European Journal of Operational Research 238 (2014) 391–403

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Invited Review

Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis

Arthur Tenenhaus [a,*], Michel Tenenhaus [b,1]

[a] SUPELEC, Plateau de moulon, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette Cedex, France
[b] HEC Paris, 1 rue de la Libération, 78351 Jouy-en-Josas Cedex, France

ARTICLE INFO

Article history:
Received 2 November 2012
Accepted 3 January 2014
Available online 13 January 2014

Keywords:
Multiblock data analysis
Multigroup data analysis
Regularized generalized canonical correlation analysis

ABSTRACT

This paper presents an overview of methods for the analysis of data structured in blocks of variables or in groups of individuals. More specifically, regularized generalized canonical correlation analysis (RGCCA), which is a unifying approach for multiblock data analysis, is extended to be also a unifying tool for multigroup data analysis. The versatility and usefulness of our approach is illustrated on two real datasets.

1. Introduction

In this paper, we consider a data matrix $\mathbf{X}$ structured in groups (partition of rows) or in blocks (partition of columns). Rows of $\mathbf{X}$ are related to individuals and columns to variables. Multiblock data analysis concerns the analysis of several sets of variables (blocks) observed on the same set of individuals. Multigroup data analysis concerns the analysis of one set of variables observed on several groups of individuals. Note that there is no established consensus in the literature on the use of the terms "multiblock" and "multigroup". Therefore these two terms are clearly defined in this paper.

In the *multiblock* framework, a column partition $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_j, \ldots, \mathbf{X}_J]$ is considered. In this case, each $n \times p_j$ data matrix $\mathbf{X}_j$ is called a block and represents a set of p_j variables observed on n individuals. The number and the nature of the variables usually differ from one block to another but the individuals must be the same across blocks. The main aim is to investigate the relationships between blocks. The data might be preprocessed in order to ensure comparability between variables and blocks. To make variables comparable, standardization is applied (zero mean and unit variance). To make blocks comparable, a possible strategy is to divide each block by $\sqrt{p_j}$ (Wold, Hellberg, Lundstedt, Sjostrom, & Wold, 1987). This two-step procedure leads to $Trace\left(\mathbf{X}_j^t\mathbf{X}_j\right) = n$ for each block.

In the *multigroup* framework, a row partition $\mathbf{X} = \left[\mathbf{X}_1^t, \ldots, \mathbf{X}_i^t, \ldots, \mathbf{X}_I^t\right]^t$ is considered. In this framework, the same set of variables is observed on different groups of individuals. Each $n_i \times p$ data matrix $\mathbf{X}_i$ is called a group. The number of individuals of each group can differ from one group to another. The main aim is to investigate the relationships among variables within the various groups. Following the proposal of Kiers and Ten Berge (1994) variables are centered and normalized (i.e. set to unit norm) within each group. This preprocessing leads to $Trace(\mathbf{X}_i^t\mathbf{X}_i) = p$ for each group.

Many methods exist for multiblock and multigroup data analysis.

Two families of methods have come to the fore in the field of multiblock data analysis. These methods rely on correlation-based or covariance-based criteria. Canonical correlation analysis (Hotelling, 1936) is the seminal paper for the first family and Tucker's inter-battery factor analysis (Tucker, 1958) for the second one. These two methods have been extended to more than two blocks in many ways:

(1) Main contributions for generalized canonical correlation analysis (GCCA) are found in Horst (1961), Carroll (1968a), Kettenring (1971), Wold (1982, 1985) and Hanafi (2007).
(2) Main contributions for extending Tucker's method to more than two blocks come from Carroll (1968b), Chessel and Hanafi (1996), Hanafi and Kiers (2006), Hanafi, Kohler, and Qannari (2010, 2011), Hanafi, Mazerolles, Dufour, and Qannari (2006), Krämer (2007), Smilde, Westerhuis, and de Jong (2003), Ten Berge (1988), Van de Geer (1984), Westerhuis, Kourti, and MacGregor (1998), Wold (1982, 1985).

* Corresponding author. Address: SUPELEC, Department of Signal Processing and Electronics Systems, Plateau de Moulon, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette Cedex, France. Tel.: +33 (0)1 69 85 14 22; fax: +33 (0)1 69 85 14 29.
E-mail addresses: Arthur.Tenenhaus@supelec.fr (A. Tenenhaus), tenenhaus@hec.fr (M. Tenenhaus).
[1] Tel.: +33 (0)1 39 67 70 00; fax: +33 (0)1 39 67 74 00.

http://dx.doi.org/10.1016/j.ejor.2014.01.008
0377-2217/

(3) Carroll (1968b) proposed the "mixed" correlation and covariance criterion. van den Wollenberg (1977) combined correlation and variance for the two-block situation (redundancy analysis). This method is extended to a multiblock situation in this paper.

Regularized generalized canonical correlation analysis (Tenenhaus & Tenenhaus, 2011) includes many of these references as particular cases.

For multigroup data analysis, we may distinguish three families of methods:

(1) Several methods combine the covariance matrices $\mathbf{S}_i$ or the correlation matrices $\mathbf{R}_i$ related to the various groups.

In a seminal paper, Levin (1966), considering the problem of simultaneous factor analysis, proposed the diagonalization of $\overline{\mathbf{R}} = (1/I)\sum_{i=1}^{I}\mathbf{R}_i$. The acronym SUMPCA$_c$ for the Levin method was proposed by Kiers (1991). Kiers and Ten Berge (1994) proposed several simultaneous component analysis methods (see paragraph 3 below): one of them (SCA-P) leads also to the diagonalization of $\overline{\mathbf{R}}$. Krzanowski (1984) proposed to carry out a multigroup PCA (MGPCA) by diagonalizing either $\mathbf{T} = \sum_{i=1}^{I}\mathbf{S}_i$ or the within group covariance matrix $\mathbf{S} = \sum_{i=1}^{I}(n_i/n)\mathbf{S}_i$.

Krzanowski (1979) proposed to use an approach similar to Carroll's GCCA (correlation criterion) for comparing group correlation matrices $\mathbf{R}_1, \ldots, \mathbf{R}_I$. Since GCCA on these matrices yields a trivial solution, Krzanowski replaced each matrix $\mathbf{R}_i$ by its k first eigenvectors and then applied GCCA on the obtained matrices.

Flury (1984) proposed a method called common principal component analysis (CPC) for I groups by supposing a special structure on the covariance matrices $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_I$ defined at the population level. In CPC, the I covariance matrices have the same eigenvectors but the eigenvalues are specific to each group: $\mathbf{\Sigma}_i = \mathbf{A}\mathbf{\Lambda}_i\mathbf{A}^t$ where $\mathbf{A}$ is orthogonal and $\mathbf{\Lambda}_i$ diagonal. Flury (1987) also proposed a partial common principal component analysis (PCPC) where only q eigenvectors of $\mathbf{\Sigma}_i$ are common to all populations. Flury's approach is based on normal-theory maximum likelihood and a complicated iterative algorithm is required. Two alternative algorithms have been proposed: (1) Krzanowski (1984) showed empirically that MGPCA and PCPC give very close results; (2) a sequential least squares solution to PCPC can be obtained by using the CCSWA algorithm (Common components and specific weight analysis) described in Hanafi et al. (2006). It is also worth mentioning that CCSWA and HPCA (Hierarchical principal component analysis described in Westerhuis et al. (1998)) are two equivalent methods (Hanafi et al., 2010).

(2) It is always possible to use a multiblock method on multigroup data by considering the transpose of each group. Eslami, Qannari, Kohler, and Bougeard (2013a) proposed to use an approach similar to Carroll's GCCA (correlation and covariance criteria) on the transpose groups $\mathbf{X}_1^t, \ldots, \mathbf{X}_I^t$. This approach has later been extended to a multiblock/multigroup situation in Eslami, Qannari, Kohler, and Bougeard (2013b).

(3) Kiers and Ten Berge (1989, 1994) and later Timmerman and Kiers (2003) proposed a generalization of PCA to a multigroup situation under the generic name of Simultaneous Component Analysis (SCA). Each data group $\mathbf{X}_i$ of dimension $n_i \times p$ is modeled by a lower rank $n_i \times p$ matrix $\widehat{\mathbf{X}}_i = \mathbf{X}_i\mathbf{W}_i\mathbf{P}_i^t$ where $\mathbf{W}_i$ is a $p \times q$ ($q < p$) weight matrix and $\mathbf{P}_i$ a $p \times q$ pattern matrix. A factor matrix $\mathbf{F}_i = \mathbf{X}_i\mathbf{W}_i$ and a loading (or structure) matrix $\mathbf{L}_i = \mathbf{R}_i\mathbf{W}_i$ are also introduced. PCA and SCA methods are about minimizing $\sum_{i=1}^{I}\|\mathbf{X}_i - \mathbf{X}_i\mathbf{W}_i\mathbf{P}_i^t\|^2$ subject to specific constraints on the weight/pattern/structure/factor matrices which are summarized in Table 1.

The reconstructed matrix $\widehat{\mathbf{X}}_i = \mathbf{X}_i\mathbf{W}_i\mathbf{P}_i^t$ is invariant up to an orthogonal (rotation) matrix $\mathbf{A}$: $\mathbf{X}_i\mathbf{W}_i\mathbf{P}_i^t = \mathbf{X}_i\mathbf{W}_i\mathbf{A}^t\mathbf{A}\mathbf{P}_i^t$. This invariance can be used to improve interpretation. Various rotation methods are described in Niesing (1997). Moreover, Niesing shows in a comparative study that SCA-P gives the best practical results.

Table 1
PCA and SCA methods.

Methods	Constraints
Separate PCA by group (Pearson (1901))	$\mathbf{P}_i = \mathbf{W}_i$ and $\mathbf{W}_i^t\mathbf{W}_i = \mathbf{I}\ \forall i$
SCA-W (Kiers & Ten Berge (1989))	$\mathbf{W}_1 = \cdots = \mathbf{W}_I$
SCA-P (Kiers & Ten Berge (1994))	$\mathbf{P}_1 = \cdots = \mathbf{P}_I$
SCA-S (Kiers & Ten Berge (1994))	$\mathbf{R}_i\mathbf{W}_i = \mathbf{L}\mathbf{D}_i\ \forall i$[a]
SCA-PF2 (Timmerman & Kiers (2003))	$\mathbf{P}_1 = \cdots = \mathbf{P}_I$ and $\mathbf{F}_i^t\mathbf{F}_i = \mathbf{D}_i\mathbf{\Phi}\mathbf{D}_i\ \forall i$[b]
SCA-IND (Timmerman & Kiers (2003))	$\mathbf{P}_1 = \cdots = \mathbf{P}_I$ and $\mathbf{F}_i^t\mathbf{F}_i = \mathbf{D}_i^2\ \forall i$
SCA-ECP (Timmerman & Kiers (2003))	$\mathbf{P}_1 = \cdots = \mathbf{P}_I$ and $(1/n_i)\mathbf{F}_i^t\mathbf{F}_i = \mathbf{I}\ \forall i$

[a] $\mathbf{L}$ is an unknown $(p \times q)$ matrix and $\mathbf{D}_i$ an unknown $(q \times q)$ diagonal matrix.
[b] $\mathbf{\Phi}$ is an unknown $(q \times q)$ positive definite matrix with unit diagonal elements.

De Roover, Ceulemans, and Timmerman (2012), De Roover, Ceulemans, Timmerman, and Onghena (2013) and De Roover, Timmerman, Van Mechelen, and Ceulemans (2013) have developed a clusterwise approach to SCA-P, SCA-PF2, SCA-IND and SCA-ECP for tracing structural differences and similarities between different groups of individuals.

Finally, Van Deun, Smilde, van der Werf, Kiers, and Van Mechelen (2009) proposed a simultaneous component analysis framework for multiblock and multigroup data analysis.

In this paper, a modified version of RGCCA, which can be applied to multiblock and multigroup data, is described. The acronym RGCCA will be kept for this more general method. This paper is organized as follows: the general optimization problem for both multiblock and multigroup data analysis is presented in Section 2. A monotonically convergent algorithm is presented in Section 3. An overview of applications of RGCCA for multiblock and multigroup data analysis is given in Sections 4 and 5. The versatility and usefulness of our approach is illustrated on two real datasets in Sections 6 and 7.

2. The optimization problem behind RGCCA for multiblock or multigroup data analysis

RGCCA for multiblock and multigroup data analyses is based on a single optimization problem that we present in this section. We consider I matrices $\mathbf{Q}_1, \ldots, \mathbf{Q}_I$. Each matrix $\mathbf{Q}_j$ is of dimension $m \times p_i$. We also associate to each matrix $\mathbf{Q}_j$ a weight column vector $\mathbf{w}_i$ of dimension p_i and a symmetric definite positive matrix $\mathbf{M}_i$ of dimensions $p_i \times p_i$. Moreover, a design matrix $\mathbf{C} = \{c_{i\ell}\}$ is defined with $c_{i\ell} = 1$ if matrices $\mathbf{Q}_i$ and $\mathbf{Q}_\ell$ are connected, and $c_{i\ell} = 0$ otherwise. The core optimization problem considered in this paper is defined as follows:

$$\begin{aligned} &\underset{\mathbf{w}_1,\ldots,\mathbf{w}_I}{\text{Maximize}} \sum_{i,\ell,i\neq\ell} c_{i\ell}\, g(\langle \mathbf{Q}_i\mathbf{w}_i, \mathbf{Q}_\ell\mathbf{w}_\ell\rangle) \\ &\text{s.c. } \mathbf{w}_i^t\mathbf{M}_i\mathbf{w}_i = 1, \quad i = 1,\ldots,I \end{aligned} \tag{1}$$

where $\langle \mathbf{x}, \mathbf{y}\rangle = \mathbf{x}^t\mathbf{y}$ is the usual scalar product and g stands for the identity, the absolute value or the square function. By setting $\mathbf{v}_i = \mathbf{M}_i^{1/2}\mathbf{w}_i$ and $\mathbf{P}_i = \mathbf{Q}_i\mathbf{M}_i^{-1/2}$ optimization problem (1) becomes

$$\begin{aligned} &\underset{\mathbf{v}_1,\ldots,\mathbf{v}_I}{\text{Maximize}} \sum_{i,\ell,i\neq\ell} c_{i\ell}\, g(\langle \mathbf{P}_i\mathbf{v}_i, \mathbf{P}_\ell\mathbf{v}_\ell\rangle) \\ &\text{s.c. } \mathbf{v}_i^t\mathbf{v}_i = 1, \quad i = 1,\ldots,I \end{aligned} \tag{2}$$

A monotone convergent algorithm can be developed for optimization problem (2). This algorithm will be presented in detail in the next section. It is worth mentioning that all the multiblock and multigroup methods to be presented in this paper are special cases of optimization problem (2).