Chemometrics and Intelligent Laboratory Systems 156 (2016) 249–259

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Overlapping Clusterwise Simultaneous Component Analysis☆

Kim De Roover [a,*], Eva Ceulemans [a], Paolo Giordani [b]

[a] KU Leuven, Belgium
[b] Sapienza University of Rome, Italy

ARTICLE INFO

Article history:
Received 23 November 2015
Received in revised form 18 April 2016
Accepted 8 May 2016
Available online 14 May 2016

Keywords:
Clusterwise simultaneous component analysis
SCA-IND
Overlapping clustering

ABSTRACT

When confronted with multivariate multiblock data (i.e., data in which the observations are nested within different data blocks that have the variables in common), it can be useful to synthesize the available information in terms of components and to inspect between-block similarities and differences in component structure. To this end, the clusterwise simultaneous component analysis (C-SCA) framework was developed across a series of papers: C-SCA partitions the data blocks into a limited number of mutually exclusive groups and performs separate SCA's per cluster. In this paper, we present a more general version of C-SCA. The key difference with the existing C-SCA methods is that the new method does not impose that the clusters are mutually exclusive, but allows for overlapping clusters. Therefore, the new method is called Overlapping Clusterwise Simultaneous Component Analysis (OC-SCA). Each of these clusters corresponds to a single component, such that all the data blocks that are assigned to a particular cluster have the associated component in common. Moreover, the more clusters a specific data block belongs to, the more complex the underlying component structure. A simulation study and an empirical application to emotion data are included in the paper.

1. Introduction

Multivariate multiblock data are a set of matrices that have either the variable (column) mode in common, whereas the entities of the observation mode differ [1], or that have the observation (row) mode in common, whereas the variables differ. Examples of columnwise-coupled multiblock data can be found in several domains of research. In psychology, one may think of multiple emotion ratings of subjects from different age groups, or inhabitants of different countries (e.g., [2,3]). In chemometrics, multiblock data may contain concentrations of chemical compounds in certain substances in different geographical areas, or measured with different measurement techniques, or from different raw material sources, etcetera (e.g., [4,5,6]). In economics, one can think of a questionnaire on work experience administered to workers belonging to different industries or countries (e.g., [7]). In marketing, an example is a survey on the liking of a food item administered to consumers of different countries (e.g., [8]). Examples of rowwise coupled multiblock data include multisource data in chemometrics (e.g., [9]). For the current paper, we will focus on columnwise coupled multiblock data. Adapting the method presented in this paper for rowwise coupled data is a possible direction for future research.

In all of the above cases, it can be useful to synthesize the available information in terms of components and to inspect similarities and differences in the component structures of the data blocks – which we will refer to as the 'within-block structures'. For this purpose, the clusterwise simultaneous component analysis (C-SCA) framework was developed in a series of papers by De Roover and colleagues [1,10]. C-SCA builds on the assumption that, based on their within-block structure, the data blocks can be partitioned into a few mutually exclusive clusters. The cluster-specific component structures are revealed by applying simultaneous component analysis (SCA) [11,12] to the data blocks that are assigned to the same cluster. C-SCA encompasses SCA and standard principal component analyses (PCA) [13,14] on the separate data blocks as special cases. The former is obtained when the number of clusters amounts to one, the latter when the number of clusters equals the number of blocks.

Several C-SCA variants have been proposed in the literature. One model feature that is varied is which particular SCA variant is used (SCA-ECP [1,10], SCA-IND [15], or SCA-P [16]), and thus, which restrictions are imposed on the block-specific component variances and correlations. Moreover, variants differ in whether or not the number of extracted components is restricted to be the same across clusters [17]. Finally, a variant has been proposed that allows some of the extracted components to be shared by all clusters (i.e., common components) and thus distinguishes between common and cluster-specific components [18].

☆ Kim De Roover is a post-doctoral fellow of the Fund for Scientific Research Flanders (Belgium). The research leading to the results reported in this paper was sponsored in part by the Belgian Federal Science Policy within the framework of the Interuniversity Attraction Poles program (IAP/P7/06), and by the Research Council of KU Leuven (GOA/15/003).

* Corresponding author at: Quantitative Psychology and Individual Differences Research Group, Tiensestraat 102, B-3000 Leuven, Belgium.
E-mail address: Kim.DeRoover@kuleuven.be (K. De Roover).

http://dx.doi.org/10.1016/j.chemolab.2016.05.002
0169-7439/

In this paper we will develop a more general version of C-SCA. The key principle of the new method is to seek for overlapping clusters, implying that a data block can be assigned to more than one cluster. Therefore, the method is called Overlapping Clusterwise Simultaneous Component Analysis (OC-SCA-IND; the reasons why we apply the SCA-IND restrictions will be elucidated in Section 2). Allowing for overlapping clusters may be helpful in many domains of research. For instance, in a cross-cultural data set, it is reasonable to think that, on the one hand, countries with the same language share a component and, on the other hand, countries with the same religion share another component, whereas countries will partially overlap in terms of religion and language.

Reconsidering the modeling features of the different C-SCA variants, OC-SCA encompasses several C-SCA variants as special cases. Regarding modeling between-block differences in the number of components, in OC-SCA-IND each cluster corresponds to one component. Consequently, the number of clusters to which a data block belongs gives an indication of the complexity of its underlying component structure. With respect to the common versus cluster-specific nature of components, the number of data blocks that is assigned to a certain cluster reflects how common or specific the corresponding component is, allowing to model different degrees of commonness and specificity.

The paper is organized as follows. In Section 2, SCA-IND and C-SCA-IND are recapitulated. Section 3 is devoted to the new OC-SCA-IND model. The estimation procedure and how to select the optimal number of clusters (which equals the number of components) are discussed in Section 4. Sections 5 and 6 report a simulation study for evaluating the performance of OC-SCA-IND and the results of a real-life application, respectively. In both cases a comparison to the SCA-IND results is included. Finally, Section 7 contains some conclusions and points of discussion.

2. (Clusterwise) simultaneous component analysis models

2.1. Data structure and preprocessing

Columnwise coupled multiblock data consist of I data blocks $\mathbf{X}_i$ $(N_i \times J)$, $i = 1, \ldots, I$, containing the scores of N_i observations on J quantitative variables. We can vertically concatenate the data blocks $\mathbf{X}_i$, $i = 1, \ldots, I$, leading to the data matrix $\mathbf{X}$ $(N \times J)$, where $N = \sum_{i=1}^{I} N_i$ denotes the total number of observations.

Prior to fitting the model to the data, these are usually preprocessed. Specifically, the data are first centered per data block to remove between-block differences in variable means, allowing us to focus on between-block differences in covariance structure. By scaling the data we subsequently eliminate artificial scale differences between variables. In SCA and C-SCA analysis, two scaling options are frequently used, namely autoscaling [19] and overall scaling [12]. In the former case every variable is normalized per data block (i.e., dividing the centered data by the block-specific standard deviations), whereas in the latter case the variables are normalized across all data blocks (i.e., dividing by the overall standard deviations). Therefore, autoscaling should be preferred when one wants to focus on the within-block correlation structure, while overall scaling is recommended to inspect the within-block covariance structure. Since the IND version of SCA will be used, which allows for between-block differences in the variances of the components, overall scaling appears to be the most natural choice in this paper.

2.2. SCA-IND

An SCA model is formulated as

$$\mathbf{X}_i = \mathbf{F}_i \mathbf{B}' + \mathbf{E}_i, i = 1, \ldots, I, \quad (1)$$

where $\mathbf{F}_i$ $(N_i \times Q)$ and $\mathbf{B}$ $(J \times Q)$ are the component score matrix of data block i and the component loading matrix, respectively, where Q denotes the number of components, and $\mathbf{E}_i$ $(N_i \times J)$ is the error matrix of data block i. As stated in the introduction, several variants have been proposed (i.e., SCA-ECP, SCA-IND, SCA-PF2, and SCA-P), that impose different restrictions on the variances and correlations of the block-specific component score matrices (for more details, see [12]). Generally speaking, the more restrictions are imposed, the less between-block differences are allowed for. Therefore, none of the variants is uniformly the best choice. Which variant is selected thus strongly depends on the data set under investigation. In this paper, we focus on SCA-IND (i.e., SCA with INDscal constraints), in which the block-specific component scores are uncorrelated. The variances of the component scores may differ across the blocks, but equal one across all blocks. Unlike SCA-ECP and SCA-P, SCA-IND has no rotational freedom (under mild assumptions), which makes interpretation simpler.

2.3. C-SCA-IND and other C-SCA variants

C-SCA models cluster the data blocks into K mutually exclusive groups and formulate a separate SCA model within each cluster. C-SCA [1,10] was originally formulated as follows:

$$\mathbf{X}_i = \sum_{k=1}^{K} p_{ik} \mathbf{F}_i^{(k)} \mathbf{B}^{(k)\prime} + \mathbf{E}_i, i = 1, \ldots, I, \quad (2)$$

where $\mathbf{F}_i^{(k)}$ is the component score matrix of data block i when assigned to cluster k, and $\mathbf{B}^{(k)}$ is the component loading matrix of cluster k. The matrices $\mathbf{F}_i^{(k)}$ and $\mathbf{B}^{(k)}$ have order $(N_i \times Q)$ and $(J \times Q)$, respectively, where Q denotes the number of cluster-specific components. Finally, the entries p_{ik} of the partition matrix $\mathbf{P}$ take values 1 (if data block i is assigned to cluster k) or 0 (otherwise). Moreover, it holds that $\sum_{k=1}^{K} p_{ik} = 1, i = 1, \ldots, I$. Hence, if $K = 1$, then $\mathbf{P} = \mathbf{1}$ (where $\mathbf{1}$ denotes a column vector of 1's) and C-SCA reduces to SCA.

Although C-SCA-ECP [1,10] and C-SCA-P versions [16] have been proposed as well, we focus here on the C-SCA-IND variant [15]. This variant has no rotational freedom and, unlike C-SCA-P, forces all important between-block differences in the correlations of the variables to show up in the clustering. Moreover, the often too restrictive C-SCA-ECP assumption of equal component variances — implying that each component gets an equal weight in the solution for each data block — is avoided.

Regarding between-block differences in the complexity of the component structure, C-SCA models generally restrict the number of components to be the same across clusters. Since this assumption is often unrealistic, De Roover et al. [17] proposed a variant that allows for different numbers of cluster-specific components $Q^{(k)}$.

Finally, since all components are cluster-specific, it can be concluded that C-SCA models strongly focus on structural differences. However, in many cases, it is reasonable to expect that next to these differences, there will also be a lot of structural similarity. To better capture both aspects — similarities and differences — a C-SCA variant was proposed that allows for common components, shared by all clusters, as well as cluster-specific ones [18]. This model is formulated as follows:

$$\mathbf{X}_i = \mathbf{F}_{i,comm} \mathbf{B}_{comm}{}' + \sum_{k=1}^{K} p_{ik} \mathbf{F}_{i,spec}^{(k)} \mathbf{B}_{spec}^{(k)}{}' + \mathbf{E}_i, i = 1, \ldots, I, \quad (3)$$

where the subscripts '*comm*' and '*spec*' indicate 'common' and 'cluster-specific', respectively. $\mathbf{F}_{i,comm}$ ($\mathbf{F}_{i,spec}^{(k)}$) and $\mathbf{B}_{comm}$ ($\mathbf{B}_{spec}$) are the common (cluster-specific) component score matrix for data block i and common (cluster-specific) component loading matrix, respectively. One drawback of CC-SCA is that the number of common components and