International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:7, No:10, 2013

Clustering of Variables Based On a Probabilistic Approach Defined on the Hypersphere

Paulo Gomes, Adelaide Figueiredo

Abstract**—**We consider *n* individuals described by *p* standardized variables, represented by points of the surface of the unit hypersphere S_{n-1}. For a previous choice of n individuals we suppose that the set of observables variables comes from a mixture of bipolar Watson distribution defined on the hypersphere. *EM* and Dynamic Clusters algorithms are used for identification of such mixture. We obtain estimates of parameters for each Watson component and then a partition of the set of variables into homogeneous groups of variables. Additionally we will present a factor analysis model where unobservable factors are just the maximum likelihood estimators of Watson directional parameters, exactly the first principal component of data matrix associated to each group previously identified. Such alternative model it will yield us to directly interpretable solutions (*simple structure*), avoiding factors rotations.

Keywords**—**Dynamic Clusters algorithm, *EM* algorithm, Factor analysis model, Hierarchical Clustering, Watson distribution.

I. Introduction

THERE is a large variety of hierarchical clustering methods that may be used to cluster either individuals or variables ([1]-[3]).

Considering that variables under study are previously standardized and then represented by points of the unit sphere in $\mathcal{R}^n$ (*n* individuals), we present a probabilistic approach for the classification of variables based on the identification of a mixture of Watson distributions. For the mixture identification, we use *EM* and dynamic clusters algorithms, which yield us a partition of the initial set of variables into K clusters of variables.

In classical approach the *p* variables are previously chosen and the *n* individuals are randomly selected from a population of individuals. In our approach the *n* individuals are previously considered and the *p* variables from each cluster (i=1, …, K) are randomly selected from a specific population of variables.

Each cluster of variables can be considered a random sample of variables from a specific Watson distribution. It means that our proposal give an innovative contribution for the so called *a priori* selection of variables problem ([4]-[7]).

We will evaluate clusters obtained by these algorithms, using measures of within-groups variability and between--groups variability.

It was shown that maximum likelihood estimate of directional parameter $\mathbf{u}_i$ of Watson distribution associated to cluster *i*, is just the first principal component of data matrix represented such cluster ([6]).

Such conclusion allows us to define an unobservable factor for each identified cluster and a factor analysis model, where factors are generally correlated and final solutions are directly interpretable avoiding factors rotation.

Paulo Gomes is with Statistics Portugal and Nova University of Lisbon, ISEGI, Portugal.

Adelaide Figueiredo is with School of Economics and LIAAD-INESC TEC, University of Porto, Portugal.

II. The Watson Distribution on the Hypersphere

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

$$f(\mathbf{x}) = \left\{ {}_1F\left(\frac{1}{2}, \frac{n}{2}, \kappa\right)\right\}^{-1} \exp\left\{\kappa(\mathbf{u}'\mathrm{x})^2\right\}, \mathbf{x} \in S_{n-1}, \mathbf{u} \in S_{n-1}, \kappa>0,$$

where ${}_1F_1\left(\frac{1}{2}, \frac{n}{2}, \kappa\right)$ is the confluent hypergeometric function defined by ${}_1F_1\left(\frac{1}{2}, \frac{n}{2}, \kappa\right) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)} \int_0^1 exp\left(\kappa t\right) t^{-1/2}(1-t)^{(n-3)/2} dt$.

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 around the principal axis $\mathbf{u}$.

If *x* comes from the bipolar Watson population $W_n(\mathrm{u}, \kappa)$, then for large κ (see [8], p. 236):

$$2\kappa\{1-(\mathbf{u}'\mathbf{x})^2\} \tilde{\sim} \chi^2_{n-1}, \kappa \to \infty$$

Let $\left[\mathbf{x}_1|\mathbf{x}_2|\cdots|\mathbf{x}_p\right]$ be a sample of variables from the bipolar Watson distribution $W_n(\mathbf{u}, \kappa)$. It can be shown ([6]), that maximum likelihood estimator of the parameter $\mathbf{u}$ is the eigenvector of the orientation matrix $\mathrm{T} = \sum_{i=1}^{p} \mathbf{x}_i \mathbf{x}_i'$ associated to 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 matrix representing such sample. Additionally, the maximum estimator of the parameter κ is the solution of the equation $Y(\kappa) = \frac{W}{p}$ where the function $Y(.)$ is defined by $Y(\kappa) = d ln\ {}_1F_1\left(\frac{1}{2}, \frac{n}{2}, \kappa\right)/d\kappa$ ([8]).

III. 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

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

where π_i, i =1, ..., K are the mixture proportions, $0 \leq \pi_i \leq 1, \forall_i, \sum_{i=1}^{K} \pi_i = 1; f(\mathbf{x}|\mathbf{u}_i, \boldsymbol{\kappa}_i)$ is the density function of the ith component of the mixture, then the density of $W_n(\mathbf{u_i}, \boldsymbol{\kappa}_i)$, and $\emptyset = (\mathbf{u}_1, \cdots, \mathbf{u}_k, \boldsymbol{\kappa}_1, \cdots, \boldsymbol{\kappa}_k, \pi_1, \cdots, \pi_k)$ is the parameter vector of the mixture.

For obtaining a partition of the set of variables $(\mathbf{x}_1|\boldsymbol{x}_2|\cdots|\mathbf{x}_p)$ into K groups of variables, we considered the *EM* algorithm and the Dynamic Clusters algorithm ([9]-[11]), To analyze the performance of the *EM* and dynamic clusters algorithms, we can use the variability measures between groups and within-groups ([12]).

The between-groups variability measure is defined by

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

and the within-groups variability measure defined by

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

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

IV. An Alternative Factor Analysis Model

The basic common factor model is usually expressed as

$$X = FA' + U,$$

where X is the p-dimensional vector of observable random variables (in our context standardized variables), F is a k-dimensional vector of unobservable variables called *common factors* , U a p-dimensional vector of unobservable variables called *unique factors* and A is pxK matrix of unknown constants called *factor loadings*. There are p unique factors and it is generally assumed that the unique part of each variable is uncorrelated with each other or with their common part. Generally it is also assumed that the factors themselves are uncorrelated. However, for interpretable purposes it is often necessary a factor rotation (orthogonal on even oblique rotation) in order to obtain a *simple structure*.

In such model we suppose the multinormality of observations and usually we estimate factors loadings by maximum likelihood methods, yielding us to the representation of variables. In a second step we estimate the common factors by regression yielding us to the representation of individuals. The two steps procedure is preceded by a choice for the number of factors in almost the cases justified by auxiliary information.

On approach consider a new factor analysis model named *common and residual factor analysis model*,

$$X = FA' + U,$$

where F is the factor matrix obtained by the K first principal component of each cluster identified, loading matrix A is obtained by regression and U represents the residual matrix, X-FA'. In our proposal the factors F can be or not correlated, so we can achieve directly interpretable solutions avoiding rotation factor.

V. Illustration

We used aggregate data at firm level provided by Portuguese Bank Association. We considered 26 Portuguese banks with information on 17 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), Share of workers with tenure: below 6 years (pten1) and between 6 and 11 years (pten2); Share of workers by commercial activity (pact1); 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) and Sales of the bank per worker (Sales).

A factor analysis solution was obtained, considering the previous assumption that each group of variables is a random sample of Watson population with specified parameters estimated by maximum likelihood method.

Since the *EM* algorithm and Dynamic Clusters algorithm require the number of components of the mixture, we applied the hierarchical clustering method based on the linear correlation coefficient and complete linkage criterion, which suggested three components. A Q-Q plot for the sample of variables represented in Fig. 1 suggests a mixture of three Watson components.

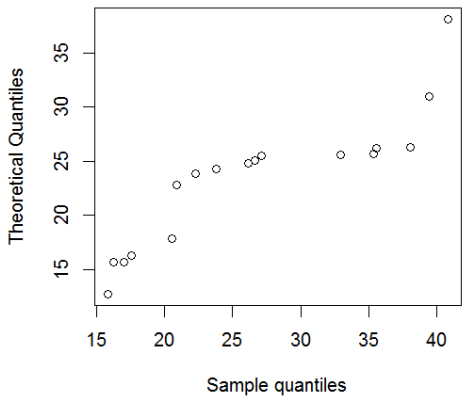

Fig. 1 Chi-square Q-Q plot for the sample of variables

Such components were obtained by *EM* algorithm and the final solution was
Group 1: { Wage, RBemp, Asset, Sales}
Group 2: { pf1, pf2, pf3, pten1, pten2, ROA, Age, profit}
Group 3: { NSeuros, Nemp, pact1, Share, Kaplab}
and matrix *F* is obtained doing

$$F=\begin{pmatrix} \textit{First principal component of group } 1 & \textit{First principal component of group } 2 & \textit{First principal component of group } 3 \end{pmatrix}$$

In Fig. 2, we can see the representation of variables from group 1 in relation to first principal component of this group.

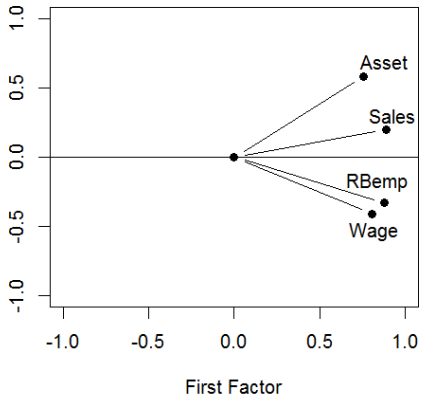

Fig. 2 Representation of variables from group 1 on first axis explaining 70.9% of total inertia of such group

The linear correlation coefficients between 3-factors and initial standardized variables (Table I) suggest a *simple structure* outcome where factors are directly interpretable, avoiding factor rotations.

TABLE I
LINEAR CORRELATIONS BETWEEN THE VARIABLES AND THE FIRST PRINCIPAL COMPONENT OF THE GROUPS

Variables	Group 1	Group 2	Group 3
Wage	0.81	--	--
RBemp	0.88	--	--
Asset	0.76	--	--
Sales	0.89	--	--
pf1	-	0.38	--
Pf2	--	0.88	--
Pf3	--	0.86	---
pten1	--	0.82	--
pten2	--	0.76	--
ROA	--	0.45	--
Age	--	0.55	--
Profit	--	0.58	--
NSeuros	--	--	0.89
Nemp	--	--	0.97
Pact1	--	--	0.72
share	--	-	0.90
kaplab	--	--	0.57

ACKNOWLEDGMENT

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-022701.

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

REFERENCES

[1] B. S. Everitt. *Cluster Analysis*, London: Arnold, 1993.
[2] E. M. Qannari, E. Vigneau, P. Luscan, A. C. Lefebvre and F. Vey. Clustering of variables: application in consumer and sensory studies. *Food Quality and Preference*, 8, 5/6, 423-428, 1997.
[3] E. Vigneau and E. M. Qannari. Clustering of variables around latent components. *Communications in Statistics - Simulation and Computation*, 32, 4, pp. 1131-1150, 2003.
[4] H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educational Psychology*, 24, pp. 417-441, 1933.
[5] Y. Escoufier. Le traitement des variables vectorielles. *Biometrics*, 29, pp. 751-760, 1973.
[6] P. Gomes. *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, 1987.
[7] A. Figueiredo. *Classificação de variáveis no contexto de um modelo probabilístico definido na n-esfera.* Tese de Doutoramento em Estatística e Investigação Operacional na especialidade de Estatística Experimental e Análise de Dados, Faculdade de Ciências, Universidade de Lisboa, 2000.
[8] K. Mardia and P. E. Jupp. *Directional Statistics*, 2nd edition, Wiley: Chichester, 2000.
[9] A. Figueiredo and P. Gomes. Power of tests of uniformity defined on the hypersphere. *Communications in Statistics -Simulation and Computation*, 22, 1, pp. 87-94, 2003.
[10] A. Figueiredo and P. Gomes. Performance of the *EM* algorithm on the identification of a mixture of Watson distributions defined on the hypersphere. *REVSTAT-Statistical Journal*, 4, 2, p. 19, 2006,
[11] A. Figueiredo and P. Gomes. Goodness-of-fit methods for the bipolar Watson distribution defined on the hypersphere. *Statistics and Probability Letters*, 76, pp. 142-152, 2006.
[12] P. Gomes and A. Figueiredo. "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*, vol. LVIII, Book 1, pp. 403-404, 1999.