components, this error rate is equal to 0. This error rate decreases or is equal to 0 when the sample size increases.

- For each algorithm, the between-groups variability increases as the separation between the two components increases and for well separated components, the between-groups variability exceeds largely the within-groups variability. The F-statistic also increases when the angle between the mean directions of the components increases.

- When the sample size increases, the between-groups variability and F-statistic increase for moderate or large separation of the components of the mixture.

4.2 Example

We used the spherical data given in Wood (1982), which consist of a set of 33 estimates of a previous magnetic pole position of the earth obtained using palaeomagnetic techniques. Each estimate is associated with a different site, the 33 sites being spread over a large of Tasmania. As the data appear to fall into two main groups, Wood (1982) estimated the parameters of a bimodal model for the data.

We obtained a partition of these data into two groups based on the estimation of a mixture of two Fisher distributions through the three variants of the EM algorithm (soft-movMF, hard-movMF and stochastic EM) described in Subsection 3.1 and dynamic clusters type algorithm described in Subsection 3.2. For obtaining the final solutions of the variants of the EM algorithm, we used the R package, movMF. For the dynamic clusters type algorithm, as it depends on the initial solution, we considered several initial partitions randomly chosen for the algorithm and for all initial partitions, the algorithm converged and the final solution obtained was the same. The final solutions obtained with the algorithms are given in the Table 3.

Table 3: Final partitions, size groups, estimates of the concentration parameters and estimate of the angle between the mean directions

Algorithm	Group	Final Partition	n_i	$\widehat{\kappa}_i$	$\widehat{\theta}$ (^{o})
Soft-movMF	1	9,10,11,12,14,15,16,23,24,30	10	21.50	33.3
	2	Remaining observations	23	36.65	
Hard-movMF	1	9,14,16,24,30	5	22.52	39.9
	2	Remaining observations	28	26.86	
Stochastic	1	9,10,11,12,14,15,16,22,23,24,30	11	20.52	31.2
EM	2	Remaining observations	22	36.47	
Dynamic	1	1,5,9,10,11,12,13,14,15,16,23,24,29,30	14	13.29	31.3
Clusters	2	Remaining observations	19	23.64	

The solutions obtained with the several algorithms do not coincide, probably because in this case the components are not well-separated (the estimated angle between the mean directions is around 30^o). But, the solutions obtained with soft-movMF, stochastic EM

and dynamic clusters algorithm are rather similar, as we may observe for these solutions, a large number of observations is stable in the partitions, i.e., 87.8% of the observations stay always together in the same group.

We compared the solutions obtained in the algorithms through the between-groups variability measure and F-statistic, where we estimated the concentration parameters. See Table 4.

Table 4: Between-groups variability measure and F-statistic for the final partitions

Algorithm	Soft-movMF	Hard-movMF	Stochastic EM	DC
Between-groups variability	21.549	22.353	26.136	20.004
F-statistic	12.048	21.000	24.549	35.088

The solution obtained in stochastic EM is preferable in what concerns to the between--groups variability, but considering the F-statistic, the solution obtained with dynamic clusters algorithm is preferable.

5 Concluding remarks

The simulations revealed that only for poorly or moderately separated components, the variants of the EM algorithm and the dynamic clusters type algorithm lead to different solutions in general. For very well separated components, the algorithms seem to originate the same result. Additionally, the larger concentration parameters associated with the components, greater is the tendency to obtain the same solution for the algorithms.

For each algorithm, as expected, the between-groups variability and the F-statistic increase when the separation between components increases or when the concentration of components increases, since these components are not badly separated (i.e, the angle θ is 30^o).

Acknowledgements

- The author is grateful to the referees of this journal for their helpful comments.

- This work is financed by the ERDF European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project FCOMP-01-0124-FEDER-037281.

References

[1] Banerjee A., Dhillon I. S., Ghosh J. and Sra S. (2005). Clustering on the unit hypersphere using von Mises-Fisher distributions, *Journal of Machine Learning Research*, 6, 1345-1382.

[2] Celeux G. and Govaert G. (1992). A classification EM algorithm for clustering and two-stochastic versions, *Computational Statistics and Data Analysis*, 14, 3, 315-332.

[3] Dempster A. P., Laird N. M. and Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society*, B, 3, 1-38.

[4] Dhillon, I. S. and Modha D. S. (2001). Concept Decompositions for large Sparse Text Data Using Clustering, *Machine Learning*, 42, 1, 143-175.

[5] Diday E. and Schroeder A. (1976). New approach in mixed distributions detection, *Révue Française D' Automatique Informatique Recherche Operationelle*, 10, 6, 75-106.

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

[7] Figueiredo, A. and Gomes, P. (2015). Clustering of variables based on Watson distribution on hypersphere: a comparison of algorithms. *Communications in Statistics - Simulation and Computation.* Special issue: Joint Meeting of y-BIS and jSPE, vol. 44, issue 10, pp. 2622-2635.

[8] Kent J. T. (1982). The Fisher-Bingham distribution on the sphere, *Journal of the Royal Statistical Society*, Series B, 44, 71-80.

[9] Harrison D., Kanji G. K. and Gadsden R. J. (1986). Analysis of variance for circular data, *Journal of Applied Statistics* **13**, 123-138.

[10] Hornik, K. and Grun B. (2014). movMF: An R Package for fitting mixtures of von Mises-Fisher distributions, *Journal of Statistical Software*, 58, 10, 1-26.

[11] Hornik K., Feinerer I., Kober M. and Buchta C. (2012). Spherical k-Means Clustering, *Journal of Statistical Software*, 50, 10, 1-21.

[12] Mardia K. V. and Jupp P. E. (2000). *Directional Statistics.* John Wiley and Sons, Chichester.

[13] Peel D., Whiten W. J. and McLachlan G. J. (2001). Fitting mixtures of Kent distributions to aid in joint set identification, *Journal of the American Statistical Association*, 96, 56-63.

[14] Redner R. and Homer W. (1984). Mixture Densities, Maximum Likelihood, EM algorithm, *Journal of the Royal Statistical Society*, B, 39, 1-38.

[15] Watson G. S. (1956). Analysis of dispersion on a sphere, *Monthly Notices of the Royal Astronomical Society Geophysical Supplement*, **7**, 153-159.

[16] Watson G. S. and Williams E. J. (1956). On the construction of significance tests on the circle and the sphere, *Biometrika*, **43**, 344-352.

[17] Wood A. (1982). A bimodal distribution on the sphere, *Applied Statistics*, 31, 1, 52-58.

[18] Wood A. (1994). Simulation of the von Mises-Fisher distribution, *Communications in Statistics - Simulation and Computation*, 23, 1, 157-164.

Appendix A

Derivation of the expected log-likelihood of the complete sample

The vectors $\mathbf{z}_i$ are independent and have multinomial distribution with parameters $(1, \pi_1, ..., \pi_k)$ and the probability density function is given by

$$g(\mathbf{z}_i|Q) = \prod_{j=1}^{k} \pi_j^{z_{ij}}.$$

The density function of $\mathbf{x}_i|\mathbf{z}_i$ is given by

$$l(\mathbf{x}_i|\mathbf{z}_i, Q) = \prod_{j=1}^{k} f(\mathbf{x}_i|\boldsymbol{\theta}_j)^{z_{ij}}.$$

Then, the density function of $(\mathbf{x}_i, \mathbf{z}_i)$ is defined by the product

$$h(\mathbf{x}_i, \mathbf{z}_i|\theta) = g(\mathbf{z}_i|Q) l(\mathbf{x}_i|\mathbf{z}_i, Q).$$

Replacing the densities $g(\mathbf{z}_i|Q)$ and $l(\mathbf{x}_i|\mathbf{z}_i, Q)$ in the previous expression, we obtain

$$h(\mathbf{x}_i, \mathbf{z}_i|Q) = \prod_{j=1}^{k} \pi_j^{z_{ij}} f(\mathbf{x}_i|\boldsymbol{\theta}_j)^{z_{ij}}.$$

The complete data log-likelihood of $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n, Z)$ is given by

$$L(Q|\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n, Z) = \ln \prod_{i=1}^{n} h(\mathbf{x}_i, \mathbf{z}_i|Q)$$

and replacing $h(\mathbf{x}_i, \mathbf{z}_i|Q)$, we obtain the expression

$$L(Q|\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n, Z) = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \ln \left[\pi_j f(\mathbf{x}_i|\boldsymbol{\theta}_j)\right].$$

The density function of $\mathbf{z}_i|\mathbf{x}_i$ given by

$$f(\mathbf{z}_i|\mathbf{x}_i, Q) = \frac{h(\mathbf{x}_i, \mathbf{z}_i|Q)}{\psi(\mathbf{x}_i|Q)}.$$

Replacing the densities h and ψ, we obtain

$$f(\mathbf{z}_i|\mathbf{x}_i, Q) = \frac{\prod\limits_{j=1}^{k} [\pi_j f(\mathbf{x}_i|\boldsymbol{\theta}_j)]^{z_{ij}}}{\sum\limits_{h=1}^{k} \pi_h f(\mathbf{x}_i|\theta_h)}.$$

The expected value of $Z_{ij}|\mathbf{x}_i$ is given by the expression

$$E(Z_{ij}|\mathbf{x}_i, Q) = \frac{\pi_j f(\mathbf{x}_i|\boldsymbol{\theta}_j)}{\sum\limits_{h=1}^{k} \pi_h f(\mathbf{x}_i|\boldsymbol{\theta}_h)}.$$

This expected value is the a-posteriori probability of $\mathbf{x}_i$ belonging to C_j, which we denote by $t_j(\mathbf{x}_i)$. Then, the expected complete data log-likelihood may be written as

$$L(Q|\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n, Z) = \sum_{i=1}^{n} \sum_{j=1}^{k} t_j(\mathbf{x}_i) \ln [\pi_j f(\mathbf{x}_i|\boldsymbol{\theta}_j)].$$

Appendix B

Derivation of the maximum likelihood estimators

First, consider the function $L_1(Q)$ subject to the constraint $\boldsymbol{\mu}_j^T \boldsymbol{\mu}_j = 1$:

$$L_1(Q) = \sum_{i=1}^{n} \sum_{j=1}^{k} t_j(\mathbf{x}_i) \left\{ \ln c_p(\kappa_j) + \kappa_j \boldsymbol{\mu}_j^T \mathbf{x}_i \right\} - \lambda_1(\boldsymbol{\mu}_j^T \boldsymbol{\mu}_j - 1),$$

where λ_1 is a Lagrange multiplier and $t_j(\mathbf{x}_i)$ is defined by (7). The maximum likelihood estimator of $\boldsymbol{\mu}_j$ is the solution of the following equation

$$\frac{\partial L_1(Q)}{\partial \boldsymbol{\mu}_j} = 0 \Leftrightarrow \sum_{i=1}^{n} t_j(\mathbf{x}_i) \kappa_j \mathbf{x}_i - 2\lambda_1 \boldsymbol{\mu}_j = 0 \Leftrightarrow \boldsymbol{\mu}_j = \frac{\kappa_j}{2\lambda_1} \sum_{i=1}^{n} t_j(\mathbf{x}_i) \mathbf{x}_i.$$

As $\boldsymbol{\mu}_j^T \boldsymbol{\mu}_j = 1$, then the Lagrange multiplier is given by

$$\lambda_1 = \frac{\kappa_j}{2\lambda_1} \left\| \sum_{i=1}^{n} t_j(\mathbf{x}_i) \mathbf{x}_i \right\|.$$

So the maximum likelihood estimator of $\boldsymbol{\mu}_j$ in the $(m+1)$th iteration, $\widehat{\boldsymbol{\mu}}_j^{(m+1)}$ is given by

$$\widehat{\boldsymbol{\mu}}_j^{(m+1)} = \frac{\sum\limits_{i=1}^{n} t_j^{(m)}(\mathbf{x}_i) \mathbf{x}_i}{\left\| \sum\limits_{i=1}^{n} t_j^{(m)}(\mathbf{x}_i) \mathbf{x}_i \right\|}, \ \textit{j=1,...,k.}$$

Second, the maximum likelihood estimator of κ_j is the solution of the equation

$$\frac{\partial L_1(Q)}{\partial \kappa_j} = 0 \Leftrightarrow \sum_{i=1}^{n} t_j(\mathbf{x}_i) \frac{c_p'(\kappa_j)}{c_p(\kappa_j)} + \sum_{i=1}^{n} t_j(\mathbf{x}_i) \boldsymbol{\mu}_j^T \mathbf{x}_i = 0.$$

Let $c_p'(\kappa_j) \big/ c_p\left(\kappa_j\right) = -A(\kappa_j)$ and then the previous equation may be written as

$$A(\kappa_j) \sum_{i=1}^{n} t_j(\mathbf{x}_i) = \sum_{i=1}^{n} t_j(\mathbf{x}_i) \boldsymbol{\mu}_j^T \mathbf{x}_i.$$

Replacing $\boldsymbol{\mu}_j$ by $\widehat{\boldsymbol{\mu}}_j^{(m+1)}$, the maximum likelihood estimator of κ_j obtained in the $(m+1)$th iteration, $\widehat{\kappa}_j^{(m+1)}$ is the solution of the equation

$$A\left(\widehat{\kappa}_j^{(m+1)}\right) = \frac{R_j}{\sum\limits_{i=1}^{n} t_j^{(m)}(\mathbf{x}_i)},\ \textit{j=1,...,k},$$

where R_j is the length of the vector $\sum_{i=1}^{n} t_j^{(m)}\left(\mathbf{x}_i\right) \mathbf{x}_i$, that is $R_j = \| \sum_{i=1}^{n} t_j^{(m)}\left(\mathbf{x}_i\right) \mathbf{x}_i \|$, i.e,

$$\widehat{\kappa}_j^{(m+1)} = A^{-1} \left(\frac{R_j}{\sum\limits_{i=1}^{n} t_j^{(m)}(\mathbf{x}_i)} \right),\ \textit{j=1,...,k}.$$

Third, consider the function, $L_2\left(Q\right)$ subject to the constraint $\sum_{j=1}^{k} \pi_j = 1$, that is, maximize $\sum_{i=1}^{n} \sum_{j=1}^{k} t_j(\mathbf{x}_i) \ln \pi_j - \lambda_2(\sum_{j=1}^{k} \pi_j - 1)$, where λ_2 is a Lagrange multiplier. The maximum likelihood estimator of π_j is the solution of the equation

$$\frac{\partial L_2(Q)}{\partial \pi_j} = 0 \Leftrightarrow \sum_{i=1}^{n} t_j\left(\mathbf{x}_i\right) \frac{1}{\pi_j} - \lambda_2 = 0.$$

Summing the last equation from $j = 1$ to k, we obtain $\lambda_2 = n$. Then the maximum likelihood estimator of π_j in the $(m+1)$th iteration, $\widehat{\pi}_j^{(m+1)}$ is given by

$$\widehat{\pi}_j^{(m+1)} = \frac{1}{n} \sum_{i=1}^{n} t_j^{(m)}\left(\mathbf{x}_i\right),\ \textit{j=1,...,k}.$$

Appendix C

Derivation of the common concentration estimator

When all concentration parameters κ_i are equal to κ, the expression $L_1(Q)$ given by (5) reduces to

$$L_1(Q) = \sum_{i=1}^{n} \sum_{j=1}^{k} t_j(\mathbf{x}_i) \left[\ln c_p\left(\kappa\right) + \kappa \boldsymbol{\mu}_j^T \mathbf{x}_i\right] - \lambda_1(\boldsymbol{\mu}_j^T \boldsymbol{\mu}_j - 1),$$

The estimator of κ is the solution of the equation

$$\frac{\partial L\left(Q\right)}{\partial \kappa}=0 \Leftrightarrow \sum_{i=1}^{n}\sum_{j=1}^{k} t_j(\mathbf{x}_i)\frac{c_p^{'}(\kappa)}{c_p(\kappa)}+\sum_{i=1}^{n}\sum_{j=1}^{k} t_j(\mathbf{x}_i)\boldsymbol{\mu}_j^T\mathbf{x}_i=0.$$

The last equation is equivalent to the following one

$$nA(\kappa)=\sum_{i=1}^{n}\sum_{j=1}^{k} t_j(\mathbf{x}_i)\frac{\left(\sum\limits_{i=1}^{n} t_j(\mathbf{x}_i)\mathbf{x}_i\right)^T}{\left\|\sum\limits_{i=1}^{n} t_j(\mathbf{x}_i)\mathbf{x}_i\right\|}\mathbf{x}_i,$$

where $A(\kappa)=c_p^{'}(\kappa)\big/c_p(\kappa)$. The maximum likelihood estimator of κ in the $(m+1)$th iteration, $\widehat{\kappa}^{(m+1)}$ is the solution of the equation

$$A(\widehat{\kappa}^{(m+1)})=\frac{\sum\limits_{j=1}^{k} R_j}{n},$$

where R_j is defined as before.