Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: http://www.tandfonline.com/loi/lssp20

Clustering of Variables Based on Watson Distribution on Hypersphere: A Comparison of Algorithms

Adelaide Figueiredo & Paulo Gomes

To cite this article: Adelaide Figueiredo & Paulo Gomes (2015) Clustering of Variables Based on Watson Distribution on Hypersphere: A Comparison of Algorithms, Communications in Statistics - Simulation and Computation, 44:10, 2622-2635, DOI: 10.1080/03610918.2014.901353

To link to this article: https://doi.org/10.1080/03610918.2014.901353

Accepted author version posted online: 19 Jun 2014.
Published online: 19 Jun 2014.

Communications in Statistics—Simulation and Computation®, 44: 2622–2635, 2015

ISSN: 0361-0918 print / 1532-4141 online
DOI: 10.1080/03610918.2014.901353

Taylor & Francis
Taylor & Francis Group

Clustering of Variables Based on Watson Distribution on Hypersphere: A Comparison of Algorithms

ADELAIDE FIGUEIREDO[1] AND PAULO GOMES[2,3]

[1]School of Economics and LIAAD-INESC TEC, University of Porto, Porto, Portugal
[2]Statistics Portugal, Lisbon, Portugal
[3]Nova University of Lisbon, ISEGI, Lisbon, Portugal

We consider n *individuals described by* p *variables, represented by points of the surface of unit hypersphere. We suppose that the individuals are fixed and the set of variables comes from a mixture of bipolar Watson distributions. For the mixture identification, we use EM and dynamic clusters algorithms, which enable us to obtain a partition of the set of variables into clusters of variables.*

Our aim is to evaluate the clusters obtained in these algorithms, using measures of within-groups variability and between-groups variability and compare these clusters with those obtained in other clustering approaches, by analyzing simulated and real data.

Keywords Dynamic clusters algorithm; *EM* algorithm; Hierarchical clustering; Principal cluster component analysis; Variable clustering; Watson distribution

Mathematics Subject Classification 62H30; 62H11; 62H12; 62H25.

1. Introduction

Clustering of variables is very useful in practical situations where there is interest in forming homogeneous groups of variables, such as in studies of preference, sensory studies, clinical trials, studies of chemical pollutants in the environment, food industry, etc. (see among others, the works of Hulshof et al., 1992; Qannari et al., 1997; Vigneau et al., 2001; Vigneau and Qannari, 2002; Carbonell et al., 2008). There is a large variety of hierarchical clustering methods that may be used to cluster either individuals or variables (see, e.g., Everitt, 1993). Other works on clustering variables include Vigneau and Qannari (2003) who considered the clustering of variables around latent components, which consists of performing a hierarchical cluster analysis, followed by a partitioning algorithm and Soffritti (1999) who suggested a hierarchical method using a multivariate association measure based on the links between canonical correlation analysis and principal component analysis, and compared the method with other possible solutions.

Received October 15, 2012; Accepted February 24, 2014
Address correspondence to Adelaide Maria de Sousa Figueiredo, Faculdade de Economia da Universidade do Porto, Agrupamento Cientifico de Matemáticae Sistemas de Informação, Rua Dr. Roberto Frias, Porto, 4200-464 Informação Portugal; E-mail: adelaide@fep.up.pt

We consider multivariate data with n individuals described by p variables. In the classical approach, it is usual to assume that the p variables are fixed and the n individuals are randomly selected from a population of individuals. Now, we consider the dual approach, where we suppose that the n individuals are fixed and the p variables are randomly selected from a population of variables. We standardize the variables to be points of the surface of the n-dimensional unit sphere, denoted by $S_{n-1} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}'\mathbf{x} = 1\}$ and this kind of data is known as directional data.

In directional data literature, there are essentially applications with data on circle S_1 and on sphere S_2 and the most used distributions are von Mises distribution for modeling circular data and Fisher distribution or Watson distribution for spherical data (e.g., Fisher et al., 1987; Mardia and Jupp, 2000). Other distribution for spherical data is Kent distribution (Kent, 1982), which was used by Peel et al. (2001) to form groups of fracture data via a model-based clustering. Recently, some applications of directional data on higher dimensions have appeared in literature, in areas such as text analysis, Biostatistics, etc. Dortet-Bernadet and Wicker (2008) considered a model-based clustering of data that lie on a unit sphere, being the inverse stereographic projection of a multivariate normal distribution considered as the directional distribution and these authors applied the clustering method to gene expression profiles. Banerjee et al. (2005) used in a model-based clustering of directional data, the von Mises–Fisher, which is an extension to higher dimensions of von Mises distribution and Fisher distribution and these authors applied the clustering method to text analysis.

We suppose that the sample of variables represented by points of the unit hypersphere is formed by k clusters of variables and each cluster comes from a bipolar Watson distribution. So we associate with the sample of variables a mixture of k bipolar Watson distributions defined on the hypersphere, as in Gomes (1987) and Figueiredo and Gomes (2006a). These authors considered an approach based on sampling of variables and introduced some new results concerning the bipolar Watson distribution, taking into account not a sample of individuals, but a sample of variables. This type of idea was discussed by Hotelling (1933) who in the context of principal components studied the convergence of the eigenvalues and eigenvectors of the covariance matrix of groups of variables randomly chosen from a population of variables, when the dimension of the groups increases. Escoufier (1973) also proposed a new coefficient for evaluating the proximity of two groups of variables, but supposing that the variables are observed.

For the identification of the mixture of k bipolar Watson distributions defined on the hypersphere, we use the *EM* algorithm proposed by Dempster et al. (1977), which was applied by Figueiredo and Gomes (2006a) and the dynamic clusters algorithm proposed by Diday and Schroeder (1976), which was presented by Gomes (1987). The identification of the mixture allows us to obtain estimates of mixture parameters and a partition of the sample of variables into clusters of variables. Each cluster is associated with a privileged direction and a concentration parameter, which measures the concentration around the privileged direction. The maximum likelihood estimate of the privileged direction of each cluster corresponds to the first principal component of the cluster. Thus, our approach is similar to the variable clustering approach based on the dynamic clusters principle, denoted by principal cluster component analysis (PCCA) proposed by Escoufier (1988). This approach was used more recently by Vigneau and Qannari (2003), who extended it, for instance, to the classification of variables taking external data into account. PCCA starts with an initial partition and then two functions f and g are successively applied until convergence is attained. The function f associates with each group of the partition, the first principal component of the group. The function g associates with each first principal component, a group formed by the variables that are more correlated with that first principal component than with others first principal components.

The goodness-of-fit methods for the bipolar Watson distribution on hypersphere, proposed by Figueiredo and Gomes (2006b) may be applied to check whether the clusters of variables obtained in the algorithms come from bipolar Watson populations. In these methods, the goodness-of-fit for the bipolar Watson distribution defined on the hypersphere for large concentration parameter is reduced to the goodness-of-fit of a chi-square distribution. To test whether a sample comes from a chi-square population, the chi-square Q–Q plot and Kolmogorov–Smirnov and chi-square tests may be used. Additionally, before applying the goodness-of-fit methods, Giné and Bingham uniformity tests may be applied to test whether the sample comes from an uniform population defined on the hypersphere. These uniformity tests can be seen in Figueiredo and Gomes (2003) as well as the power of the tests against a bipolar Watson population or a mixture of bipolar Watson populations.

In Section 2, we review the distribution used in this article, the bipolar Watson distribution defined on the hypersphere and some useful properties of this distribution, including the maximum likelihood estimates of its parameters. In Section 3, we present the two algorithms used for the identification of the mixture. In Section 4, after defining the variability measures between-groups and within-groups, we report on a simulation study, in which we use these measures to evaluate the solutions obtained in EM algorithm and dynamic clusters algorithm and compare them to the solutions obtained with a hierarchical clustering method and PCCA, for various dimensions of the sphere, different sample sizes, and also different parameters of the mixture components. In Section 5, we compare the two proposed methods with a hierarchical clustering method and PCCA, using real data. Finally, in Section 6 we conclude the article.

2. The Watson Distribution on the Hypersphere

We consider a particular case of Watson distribution defined on the hypersphere, the bipolar Watson distribution on hypersphere, denoted by $W_n(\mathbf{u}, \kappa)$, with probability density function given by

$$f(\mathbf{x}) = \left\{ {}_1F_1\left(\frac{1}{2}, \frac{n}{2}, \kappa\right)\right\}^{-1} \exp\{\kappa(\mathbf{u}'\mathbf{x})^2\}, \quad \mathbf{x} \in S_{n-1}, \quad \mathbf{u} \in S_{n-1},\ \kappa > 0, \tag{2.1}$$

where ${}_1F_1(0.5, n/2, \kappa)$ is the confluent hypergeometric function defined by

$${}_1F_1\left(\frac{1}{2}, \frac{n}{2}, \kappa\right) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)} \int_0^1 \exp(\kappa t)\, t^{-05}\, (1-t)^{(n-3)/2}\, \mathrm{d}t. \tag{2.2}$$

This distribution has two parameters: a directional parameter $\mathbf{u}$ and a concentration parameter κ, which measures the concentration around $\mathbf{u}$. It is rotationally symmetric about the principal axis $\mathbf{u}$.

If $\mathbf{x}$ comes from the bipolar Watson population $W_n(\mathbf{u}, \kappa)$, then for large κ (see Mardia and Jupp, 2000, p. 236):

$$2\kappa\{1 - (\mathbf{u}'\mathbf{x})^2\} \quad \dot{\sim} \quad \chi^2_{n-1}, \quad \kappa \to \infty. \tag{2.3}$$

Let $[\mathbf{x}_1|\mathbf{x}_2|...|\mathbf{x}_p]$ be a sample of variables from the bipolar Watson distribution defined on the hypersphere $W_n(\mathbf{u}, \kappa)$.

The maximum likelihood estimator of the parameter $\mathbf{u}$ is the eigenvector of the orientation matrix $T = \sum \mathbf{x}_i \mathbf{x}_i'$ associated with the largest eigenvalue w. So it follows that the maximum likelihood estimator of the directional parameter $\mathbf{u}$ based on the sample of variables is the first principal component of the sample.

The maximum likelihood estimator of the parameter κ is the solution of the equation $Y(\widehat{\kappa}) = \frac{w}{p}$, where the function $Y(.)$ is defined by $Y(\kappa) = \mathrm{d}\ln_1 F_1(0.5, n/2, \kappa)/\mathrm{d}\kappa$.

For more details about this distribution, see, for example, Mardia and Jupp (2000) and Fisher et al. (1987).

3. Identification of a Mixture of Bipolar Watson Distributions

The probability density function of a mixture with k bipolar Watson components on hypersphere is given by

$$\varphi(\mathbf{x}; \boldsymbol{\phi}) = \sum_{i=1}^{k} \pi_i f(\mathbf{x}|\mathbf{u}_i, \kappa_i), \quad \mathbf{x} \in S_{n-1}, \quad \mathbf{u}_i \in S_{n-1}, \ \kappa_i > 0, \tag{3.1}$$

where π_i, $i = 1, ..., k$ are the mixture proportions, $0 \leq \pi_i \leq 1$, $\forall_i$, $\sum_{i=1}^{k} \pi_i = 1$; $f(\mathbf{x}|\mathbf{u}_i, \kappa_i)$ is the density function of the ith component of the mixture, that is, the density of $W_n(\mathbf{u}_i, \kappa_i)$ distribution given by (2.1) and $\boldsymbol{\phi} = (\mathbf{u}_1, ..., \mathbf{u}_k, \kappa_1, ..., \kappa_k, \pi_1, ..., \pi_k)$ is the parameter vector of the mixture.

For obtaining a partition of the set of variables $[\mathbf{x}_1|\mathbf{x}_2|...|\mathbf{x}_p]$ into homogeneous groups of variables, we consider the *EM* algorithm and the dynamic clusters algorithm, briefly described bellow.

3.1. EM *Algorithm*

The *EM* algorithm is applied to solve the likelihood equations in the mixture parameters estimation.

The algorithm proceeds iteratively in two steps *E*- Estimation and *M*- Maximization.

The algorithm starts with an initial solution for instance to the parameter vector of the mixture: $\boldsymbol{\phi}^0 = (\mathbf{u}_1^0, ..., \mathbf{u}_k^0, \kappa_1^0, ..., \kappa_k^0, \pi_1^0, ..., \pi_k^0)$. In the mth iteration, the two steps are:

E-Step

Use estimates $\boldsymbol{\phi}^{(m)}$ of the mixture parameters in the mth iteration to estimate the posterior probability of $\mathbf{x}_i$ belonging to the jth mixture component:

$$t_j^{(m)}(\mathbf{x}_i) = \frac{\pi_j^{(m)} f\left(\mathbf{x}_i|\mathbf{u}_j^{(m)}, \kappa_j^{(m)}\right)}{\sum_{h=1}^{k} \pi_h^{(m)} f\left(\mathbf{x}_i|\mathbf{u}_h^{(m)}, \kappa_h^{(m)}\right)}, \ j = 1, ..., k, \ i = 1, ..., p. \tag{3.2}$$

M-Step

Use the estimates $t_j^{(m)}(\mathbf{x}_i)$ given by (3.2) to maximize the logarithm of the likelihood function.

The estimate of $\mathbf{u}_j$ in the $(m+1)$th iteration is the eigenvector associated with the largest eigenvalue w_j of the matrix $\sum_{i=1}^{p} t_j^{(m)}(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'$, that is,

$$\left(\sum_{i=1}^{p} t_j^{(m)}(\mathbf{x}_i)\, \mathbf{x}_i \mathbf{x}_i'\right) \widehat{\mathbf{u}}_j^{(m+1)} = w_j \widehat{\mathbf{u}}_j^{(m+1)}, \ j = 1, ..., k; \tag{3.3}$$

the estimate of κ_j in the $(m+1)$-th iteration is the solution of the equation

$$Y\left(\widehat{\kappa}_j^{(m+1)}\right) = \frac{w_j}{\sum_{i=1}^{p} t_j^{(m)}(\mathbf{x}_i)},\ j = 1, ..., k; \tag{3.4}$$

and the estimate of π_j in the $(m+1)$th iteration is given by

$$\widehat{\pi}_j^{(m+1)} = \frac{\sum_{i=1}^{p} t_j^{(m)}(\mathbf{x}_i)}{p},\ j = 1, ..., k. \tag{3.5}$$

A partition $(P_1, ..., P_k)$ of the set of variables is obtained assigning the variable $\mathbf{x}_i$ to the component for which the posterior probability is the largest:

$$P_j = \{\mathbf{x}_i : t_j(\mathbf{x}_i) = \max t_h(\mathbf{x}_i),\ h = 1, ..., k\}, \tag{3.6}$$

and if $t_j(\mathbf{x}_i) = t_h(\mathbf{x}_i)$, consider $\mathbf{x}_i \in P_j$ if $j < h$.

3.2. *Dynamic Clusters Algorithm*

The aim is to determine a partition of the set of variables into k groups $(P_1, ..., P_k)$, so that P_j group can be considered to come from bipolar Watson $W_n(\mathbf{u}_j, \kappa_j)$ subpopulation.

The algorithm starts with a initial partition $(P_1^0, ..., P_k^0)$ or a set of values $(\lambda_1^0, ..., \lambda_k^0)$ for the parameters of the bipolar Watson components.

Two functions f and g are successively applied until convergence is attained. The f function associates to a partition, a set of values for the parameters of the bipolar Watson distributions:

$$f : (P_1, ..., P_k) \to (\lambda_1, ..., \lambda_k),$$

where $\lambda_j = (\mathbf{u}_j, \kappa_j)$ are the parameters of the bipolar Watson distribution associated with P_j group, estimated through maximum likelihood, based on P_j group.

The g function associates a partition to a set of values for parameters

$$g : (\lambda_1, ..., \lambda_k) \to (P_1, ..., P_k),$$

where P_j group of the partition is composed by variables that are closer to the bipolar Watson distribution with parameter $\lambda_j = (\mathbf{u}_j, \kappa_j)$ than with other parameters.

We consider that a variable $\mathbf{x}$ is close to distribution with $\lambda = (\mathbf{u}, \kappa)$ parameter if $f_\lambda(\mathbf{x})$ density of the bipolar Watson distribution with λ parameter is large, and then the distance between $\mathbf{x}$ and λ is defined in the following way

$$D(\mathbf{x}, \lambda) = \ln\left(\frac{C}{f_\lambda(\mathbf{x})}\right), \tag{3.7}$$

where the constant C must be chosen in the following way:

$$C \geq \max\left\{f_{\lambda_j}(\mathbf{x}),\quad j = 1, ..., k, \forall_{\mathbf{x}}\right\}, \tag{3.8}$$

so that the distance definition results in a set of limited inferiorly values. The criterion to be optimized is a function of P^* partition and parameters obtained in convergence:

$$\sum_{1 \leq i \leq k} \sum_{\mathbf{x} \in P_i^*} D(\mathbf{x}, \lambda_i). \tag{3.9}$$

4. Performance of the Algorithms

Figueiredo and Gomes (2006a) studied the properties of the maximum likelihood estimators obtained through *EM* algorithm for a mixture of k bipolar Watson distributions defined on the hypersphere. Next, we report on a simulation study to analyze the performance of the EM and Dynamic Clusters algorithms in clustering variables, that is, in evaluating the solutions obtained by these algorithms, through the variability measures between-groups and within-groups defined in Mardia and Jupp (2000, p. 240; see also Gomes and Figueiredo, 1999). The between-groups variability measure is defined by the expression

$$\sum_{i=1}^{k} \widehat{\lambda}_i - \widehat{\lambda} = \sum_{i=1}^{k} \sum_{j=1}^{p_i} \widehat{\kappa}_i \left\{ \left(\widehat{\mathbf{u}}_i' \mathbf{x}_{ij} \right)^2 - \left(\widehat{\mathbf{u}}' \mathbf{x}_{ij} \right)^2 \right\}$$

and the within-groups variability measure is defined by the expression

$$\sum_{i=1}^{k} \left(\widehat{\kappa}_i \, p_i - \widehat{\lambda}_i \right) = \sum_{i=1}^{k} \sum_{j=1}^{p_i} \widehat{\kappa}_i \left\{ 1 - \left(\widehat{\mathbf{u}}_i' \mathbf{x}_{ij} \right)^2 \right\},$$

where $X_i = [\mathbf{x}_{i1} | \mathbf{x}_{i2} | ... | \mathbf{x}_{ip_i}]$ represents the sample of p_i variables of ith subpopulation with $W_n(\mathbf{u}_i, \kappa_i)$ distribution, $i = 1, ..., k$, $p = \sum_{i=1}^{k} p_i$ denotes the total number of variables, $\widehat{\mathbf{u}}_i$ is the eigenvector associated with the largest eigenvalue $\widehat{\lambda}_i$ of $\widehat{\kappa}_i X_i X_i'$, $i = 1, ..., k$, $\widehat{\mathbf{u}}$ is the eigenvector associated with the largest eigenvalue $\widehat{\lambda}$ of $\sum_{i=1}^{k} \widehat{\kappa}_i X_i X_i'$ and $\widehat{\kappa}_i$ is the maximum likelihood estimate of the concentration parameter κ_i associated with the ith bipolar Watson component, $i = 1, ..., k$.

Under the null hypothesis of equality of the directional parameters of Watson distributions, and for large concentration parameters, the statistic $F = \frac{(\sum_{i=1}^{k} \widehat{\lambda}_i - \widehat{\lambda}) / (k-1)(n-1)}{\sum_{i=1}^{k} (\widehat{\kappa}_i \, p_i - \widehat{\lambda}_i) / (p-k)(n-1)}$ has an approximately $F_{(k-1)(n-1),(p-k)(n-1)}$ distribution.

We used a rejection-type method proposed by Huo (1984) for the simulation of the bipolar Watson distribution defined on the hypersphere. As EM algorithm and Dynamic Clusters algorithm depend on the initial solution, we compared both algorithms for the same initial solution. We took for the initial solution of the algorithms in each case, the partition obtained in the hierarchical clustering method, using the linear correlation coefficient as a similarity measure between variables and the complete linkage criterion (furthest neighbor) as an aggregation criterion.

First, we generated simulated samples from mixtures with equal proportions of two bipolar Watson $W_n(\mathbf{u}_1, \kappa)$ and $W_n(\mathbf{u}_2, \kappa)$ distributions, with common concentration parameter κ. We considered different sphere dimensions ($n = 10, 20, 30$); various sample sizes p for each n ($p = 20, 30$ for $n = 10$, $p = 30$ for $n = 20$ and $p = 50$ for $n = 30$); several values of the concentration parameter associated with bipolar Watson components ($\kappa = 10, 30, 50, 100$) and different angles between directional parameters of the components ($\theta = 18^\circ, (18^\circ), 90^\circ$).

For each final solution obtained in the algorithms, we calculated the variability measures between-groups and within-groups, which are represented in Figure 1 for $n = 10$, $p = 20$ and $n = 10$, $p = 30$ and in Figure 2 for $n = 20$, $p = 30$ and $n = 30$, $p = 50$. The solutions obtained by the algorithms were equal, and consequently the variability measures coincided for both algorithms, except for very overlapped components, that is, when the angle between directional parameters associated with the components was very small ($\theta = 18^o$) or for components with small concentration parameter for a sphere dimension n

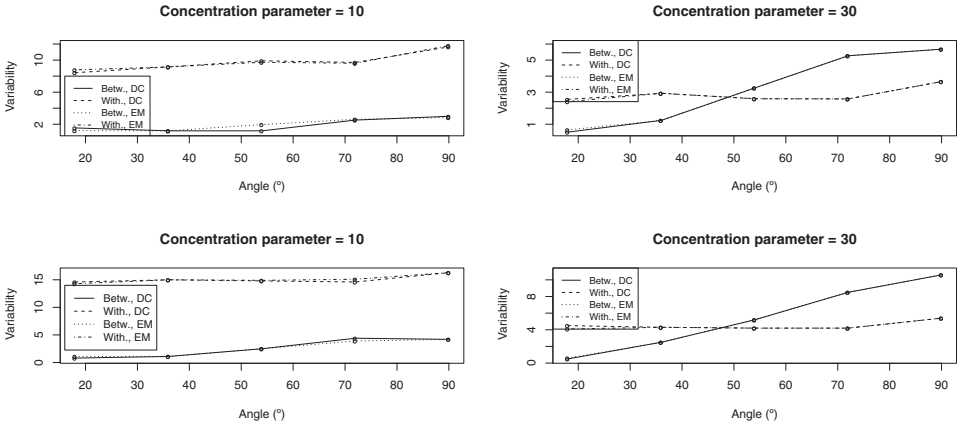

Figure 1. Between-groups and within-groups variability for the solutions obtained in EM and DC algorithms for $n = 10$, $p = 20$ (above) and $n = 10, p = 30$ (below).

($\kappa = 10$). Consequently, in general in each graph of Figures 1 and 2, the respective lines are overlapped for both algorithms, being visible only two instead of four lines.

From Figures 1 and 2, we observe that, except when the common concentration parameter is small for each sphere dimension n ($\kappa = 10$ for $n = 10$, $\kappa = 30$ for $n = 20$ and $\kappa = 50$ for $n = 30$), the between-groups variability increases substantially as the separability between components increases. Additionally, for mixtures with well-separated components, the between-groups variability exceeds largely the within-groups variability, while for poorly separated mixtures, within-groups variability is larger than between-groups variability.

Second, we generated samples from mixtures with equal proportions of two bipolar Watson $W_n(\mathbf{u}_1, \kappa_1)$ and $W_n(\mathbf{u}_2, \kappa_2)$ distributions, with different concentration parameters

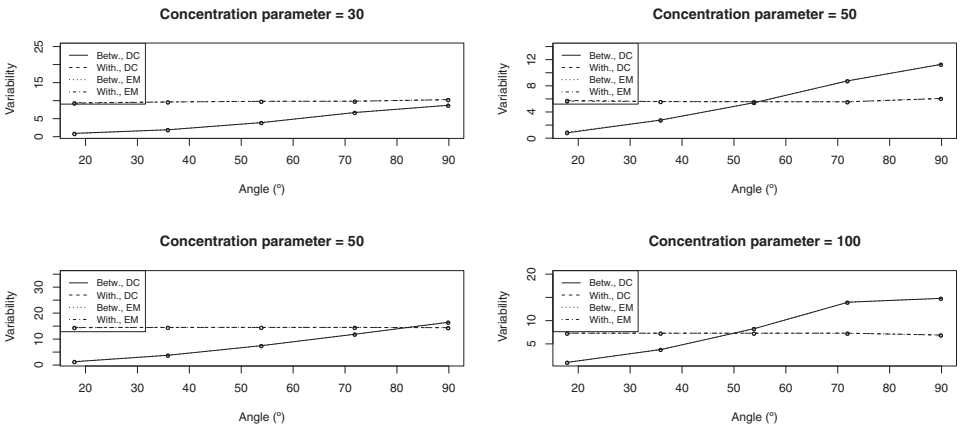

Figure 2. Between-groups and within-groups variability for the solutions obtained with EM and DC algorithms for $n = 20$, $p = 30$ (above), $n = 30$, $p = 50$ (below).

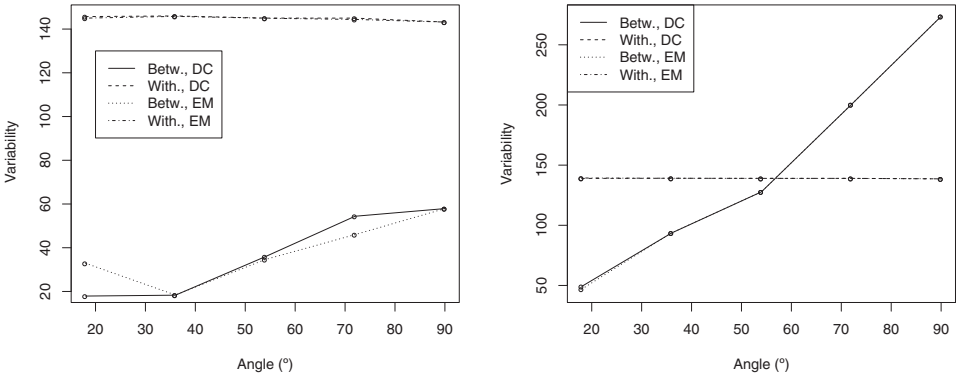

Figure 3. Between-groups and within-groups variability for the solutions obtained with EM and DC algorithms for $n = 10$, $p = 30$ and $\kappa_1 = 10$, $\kappa_2 = 20$ (left), $\kappa_1 = 20$, $\kappa_2 = 30$ (right).

κ_1 and κ_2. For the solutions obtained in the algorithms, we calculated variability measures between-groups and within-groups. We considered $n = 10$, $p = 30$, different angles between directional parameters of the components, $\theta = 18^\circ, (18^\circ), 90^\circ$ and different concentration parameters ($\kappa_1 = 10$, $\kappa_2 = 20$ or $\kappa_1 = 10$, $\kappa_2 = 30$).

Both algorithms gave the same solution except for poorly-separated components (that is, small angle between the directional parameters of the components, $\theta = 18^\circ, 36^\circ$) or for components with relatively small concentration parameters ($\kappa_1 = 10$, $\kappa_2 = 20$ for $n = 10$).

So the variability measures coincided in many cases and consequently we observe in general two instead of four lines in each graph of Fig. 3.

Similarly to the case of components with common concentration parameter, we observe that, except for small concentration parameters ($\kappa_1 = 10$ and $\kappa_2 = 20$ for $n = 10$), the between-groups variability increases rapidly as the angle between directional parameters increases. In addition, for well-separated components, the between-groups variability is larger than the within-groups variability and for poorly separated components, the within-groups variability exceeds the between-groups variability. Although these algorithms have similar performance in many cases, we can choose one of them, depending on the advantages and disadvantages of each. The EM algorithm has the advantage of providing strongly consistent estimators (Redner and Walker, 1984), with asymptotic normal distribution, while the estimators obtained with the dynamic clusters algorithm are not convergent. On the other hand, dynamic clusters algorithm converges rapidly to an optimum local, while EM algorithm may converge slowly to the optimum local.

Finally, we compare the solutions obtained in EM algorithm, dynamic clusters (DC) algorithm, PCCA and hierarchical clustering based on the linear correlation coefficient and complete linkage criterion. The solution obtained in the hierarchical clustering method was the initial solution in EM and DC algorithms and PCCA. For this purpose, we calculated the value of F- statistic for these solutions in the following cases: $n = 10$, $p = 20$, $\kappa = 30$, $n = 10$, $p = 30$, $\kappa = 10, 30$ and $n = 20$, $p = 30$, $\kappa = 30$. See Fig. 4.

The F-statistic always took the lowest value for the hierarchical clustering solution and consequently, this solution was the worst solution. The solutions obtained with EM algorithm, DC algorithm and PCCA were equal or very similar and so the value of F-statistic was equal or approximately equal, except when the concentration parameters are low or the components are badly separated.

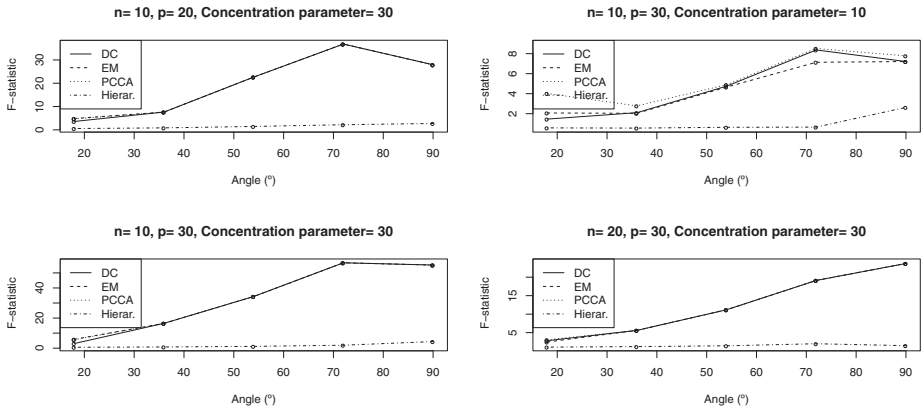

Figure 4. F-statistic for the solutions obtained with EM and DC algorithms, PCCA and Hierarchical Clustering method.

5. Example

We used aggregate data at firm level provided by *Associação Portuguesa de Bancos*. We considered 26 Portuguese banks with information on 20 variables that describe both the labor and product markets of the banking sector. These variables are: Share of workers by occupation: managerial (pf1), technical (pf2), administrative (pf3) and auxiliary (pf4); Share of workers with tenure: below 6 yr, (pten1), between 6 and 11 yr (pten2) and greater than 11 yr (pten3); Share of workers by main activity: commercial (pact1) and other (pact2); Net situation of the bank (NSeuros), Number of employees per bank (Nemp), Tax of return of the investment (ROA), Market share (Share), Age of the bank (Age), Wage, Profit per worker, real (Profit), Capital labor ratio (Kaplab), Profit per worker, non real (RBemp), Asset per worker (Asset), Sales of the bank per worker (Sales). Once that variables related with share of workers by occupation, main activity and seniority, sum to one, we selected, for analysis, only three variables describing the occupational categories, one variable describing the main activity and two variables describing the seniority categories.

Although in this case, the sample size is not large compared to the dimension of the sphere ($p = 17$ and $n = 26$) and also the concentration parameter estimate of the bipolar Watson distribution associated with the sample of variables is not very high ($\widehat{\kappa} = 20.459$), we obtained the chi-square Q–Q plot for the sample of variables, which is represented in Fig. 5 and suggests a mixture of three Watson components.

Since the EM algorithm and the dynamic clusters algorithm require the number of components of the mixture to be known, we applied the hierarchical clustering method based on the linear correlation coefficient and complete linkage criterion, which also suggested three components. The solution obtained was:

Group 1 = {Wage,RBemp,Asset,Sales,Kaplab}
Group 2 = {pf1,pf2,pten1,pten2,ROA,profit}
Group 3 = {NSeuros,Nemp,pact1,Share,Age,pf3}.

This solution was taken as the initial solution of the EM and dynamic clusters algorithms and the final solution obtained with EM algorithm ($EM(1)$) was

Group 1 = {Wage,RBemp,Asset,Sales}
Group 2 = {pf1,pf2,pf3,pten1,pten2,ROA,Age,profit}

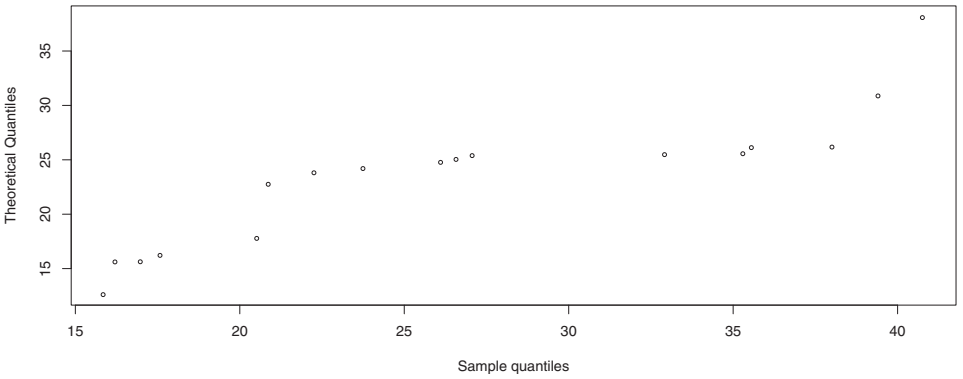

Figure 5. Chi-square Q–Q plot for the sample of variables.

Group 3 = {NSeuros,Nemp,pact1,Share,Kaplab}
and with dynamic clusters algorithm (DC) was
Group 1 = {Wage,RBemp,Kaplab,Asset,Sales}
Group 2 = {pf1,pf2,pf3,pten1,pten2,ROA, profit}
Group 3 = {NSleuros,Nemp,pact1,Share,Age}.

As the previous algorithms depend on the initial solution, we also considered several initial solutions chosen at random for EM algorithm, and we took the final solution that minimized the normalized entropy criterion, NEC (Celeux and Soromenho, 1996). The same solution was obtained using Akaike Information Criterion, AIC (Akaike, 1974), Bayesian Information Criterion, BIC (Schwarz, 1978) and Approximate Weight of Evidence criterion, AEW (Banfield and Raftery, 1993), that is, these criteria were minimized for the same solution. The solution obtained ($EM(2)$) was

Group 1 = {NSeuros,Nemp,Share}
Group 2 = {pf2,pf3}
Group 3 = {pf1, pact1,pten1,pten2,ROA,Age, Wage,RBemp,Kaplab,Profit,Asset,Sales} .

Table 1
Sizes and concentration parameter estimates of the groups, percentage of variance explained by the first principal component of each group, variability measures and F- statistic for $EM(1)$, DC and $EM(2)$ solutions

	EM (1)			DC			EM (2)		
Group	1	2	3	1	2	3	1	2	3
p_i	4	8	5	5	7	5	3	2	12
$\widehat{\kappa_i}$	42.51	24.91	39.55	30.96	26.45	39.70	159.98	583.23	21.32
% variance	70	47	68	58	50	68	92	98	37
Bet.-groups var.		151.05			124.68			396.08	
Wit.-groups var.		220.32			220.09			223.58	
F- statistic		4.80			3.97			12.40	

Table 2
Sizes and concentration parameter estimates of the groups, percentage of variance explained by the first principal component of each group, variability measures and F-statistic for the solutions obtained in hierarchical clustering and PCCA

	Hierarchical clustering			PCCA		
Group	1	2	3	1	2	3
p_i	5	6	6	4	6	7
$\widehat{\kappa}_i$	30.96	24.93	32.74	42.51	30.19	28.48
% variance	58.30	47.03	60.71	70.07	57.16	54.35
Between-groups var.		104.75			143.36	
Within-groups var.		220.97			219.53	
F-statistic		3.32			4.57	

Considering that clusters' sizes are really small, we did not apply the goodness-of–fit methods for the bipolar Watson distribution defined on the hypersphere proposed by Figueiredo and Gomes (2006b) to check whether the clusters of variables obtained in the algorithms come from bipolar Watson populations. For the same reason, we did not apply the uniformity tests.

Table 3
Linear correlations between the variables and the first principal component of the clusters of $EM(1)$, DC and $EM(2)$ solutions

	EM (1)			DC			EM (2)		
Variables/Group	1	2	3	1	2	3	1	2	3
Wage	**0.81**	—	—	**0.80**	—	—	—	—	**0.57**
RBemp	**0.88**	—	—	**0.86**	—	—	—	—	**0.69**
Asset	**0.76**	—	—	**0.75**	—	—	—	—	**0.76**
Sales	**0.89**	—	—	**0.90**	—	—	—	—	**0.86**
pf1	—	0.38	—	—	0.37	—	—	—	0.31
pf2	—	**0.88**	—	—	**0.89**	—	—	**−0.99**	—
pf3	—	**−0.86**	—	—	**−0.89**	—	—	**0.99**	—
pten1	—	**0.82**	—	—	**0.77**	—	—	—	0.75
pten2	—	**0.76**	—	—	**0.69**	—	—	—	0.66
ROA	—	0.45	—	—	0.55	—	—	—	0.10
Age	—	0.55	—	—	—	−0.58	—	—	−0.66
Profit	—	0.58	—	—	0.56	—	—	—	0.36
NSeuros	—	—	**0.89**	—	—	**−0.90**	**0.96**	—	—
Nemp	—	—	**0.97**	—	—	**−0.98**	**0.97**	—	—
pact1	—	—	0.72	—	—	−0.71	—	—	−0.73
Share	—	—	**0.90**	—	—	**−0.88**	**0.95**	—	—
Kaplab	—	—	0.57	0.49	—	—	—	—	0.38

Table 4
Linear correlations between the variables and the first principal component of the clusters obtained with Hierarchical Clustering method and PCCA

	Hierarchical clustering			PCCA		
Variables/Group	1	2	3	1	2	3
Wage	**0.80**	—	—	**0.81**	—	—
RBemp	**0.86**	—	—	**0.88**	—	—
Asset	**0.75**	—	—	**0.76**	—	—
Sales	**0.90**	—	—	**0.89**	—	—
pf1	—	0.29	—	—		−0.42
pf2	—	**0.80**	—	—	**0.89**	—
pf3	—		−0.56	—	**−0.86**	—
pten1	—	**0.80**	—	—	**0.77**	—
pten2	—	**0.70**	—	—	0.67	—
ROA	—	0.64	—	—	0.60	—
Age	—	—	−0.57	—	—	0.58
Profit	—	0.74	—	—	0.70	—
NSeuros	—	—	**−0.88**	—	—	**0.87**
Nemp	—	—	**−0.97**	—	—	**0.97**
pact1	—	—	−0.71	—	—	0.73
Share	—	—	**−0.88**	—	—	**0.85**
Kaplab	0.49	—	—	—	—	−0.58

From Table 1, taking into account the variability measures, or more precisely the F-statistic (12.4, 4.8 and 3.97 for $EM(2)$, $EM(1)$ and DC solutions, respectively) and the average percentage of variance explained by the first principal component of the groups (75.7%, 61.7% and 58.7% for $EM(2)$, $EM(1)$ and DC solutions, respectively), we conclude that the solutions obtained with EM algorithm are preferable to the solution obtained with dynamic clusters algorithm.

Finally, we compare $EM(1)$, $EM(2)$, and DC solutions with the hierarchical clustering solution, previously given, and with the solution obtained with PCCA, using the hierarchical clustering solution as initial solution. This last solution is given by

Group 1 = {Wage, RBemp,Asset, Sales}
Group 2 = {pf2,pf3,pten1,pten2,ROA,Profit}
Group 3 = {pf1,Age,Sleuros,Nemp,pact1,Share,Kaplab} .

From Table 2, comparing the solutions obtained in the hierarchical clustering and PCCA with the previous ones, in terms of F-statistic (3.33 and 4.57 for hierarchical clustering and PCCA, respectively) and average percentage of variance explained by the first principal components of groups (55.35% and 60.53% for the hierarchical clustering method and PCCA, respectively), we observe that the solution obtained with hierarchical clustering method has the worst performance and the solutions obtained with EM algorithm are those that have better performance.

In Tables 3 and 4, we indicate the linear correlations between the variables of each group and the first principal component of the group, for all solutions. Although the solutions

obtained are not equal, the variables quite correlated with the first principal component of the respective groups are clustered together in almost all solutions. In fact, NSeur, Nemp, Share are always together in a cluster, pf2 and pf3 are always together in another cluster (except for the solution obtained with hierarchical clustering method) and also, Asset, Sales, Wage, RBemp are always together in a cluster.

Thus, in general, one first component opposes technical to administrative functions of workers; other first component is associated with net situation, number of employees, and share of the banks, and the other first component is associated with assets per worker, profit per worker, nonreal, wage and sales of the bank per worker.

6. Conclusion

We proposed an approach for clustering of variables, based on the identification of a mixture with bipolar Watson components defined on the hypersphere, through EM and dynamic clusters algorithms.

Both algorithms gave the same solution for simulated data from a mixture of two bipolar Watson components, except for very small concentration parameters or very overlapped components. Concerning variability measures between-groups and within-groups, EM and dynamic clusters algorithms presented identical performance to PCCA and better performance than hierarchical clustering method for simulated and real data.

Acknowledgments

The authors thank Natália Monteiro from University of Minho, Portugal, for the data used in this work.

The authors also thank the helpful comments and suggestions given by the referees and associate editor of this journal.

Funding

This work is funded (or part-funded) by the ERDF European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT-Portuguese Foundation for Science and Technology within project FCOMP - 01-0124-FEDER-037281.

References

Akaike, H. (1974). A new look at statistical model identification. *IEEE Trans. Automat. Contr.* 19: 716–723.

Banerjee, A., Dhillon, I., Ghosh, J., SRA, S. (2005). Clustering on the unit hypersphere using von Mises–Fisher distributions. *Journal of Machine Learning Research* 6:1345–1382.

Banfield, J. D., Raftery, A. E. (1993). Model-Based Gaussian and non-Gaussian Clustering. *Biometrics* 49:803–821.

Carbonell, L., Izquierdo, L., Carbonell, I., Costell, E. (2008). Segmentation of food consumers according to their correlations with sensory attributes projected on preference spaces. *Food Quality and Preferences* 19:71–78.

Celeux, G., Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification* 13(2):195–212.

Dempster, A. P., Laird, N. M., 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.

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

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

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics* 29:751–760.

Escoufier Y., (1988). Beyond correspondence analysis. In: *Classification and Related Methods of Data Analysis* (H. H. Bock, ed), pp. 505–514. Elsevier B. V. (North-Holland).

Everitt, B. S. (1993). *Cluster Analysis*. London: Arnold.

Figueiredo, A., Gomes, P. (2003). Power of tests of uniformity defined on the hypersphere. *Communications in Statistics: Simulation and Computation* 22(1):87–94.

Figueiredo, A., Gomes, P. (2006a). Performance of the EM algorithm on the identification of a mixture of Watson distributions defined on the hypersphere. *REVSTAT–Statistical Journal* 4(2):19.

Figueiredo, A., Gomes, P. (2006b). Goodness-of-fit methods for the bipolar Watson distribution defined on the hypersphere. *Statistics and Probability Letters* 76:142–152.

Fisher, T., Lewis, T., Embleton, B. (1987). *Statistical Analysis of Spherical Data*. Cambridge: Cambridge University Press.

Gomes, P. (1987). *Distribution de Bingham sur la n-sphere: une nouvelle Approche de l' Analyse Factorielle*, Thèse D'État. Université des Sciences et Techniques du Languedoc-Montpellier.

Gomes, P., Figueiredo, A. (1999). A new probabilistic approach for the classification of normalized variables. *In Contributed Papers of the Bulletin of the 52nd Session of the International Statistical Institute* LVIII(Book 1):403–404.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educational Psychology* 24:417–441.

Hulshof, K., Wedel, M., Lowik, M., Kok, F., Kistemaker, C., Hermus, R., Hoor, F., Ockhuizen, Th (1992). Clustering of dietary variables and other lifestyle factors. *Journal of Epidemiology and Community Health* 46:417–424.

Huo, V. (1984). *Small Samples from Bingham Distributions*, PhD thesis, University of Minnesota.

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

Kiers, H. L. (1991). *Principal Components Analysis: In Optimal Clusters of Variables*. University of Groningen, pp. 1–26.

Mardia, K. V., Jupp, P. E. (2000). *Directional Statistics*, 2nd ed. Chichester: Wiley. bibitem Peel, D., Whiten, W. J., McLachlan, G. J. (2001). Fitting mixtures of Kent distributions to aid in joint set identification. *Journal of the American Statistical Association* 96:56–63.

Qannari, E. M., Vigneau, E., Luscan, P., Lefebvre, A. C., Vey, F. (1997). Clustering of variables: application in consumer and sensory studies. *Food Quality and Preference* 8(5/6): 423–428.

Redner, R., Walker, H. (1984). Mixture densities, Maximum-Likelihood and the EM algorithm. *SIAM Review* 26(2):195–237.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6:461–464.

Soffritti, G. (1999). Hierarchical clustering of variables: a comparison among strategies of analysis. *Communications in Statistics: Simulation and Computation* 28(4):977–999.

Vigneau, E., Qannari, E. M. (2002). Segmentation of consumers taking account of external data. A clustering of variables approach. *Food Quality and Preference* 13:515–521.

Vigneau, E., Qannari, E. M. (2003). Clustering of variables around latent components. *Communications in Statistics: Simulation and Computation* 32(4):1131–1150.

Vigneau, E., Qannari, E. M., Punter, P. H., Knoops, S. (2001). Segmentation of a panel of consumers using clustering of variables around latent directions of preference. *Food Quality and Preference* 12:359–363.