Hierarchical Clustering and Dimensionality Reduction for Big Data

Clustering e Riduzione Dimensionale Gerarchici per Dati di Grandi Dimensioni

Carlo Cavicchia, Maurizio Vichi and Giorgia Zaccaria

Abstract The development of new technologies and methods of data collection produces the necessity to summarise the large quantity of information that is available. Usually, we face a data matrix $\mathbf{X}$ of size $(n \times J)$, corresponding to n statistical units and J quantitative variables, where n and J are very large. Clustering is the analysis which identifies homogeneous clusters of units, thus it might be meant as a way to reduce their dimension. Dimensionality reduction techniques are methods to obtain latent dimensions (less than manifest variables), so they reduce the dimensionality of the variables space. In this paper, we apply *Double Hierarchical Parsimonious Means Clustering* [2] in order to get a simultaneous hierarchical parsimonious clustering of units - aggregated around centroids - and dimensionality reduction of variables - aggregated around components - on *Asia-Europe Meeting* (ASEM) data set. The model is estimated by using the LS method and an efficient coordinate descent algorithm is given. The goodness of fit of the double hierarchical parsimonious trees can be computed to assess the quality of the two hierarchical partitions.

Key words: clustering, dimensionality reduction, big data, hierarchy.

Carlo Cavicchia
University of Rome La Sapienza, Piazzale Aldo Moro 5, Rome,
e-mail: carlo.cavicchia@uniroma1.it

Maurizio Vichi
University of Rome La Sapienza, Piazzale Aldo Moro 5, Rome,
e-mail: maurizio.vichi@uniroma1.it

Giorgia Zaccaria
University of Rome La Sapienza, Piazzale Aldo Moro 5, Rome,
e-mail: giorgia.zaccaria@uniroma1.it

1 Introduction

In recent years, with the data revolution and the use of new technologies, phenomena are frequently described by a huge quantity of information useful for making strategical decisions. A priority for policymakers is having simple statistical methods useful to synthesise all the available information. Different levels of synthesis are required by stakeholders in order to describe properly different phenomena.

Cluster analysis is a field of study which tries to identify homogeneous clusters of units. Hierarchical clustering methods are well-known and widely used for producing a hierarchy of statistical units, clustered in $(n-1)$ nested partitions. Most of these methods take into account an objects-by-objects dissimilarity matrix, by setting a priori the kind of metrics and linkage in order to measure and update the distance between items, respectively. These clustering methods are heuristic and they do not underpin a model for the dissimilarity data or an objective function that can be optimised. [7] proposed a method that extends *K-means* to the case of hierarchical clustering estimating the objective function via least squares.

Dimensionality reduction methods (e.g. Principal Component Analysis (PCA) and Factor Analysis (FA)) are usually implemented to obtain a straightforward interpretation of the data. These methodologies are sometimes not able to get the real structure of the data and their relationships, i.e. a hierarchical correlation structure.

[3] proposed a hierarchical extension of *Disjoint Principal Component* [4] in order to build composite indicators.

In this paper, we apply the *Double Hierarchical Parsimonious Means Clustering* [2] in order to get a simultaneous hierarchical partitions of units - represented by centroids - and of variables - represented by components - on the *Asia-Europe Meeting* (ASEM) dataset. The aim of this research is to build a composite indicator for ASEM taking into account a hierarchical set of nested partitions of countries.

The paper is organised as follows. In Section 2 the model is presented and in Section 3 it is applied on the ASEM dataset. Finally, in Section 4 some conclusions end the paper.

2 Methodology

In the era of big data, the need to synthesise information is even more crucial. Clustering and dimensionality reduction are considered in order to synthesise large quantity of data. Both for units and for variables, it is worthy to identify clusters or classes of objects that represent homogeneous features. On one hand, the huge amount of data holds much more information than previously and millions of statistical units are available; on the other hand, it becomes necessary to understand if this information might be transformed into statistical knowledge.

The syntheses of objects and variables are usually achieved according to sequential or simultaneous approaches, as the *tandem analysis* or the *Clustering and Disjoint Principal Component Analysis* (CDPCA) proposed by [8], respectively. Many authors have criticised the former method since it brings about a masking of the taxonomic information of the data. However, the simultaneous approach does not allow to inspect the hierarchical relationships between dimensions of a multidimensional phenomenon, whenever they exist.

In the specialised literature, many methodologies have been developed to simplify the complete hierarchies ([5]) and to build parsimonious trees ([6]), both for units and variables. In case of big data, the parsimony property is fundamental to interpret the results.

[2] studied a new hierarchical simultaneous model-based approach to cluster objects and to identify new latent concepts, each one associated to a group of variables. The methodology is based upon the CDPCA, starting from a fixed number of clusters K and components Q and reducing these values by one at each hierarchical level. Formally,

$$\mathbf{X} = \mathbf{U}_k \mathbf{M}_{kq} \mathbf{V}'_q \mathbf{B}_q + \mathbf{E}_k \qquad \forall k = K, ..., 1, q = Q, ..., M, \tag{1}$$

where $\mathbf{X}$ is a $(n \times J)$ data matrix - with n statistical units and J quantitative variables- , $\mathbf{U}_k$ and $\mathbf{V}_q$ are the membership matrices for units and variables, respectively, $\mathbf{B}_q$ is the matrix of weights and $\mathbf{M}_{kq}$ is the centroids matrix in the reduced space.

The model (1) is subject to the classical constraints on membership matrices for partitioning and, according to [8], on the reparametrization of the loading matrix $\mathbf{A}_q$ into the product of two matrices, i.e. $\mathbf{B}_q$ and $\mathbf{V}_q$. Furthermore, a constrain on nested partitions has been added to the model (1).

Eq.(1) represents the reflective part of the model with $Q - M + 1$ hierarchical levels. M identifies the number of the bottom-up level of the hierarchy at which the model becomes formative, i.e. the M components are merged into a unique measure of synthesis, and it is selected according to a statistical test.

The model is estimated in a least-squares semi-parametric framework in which a quadratic loss function is minimised and it is implemented with a coordinate descent algorithm. The latter is efficient in real applications.

3 Application: ASEM Index - International Sustainable Connectivity

Asia-Europe Meeting (ASEM) Sustainable Connectivity Index is aimed at measuring connectivity among countries, people and societies in an economic sense (e.g. transport links, energy, trade,...) and in a social sense (e.g. migration, linkage, cultural connection,...). The data comprises 51 countries[1] - 30 European and 21 Asian - and 49 indicators.

[1] Source: [1].

The indicators are grouped in two indexes, *Connectivity* and *Sustainability*, with 5 and 3 dimensions respectively, as shown in Figure 1.

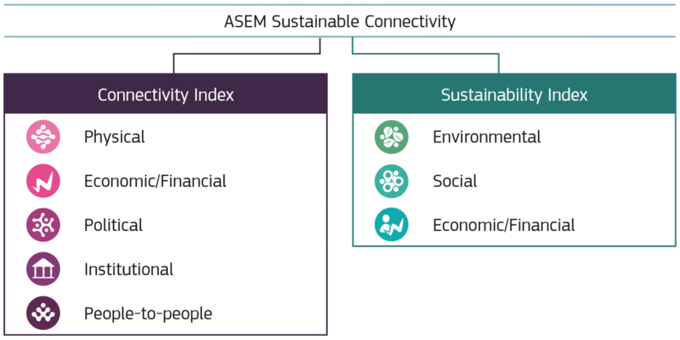

Fig. 1 ASEM Sustainable Connectivity Conceptual Framework[1].

The methodology described in Section 2 has been implemented on this data set for clustering countries and build a composite indicator, i.e. a measure of synthesis, from the 49 indicators. With respect to the construction of the variables hierarchy, two research approaches are defined: confirmatory and exploratory. In the former, the 8 dimensions are fixed (Figure 1), i.e. the partition of the manifest variables at the eighth hierarchical level is constrained. In the latter, all the constraints are relaxed and the initial parsimonious number of variable groups is pinpointed according to the unidimensionality of the components. In both cases, the optimal solution of the model corresponds to 6 clusters of statistical units. Before analysed the results and in order to measure the internal reliability of the two proposed indices, the Cronbach's α has been computed: the Connectivity index has $\alpha = 0.94$ and the Sustainability one has $\alpha = 0.37$. Thus, the former seems to be very consistent, whereas the latter turns out to be not reliable.

In the confirmatory approach, the model (1) underpins the theoretical double composite indicators approach with $M = 2$, whose corresponding components are

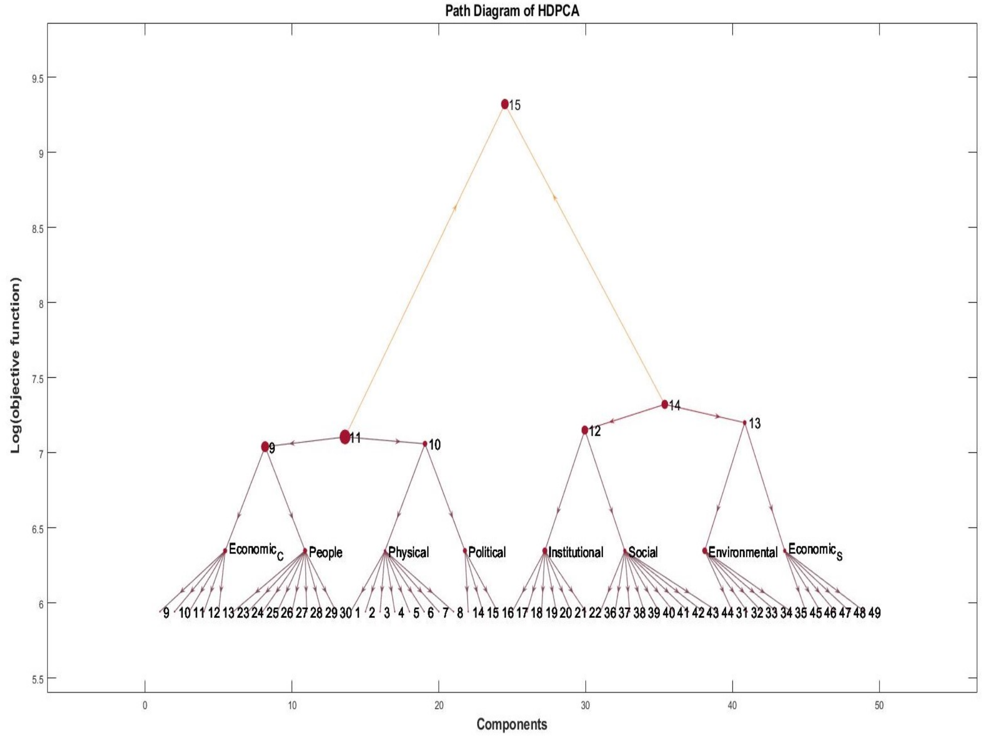

Fig. 2 Conrmatory Approach on ASEM Data Set. Variables Hierarchy.

merged together in a formative way. The partition obtained at this hierarchical level is equal to the one proposed, except for the dimension *Institutional* - which belongs to the Sustainability index group according to the model (1) - as shown in Figure 2. The Cronbach's α shows improvements for this group, passing from 0.37 to 0.67. Moreover, only one dimension is not unidimensional (*Political*) - the unidimensionality is assessed according to the magnitude of the "restricted"[2] covariance matrix second eigenvalue.

In the exploratory approach, the model (1) pinpoints $Q = 3$ unidimensional components and it underpins again the theoretical model identifying $M = 2$. The three groups are composed by the following variables of the theoretical domains:

- $4/8$ *Physical*, $5/5$ *Economic/Financial* (Connectivity), $1/3$ *Political*, $7/8$ *People-to-people*.
- $4/8$ *Physical*, $2/3$ *Political*, $6/6$ *Institutional*, $1/8$ *People-to-people*, $1/5$ *Environmental*, $9/9$ *Social*.
- $4/5$ *Environmental*, $5/5$ *Economic/Financial* (Sustainability).

[2] It refers to the manifest variables of the data matrix associated to a component.

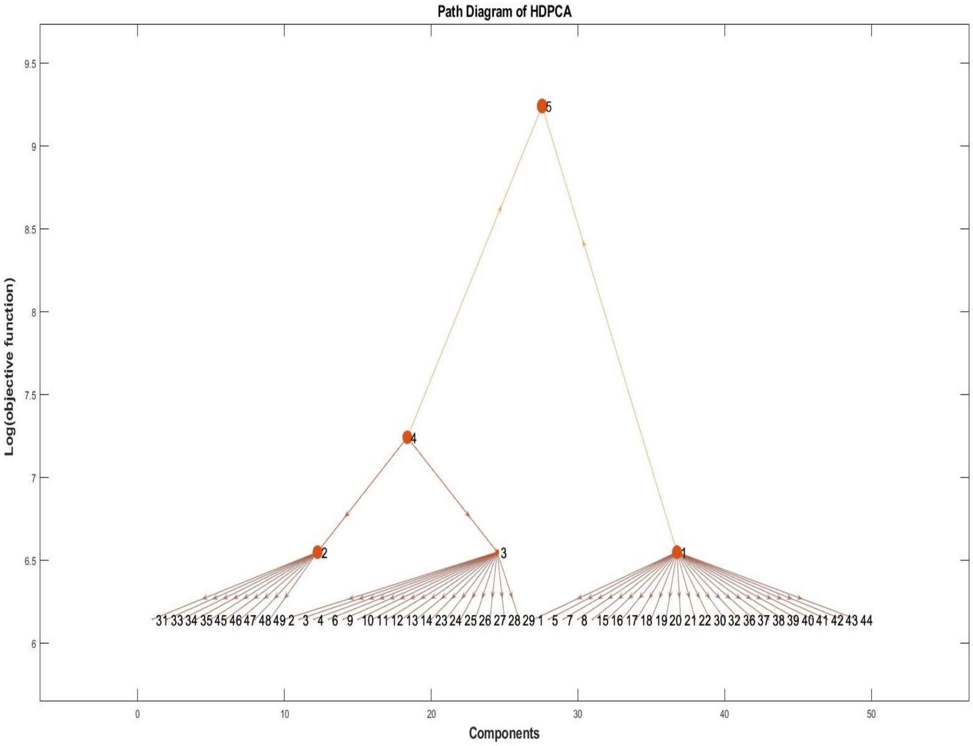

Fig. 3 Exploratory Approach on ASEM Data Set. Variables Hierarchy.

The components related to the aforementioned groups seem to be coherent with the confirmatory results. Indeed, the first group is mainly composed by three domains of the Connectivity index, the third by two dimensions of the Sustainability index, and the second one puts together the *Institutional* domain with many variables pertaining to the Connectivity index. The Cronbach's α are 0.96, 0.78 and 0.94, respectively.

The clustering of the statistical units returns the same results both for the confirmatory and the exploratory approach. The optimal number of clusters turns out to be equal to 6, according to the best solution of the model (1) as represented in Figure 4 by the red line.

The six clusters are pinpointed by the following countries:

1. Austria, Belgium, Denmark, Finland, Ireland, Luxembourg, Netherlands, Norway, Sweden, Switzerland, Australia, New Zealand, Singapore.
2. Brunei Darussalam, Kazakhstan, Mongolia, Russian Federation.
3. Bulgaria, Croatia, Cyprus, Czech Republic, Estonia, Greece Hungary, Latvia, Lithuania, Malta, Poland, Portugal, Romania, Slovakia, Slovenia.
4. Italy, Spain, Japan, Korea.
5. France, Germany, United Kingdom, China.

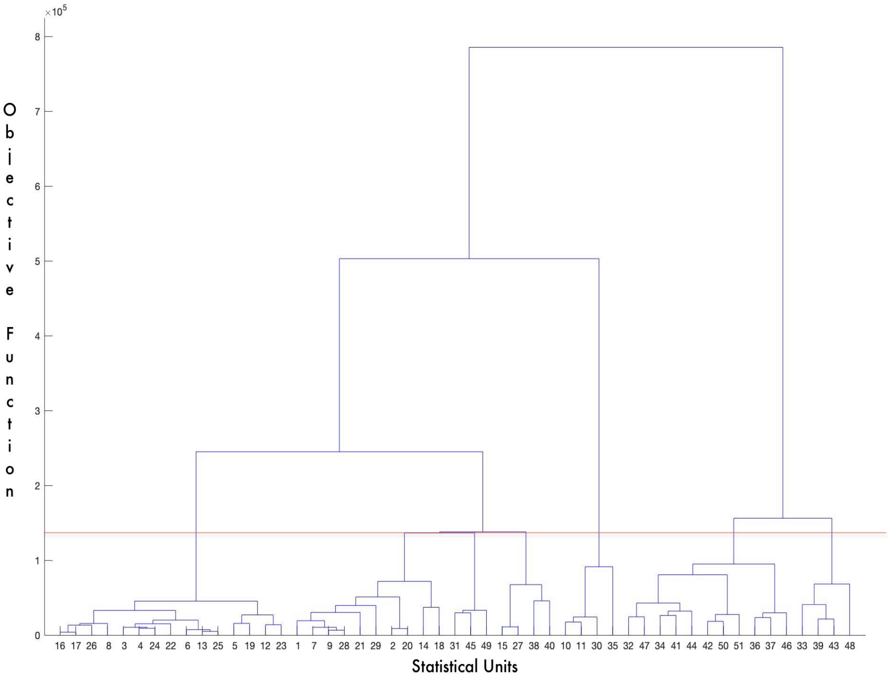

Fig. 4 Unit Clustering of the ASEM Data Set. Partition in 6 clusters of units (red line).

6. Bangladesh, Cambodia, India, Indonesia, Lao PDR, Malaysia, Myanmar, Pakistan, Philippines, Thailand, Vietnam.

The hierarchical levels from the sixth, i.e. that one with 6 clusters of countries, upwards are firstly defined by the aggregations of some of the European countries - 1 and 5 - and the Asian countries - 2 and 6. Then, the former and the remaining clusters of the European countries groups are lumped together, coherently with their geo-political distribution.

4 Conclusions

Clustering and dimensionality reduction are widely used analyses and their applications might be in several areas. Both for statistical units and for variables, the process of reduction often has a hierarchically nested shape which can be represented with a graphical configuration of a tree.

The hierarchy-shape is perfect to represent multidimensional concepts, starting from more specific ones up to the most general one, and to understand the underlining interconnections. A hierarchical approach permits to stop the analysis at the level the researcher considers optimal and it allows the researcher to investigate all the interconnections among items (i.e., variables and/or statistical units). In this paper, we applied the model proposed by [2] in order to get the optimal number of clusters and the optimal dimensions of the *Asia-Europe Meeting* (ASEM) data. The presence of an objective function permitted us to test a given theory and then to propose a new framework given by the study of the relations behind the data. The result is a deep study of the structure of the data and a reduced data matrix in both the dimensions (i.e, variables and/or statistical units).

References

1. Becker, W., Dominguez-Torreiro, M., Neves, A.R., Tacao Moura, C. J., Saisana, M.: Exploring ASEM Sustainable Connectivity What brings Asia and Europe together?, ISBN 978-92-79-99726-6, doi:10.2760/738153, PUBSY JRC112998 (2019)
2. Cavicchia, C., Vichi, M., Zaccaria, G.: Double Hierarchical Parsimonious Means Clustering. Unpublished manuscript
3. Cavicchia, C., Vichi, M., Zaccaria, G.: Hierarchical Disjoint Principal Component Analysis. Unpublished manuscript
4. Ferrara, C. and Martella, F. and Vichi, M.: Dimensions of Well-Being and Their Statistical Measurements. Studies in theoretical and applied statistics. 85-99 (2016)
5. Gordon, A. D.: Classification. Chapman & Hall/CRC, 2^{nd} Edition (1999)
6. Hartigan, J. A.: Representation of Similarity Matrices by Trees. Journal of the American Statistical Association. **62:320**, 1140–1158 (1967)
7. Vichi, M., Groenen, P.K., Cavicchia, C.: Hierarchical Means Clustering. Unpublished manuscript
8. Vichi, M., Saporta, G.: Clustering and Disjoint Principal Component Analysis. Computational Statistics and Data Analysis. **53**, 3194–3208 (2009)