Computational Statistics and Data Analysis 54 (2010) 1858–1871

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Factorial and reduced K-means reconsidered

Marieke E. Timmerman[a,*], Eva Ceulemans[b], Henk A.L. Kiers[a], Maurizio Vichi[c]

[a] *Heymans Institute of Psychology, University of Groningen, Grote Kruisstraat 2/1, 9712TS Groningen, The Netherlands*

[b] *Department of Educational Sciences, Katholieke Universiteit Leuven, Andreas Vesaliusstraat 2, B-3000 Leuven, Belgium*

[c] *Dipartimento di Statistica, Probabilità e Statistiche Applicate, Sapienza Università di Roma, P. le A. Moro 5, I-00185 Rome, Italy*

ARTICLE INFO

Article history:
Received 8 September 2008
Received in revised form 12 February 2010
Accepted 13 February 2010
Available online 19 February 2010

Keywords:
Cluster analysis
Factorial model
K-means
Dimensional reduction

ABSTRACT

Factorial K-means analysis (FKM) and Reduced K-means analysis (RKM) are clustering methods that aim at simultaneously achieving a clustering of the objects and a dimension reduction of the variables. Because a comprehensive comparison between FKM and RKM is lacking in the literature so far, a theoretical and simulation-based comparison between FKM and RKM is provided. It is shown theoretically how FKM's versus RKM's performances are affected by the presence of residuals within the clustering subspace and/or within its orthocomplement in the observed data. The simulation study confirmed that for both FKM and RKM, the cluster membership recovery generally deteriorates with increasing amount of overlap between clusters. Furthermore, the conjectures were confirmed that for FKM the subspace recovery deteriorates with increasing relative sizes of subspace residuals compared to the complement residuals, and that the reverse holds for RKM. As such, FKM and RKM complement each other. When the majority of the variables reflect the clustering structure, and/or standardized variables are being analyzed, RKM can be expected to perform reasonably well. However, because both RKM and FKM may suffer from subspace and membership recovery problems, it is essential to critically evaluate their solutions on the basis of the content of the clustering problem at hand.

1. Introduction

Cluster analysis aims at assigning a number of objects to a limited number of homogeneous classes. This is often done on the basis of the objects' scores on multiple variables. The inclusion of variables in a cluster analysis that hardly reflect, or even do not reflect the clustering structure may hinder or even completely obscure the recovery of the underlying cluster structure (e.g. Milligan, 1996). To deal with those problems, various approaches can be taken. One may use variable selection (e.g. Steinley and Brusco, 2008) or variable weighting (e.g. Milligan and Cooper, 1988), implying that some variables are discarded from the cluster analysis or are given less weight. An alternative is the subspace clustering approach, which rests on the assumption that the cluster centroids are located in a subspace of the variables.

A relatively easy subspace clustering approach, which does not require distributional assumptions, uses component analysis. The first attempt in this direction was a two-step procedure, where a principal component analysis was followed by a cluster analysis. After various warnings against this so-called tandem clustering (e.g. Arabie and Hubert, 1994), De Soete and Carroll (1994) proposed Reduced K-means analysis (RKM), which appeared to equal the earlier proposed Projection Pursuit Clustering (Bock, 1987). RKM simultaneously searches for a clustering of the objects, based on the K-means criterion (MacQueen, 1967), and a dimension reduction of the variables, based on component analysis. The notion that RKM may fail

[*] Corresponding author. Tel.: +31 50 3636255; fax: +31 50 3636304.
E-mail address: m.e.timmerman@rug.nl (M.E. Timmerman).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.02.009

to recover the clustering of the objects when the data contain much variance in directions orthogonal to the subspace of the data in which the clusters reside, led to the proposal of Factorial K-means analysis (FKM Vichi and Kiers, 2001).

A comprehensive comparison between FKM and RKM is lacking, both theoretically and empirically. After all, the simulation study that Vichi and Kiers (2001) conducted to support their claims about the superior performance of FKM over RKM consisted of the analysis of a single simulated data set only. This means that it is unknown when RKM and/or FKM would yield good insight into the cluster structure in empirical data, and hence when to use RKM or FKM. The purpose of the present article is to clarify this issue, so that FKM and RKM can be sensibly used in practice. It will be shown that FKM and RKM serve different goals. Therefore FKM and RKM complement each other: FKM is of use when RKM fails, and vice versa.

The remainder of the paper is organized as follows. Section 2 recapitulates the RKM and FKM models. Section 3 provides a theoretical comparison of the performance of RKM and FKM. Section 4 presents a simulation study to examine the conjectures from Section 3, and to evaluate the performance of RKM and FKM in various conditions. The use of FKM and RKM is illustrated with an empirical example in Section 5. The paper closes with a discussion of the reported findings (Section 6).

2. Factorial and reduced K-means

2.1. Notation

The following notation is adopted in this paper:

I	Number of objects, indexed $i = 1, \dots, I$.
J	Number of variables, indexed $j = 1, \dots, J$.
C	Number of clusters, indexed $c = 1, \dots, C$.
Q	Number of components to which the variables are reduced, indexed $q = 1, \dots, Q$.
$\mathbf{X}$	An $I \times J$ matrix containing the observed scores of I objects on J variables. Variables are supposed to be centered. In the case they have different units of measurements they are commonly standardized, i.e., so that they have a mean of zero and unit variances.
$\mathbf{U}$	A binary $I \times C$ membership matrix, which specifies to which cluster each object belongs, i.e., $u_{ic} = 1$ if object i belongs to cluster c, and $u_{ic} = 0$ otherwise; $\sum_{c=1}^{C} u_{ic} = 1$.
$\mathbf{F}$	$C \times Q$ centroid matrix, where f_{cq} is the centroid score of cluster c on component q.
$\mathbf{A}$	$J \times Q$ columnwise orthonormal loading matrix; i.e., $\mathbf{A}'\mathbf{A} = \mathbf{I}_Q$.
$\mathbf{A}^{\perp}$	$J \times (J - Q)$ columnwise orthonormal matrix for which it holds that $\mathbf{A}'\mathbf{A}^{\perp} = \mathbf{0}$.
$\mathbf{EA}'$	$I \times Q$ subspace residual matrix.
$\mathbf{E}^{\perp}\mathbf{A}^{\perp\prime}$	$I \times (J - Q)$ complement residual matrix.
$\mathbf{T}$	$Q \times Q$ orthonormal rotation matrix.
$\mathbf{0}_{I\times J}$	$I \times J$ matrix consisting of zeros.
$\mathbf{I}_J$	$J \times J$ identity matrix.
diag($\mathbf{c}$)	Diagonal matrix with diagonal elements equal to the elements of vector $\mathbf{c}$.

2.2. Loss functions and decomposition rules

Both FKM and RKM aim at simultaneously finding an optimal partitioning of the objects and an optimal reduction of the variables. As such, FKM as well as RKM decompose the data matrix $\mathbf{X}$ into a membership matrix $\mathbf{U}$, a columnwise orthonormal loading matrix $\mathbf{A}$ that reveals the extent to which the variables express the clustering structure, and a centroid matrix $\mathbf{F}$ that contains the scores of the cluster centroids on the Q components. To illustrate the interpretation of these matrices, we make use of the following artificial example, with $I = 4, J = 5, C = 2$, and $Q = 2$:

$$\mathbf{U} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \qquad \mathbf{F} = \begin{bmatrix} 1.2 & 1.3 \\ -0.1 & 0.2 \end{bmatrix} \qquad \mathbf{A} = \begin{bmatrix} 0.7 & 0.0 \\ 0.7 & 0.0 \\ 0.0 & 0.7 \\ 0.0 & 0.7 \\ 0.0 & 0.0 \end{bmatrix}.$$

The membership matrix $\mathbf{U}$ shows that Objects 1 and 2 belong to one cluster, and Objects 3 and 4 to another. The centroid matrix contains the centroids of the two clusters, in a two-dimensional space. The loading matrix $\mathbf{A}$ shows that Variables 1 and 2 are associated to the first dimension, Variables 3 and 4 to the second, and Variable 5 does not reflect the clustering structure at all. Such a variable is commonly denoted as a noise or masking variable. Note that, because the size of loadings express the relative importance of the variables concerned in the clustering, FKM and RKM are related to variable selection and weighting approaches. The essential difference is that in the latter two approaches the variables are selected and weighted before cluster analysis, whereas FKM and RKM weight, select and cluster simultaneously.