Behaviormetrika (2019) 46:239–241
https://doi.org/10.1007/s41237-019-00092-6

PREFACE

Special feature: dimension reduction and cluster analysis

Michel van de Velden[1] · Alfonso Iodice D'Enza[2] · Michio Yamamoto[3]

Published online: 13 September 2019

Dimension reduction and cluster analysis have a long history in multivariate data analysis. Dimension reduction methods typically concern themselves with a reduction in the variable space through either selection of variables or the construction of new variables as combinations of the original ones. Cluster analysis aims to detect groups of similar observations thus reducing the row space. Although these methods typically consider different objective functions, the ultimate goals, e.g., detecting and summarizing relevant properties and relationships in the data, are typically similar if not identical. Consequently, dimension reduction and cluster analysis are frequently combined.

One way to combine dimension reduction and cluster analysis is to perform the analyses sequentially. In particular, a common procedure is to first perform dimension reduction and then apply cluster analysis to the reduced data. However, Bock (1987), Van Buuren and Heiser (1989) and De Soete and Carroll (1994), already observed that such sequential analyses may not be optimal due to the differences between the objective functions corresponding to the dimension reduction and cluster analysis parts. An alternative to such sequential methods, is to formulate an objective that incorporates the two, dimension reduction and clustering, jointly. We refer to such methods as joint dimension reduction and clustering and they are the topic of this special issue.

Recently, there appears to be an increased interest in joint methods. See, for example, (Hwang et al. 2006; Yamamoto and Hwang 2014, 2017; Vichi and Kiers 2001; van de Velden et al. 2017). In this special issue, we present several contributions covering various aspects of joint dimension reduction and cluster analysis methods.

The five papers comprising this issue cover several theoretical and applied aspects of cluster analysis and dimension reduction. In the first paper of the

✉ Michel van de Velden
vandevelden@ese.eur.nl

[1] Econometric Institute, Erasmus University Rotterdam, Rotterdam, The Netherlands

[2] Department of Political Sciences, Universit`a degli Studi di Napoli Federico II, Naples, Italy

[3] Department of Human Ecology, Graduate School of Environmental and Life Science, Okayama University, Okayama, Japan

special issue, Vichi, Vicari and Kiers, a comprehensive framework is given covering various variants for joint dimension reduction and clustering methods. The final method the authors propose, called CDR: Clustering and Dimension Reduction, allows a simultaneous dimension reduction and cluster analysis of data consisting of both qualitative (nominal and ordinal) and quantitative variables.

The contribution by Durieux and Wildemans, gives a more applied view of the special issue's topic. In particular, the paper is concerned with the analysis of fMRI data. As such data are typically three-way and of very high dimensionality, classical (two-way) clustering methods are inadequate. A two-step method combining independent component analysis and Ward's hierarchical clustering is proposed that outperforms alternatives such as a combined principal component analysis and cluster analysis approach.

In multivariate data analysis, the data can have class information about subjects and/or variables, which is called external information. When the external information exists, it is more efficient to take it into consideration in the data analysis. Constrained principal component analysis (CPCA) is one of the methods that deals with the external information efficiently (Takane and Shibayama 1991; Takane et al. 1995; Takane and Hunter 2001; Hwang and Takane 2002). In their contribution to this special issue, Yamagishi, Tanioka, and Yadohisa develop constrained nonmetric principal component analysis (CNPCA) that can deal with multivariate categorical data when class information of both subjects and variables exists. Their method can be considered as an extension of CPCA to categorical data using optimal scaling which is a well-established technique of quantification (see, e.g., Takane et al. 1979). A numerical study shows that the proposed method can have a better fit to the data, than the existing method. In addition, the analysis of purchase data shows that the proposed method can provide new insights into the data.

Preference rankings, expressed by a set of judges over a set of alternatives, are rather special data structures: ad-hoc statistical methods have been proposed to either describe the ranking structure or model the ranking process and the population of judges (see, e.g., Marden 2014). When it is fair to assume heterogeneity among the judges, the analysis typically aims to identify homogeneous subpopulations of judges. D'Ambrosio and Heiser propose a soft clustering algorithm for preference rankings that follows a probabilistic clustering approach (Ben-Israel and Iyigun 2008): their so-called K-median cluster component analysis is evaluated on both real and synthetic data and compared to other clustering methods for ranking data.

The assessment of the quality of a clustering solution is an important and difficult problem in cluster analysis. In fact, comparing different clustering solutions using a single overall index, is already a complex matter. In their contribution, van der Hoef and Warrens provide a by-cluster decomposition of a class of normalizations of mutual information. The authors show that the overall index corresponds to a summary statistic of information related to individual clusters and, therefore, recommend to use individual cluster measures rather than overall measures for improved interpretability. A further motivation for using individual rather than overall measures, is the fact that the latter tend to be heavily affected by cluster size imbalance.

We believe that the five papers of this special issue provide a broad and varied view on several topics related to joint dimension reduction and cluster analysis and we expect them to generate more high quality research into this area.

References

Ben-Israel A, Iyigun C (2008) Probabilistic d-clustering. J Classif 25(1):5

Bock H (1987) On the interface between cluster analysis, principal component analysis, and multidimensional scaling. In: Gupta A, Bozdogan H (eds) Multivariate statistical modeling and data analysis. Springer, Berlin, pp 17–34

De Soete G, Carroll JD (1994) K-means clustering in a low-dimensional euclidean space. In: Diday E, Lechevallier Y, Schader M, Bertrand P, Burtschy B (eds) New approaches in classification and data analysis. Springer, Berlin, pp 212–219

Hwang H, Takane Y (2002) Generalized constrained multiple correspondence analysis. Psychometrika 67(2):211–224. https://doi.org/10.1007/BF02294843

Hwang H, Dillon WR, Takane Y (2006) An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents. Psychometrika 71(1):161–171

Marden JI (2014) Analyzing and modeling rank data. Chapman and Hall/CRC, Boca Raton

Takane Y, Hunter MA (2001) Constrained principal component analysis: a comprehensive theory. Appl Algebra Eng Commun Comput 12(5):391–419. https://doi.org/10.1007/s002000100081

Takane Y, Shibayama T (1991) Principal component analysis with external information on both subjects and variables. Psychometrika 56(1):97–120. https://doi.org/10.1007/BF02294589

Takane Y, Kiers HAL, de Leeuw J (1995) Component analysis with different sets of constraints on different dimensions. Psychometrika 60(2):259–280. https://doi.org/10.1007/BF02301416

Takane Y, Young FW, de Leeuw J (1979) Nonmetric common factor analysis: an alternating least squares method with optimal scaling features. Behaviormetrika 6(6):45–56

Van Buuren S, Heiser WJ (1989) Clustering n objects into k groups under optimal scaling of variables. Psychometrika 54(4):699–706

van de Velden M, Iodice D'Enza A, Palumbo F (2017) Cluster correspondence analysis. Psychometrika 82(1):158–185

Vichi M, Kiers H (2001) Factorial *k*-means analysis for two-way data. Comput Stat Data Anal 37(1):49–64

Yamamoto M, Hwang H (2014) A general formulation of cluster analysis with dimension reduction and subspace separation. Behaviormetrika 41(1):115–129

Yamamoto M, Hwang H (2017) Dimension-reduced clustering of functional data via subspace separation. J Classif 34(2):294–326