Clustering and Disjoint Principal Component Analysis

Maurizio Vichi, Gilbert Saporta

PII: S0167-9473(08)00293-4
DOI: 10.1016/j.csda.2008.05.028
Reference: COMSTA 4056

To appear in: *Computational Statistics and Data Analysis*

Please cite this article as: Vichi, M., Saporta, G., Clustering and Disjoint Principal Component Analysis. *Computational Statistics and Data Analysis* (2008), doi:10.1016/j.csda.2008.05.028

Clustering and Disjoint Principal Component Analysis

Maurizio Vichi[1],
Dpt. Statistics, Probability, and Applied Statistics,
University “La Sapienza”, P.le A. Moro 5, 00185 Rome, Italy
maurizio.vichi@uniroma1.it

Gilbert Saporta,
CEDRIC, CNAM,
292 rue Saint Martin, 75003 Paris, France,
saporta@cnam.fr

Abstract: A constrained principal component analysis, which aims at a simultaneous clustering of objects and a partitioning of variables is proposed. The new methodology allows to identify components with maximum variance, each one a linear combination of a subset of variables. All the subsets form a partition of variables. Simultaneously, a partition of objects is also computed maximizing the between cluster variance. The methodology is formulated in a semi-parametric least-squares framework as a quadratic mixed continuous and integer problem. An alternating least-squares algorithm is proposed to solve the clustering and disjoint PCA. Two applications are given to show the features of the methodology.

Keywords: Principal Component Analysis, Cluster Analysis of variables, rotation of PCA

1. Introduction

The reduction and synthesis of objects and variables is one of the most used analyses for exploring and mining the observed data. The aim is to detect the most relevant information which allows an appropriate interpretation of the data. Dimensionality reduction of objects (units) is generally achieved by a cluster analysis methodology, and frequently by a partitioning algorithm. In this last case, together with an optimal partition, also a reduced set of *centroids* is detected, i.e., generally non-observable objects representing mean profiles of units belonging to clusters. The reduction of variables is sometimes obtained by a clustering procedure by using specific measures of proximities between variables, such as correlation or association measures. More frequently, variable reduction is provided by principal component analysis (PCA) or by a factorial technique (FA) in order to detect non-observable dimensions (components, factors) summarizing the common information in the data set.

The reduction of objects and variables can be obtained by applying the two techniques sequentially. This is often done by first carrying out a PCA and subsequently applying a clustering algorithm on the component scores of the objects, thereby using the first few components only. However, De Sarbo et al. (1990), De Soete & Carroll (1994), Vichi & Kiers (2001) warn against this approach, called "*tandem analysis*", because PCA or FA may identify dimensions that do not necessarily contribute much to perceive the clustering structure in the data and that, on the contrary, may obscure or mask the taxonomic information.

In an attempt to focus on the main information in the data conveyed in a limited number of dimensions, De Soete & Carroll (1994) proposed a modified *k*-means procedure such that each

[1] Corresponding author. The paper was written when Maurizio Vichi was visiting CEDRIC, CNAM, Paris, France

cluster is represented by a centroid in a low-dimensional space, chosen so that it is the closest point to the objects associated with the cluster at hand. Vichi and Kiers (2001) proposed a methodology that allows one to select the most relevant components for the classification. A representation in a reduced number of dimensions can be given to help the interpretation of the relationships within the set of variables and the set of objects. Vichi (2002) introduced a model that partitions the objects along a set of centroids and variables along a set of components, so as to minimize the within cluster deviance of components in the reduced space. Properties of the relationship between these methods will be discussed into more details later in this paper. Various alternative methods combining cluster analysis and the search for a low-dimensional representation have been also proposed in the multidimensional scaling or unfolding context (e.g., Heiser, 1993, De Soete and Heiser, 1993, Heiser and Groenen, 1997).

In this paper a new methodology is proposed to cluster objects along a set of centroids and variables along a set of components in order to maximize the between cluster deviance of the components in the reduced space. Here, we require to maximize the between variance to have a classification of the variables identified via components with maximal variance. An efficient alternating least-squares algorithm is also given. The new methodology is named Clustering and Disjoint Principal Component Analysis (CDPCA), because we require to find components associated to disjoint classes of variables.

The overview of the paper is as follows. In section 2, for convenience of the reader the notation common to all sections is listed. Section 3 is devoted to the discussion of the clustering and disjoint PCA model. In section 4 the least-squares estimation of the model given in section 3 is discussed and a general alternating least-squares algorithm is described. Performances of the clustering and disjoint PCA are evaluated in section 5 by applying the new methodology to a first data set describing short-term macroeconomic scenario of OECD countries and a second data set analyzing the socio-economic characteristics of 103 countries according to 12 variables. In section 6 some conclusions follow.

2. Notation

For the convenience of the reader the notation and terminology common to all sections is listed here.

I,J,P,Q, number of: objects, variables, clusters of objects, clusters of variables, respectvely;

$\mathbf{X}=[x_{ij}]$ ($I \times J$) two-way two-mode (objects and variables) data matrix, describing the J-variate profiles of I objects. The variables to be analyzed are supposed commensurate, and therefore if they are expressed by different units of measurements they are standardized to have mean zero and unit variance;

$\mathbf{E}=[e_{ij}]$ ($I \times J$) error matrix;

$\mathbf{U} = [u_{ip}]$ ($I \times P$) binary object membership matrix defining a partition of the objects into P clusters, where u_{ip}=1 if the i^{th} object belongs to cluster p^{th}; u_{ip}=0, otherwise. Matrix **U** is row stochastic, i.e., has no negative elements that sum to one per row, and thus, has only one nonzero element per row;

$\mathbf{V} =[v_{jq}]=[\mathbf{v}_q]$ ($J \times Q$) binary matrix defining a partition of variables into Q clusters, where v_{jq}=1 if the j^{th} variable belongs to q^{th} cluster, v_{jq}=0, otherwise. Matrix **V** has only one nonzero element per row;

$\bar{\mathbf{X}}$ ($P \times J$) object centroids matrix $[\bar{\mathbf{x}}_1,\ldots,\bar{\mathbf{x}}_p,\ldots,\bar{\mathbf{x}}_P]'$, where $\bar{\mathbf{x}}_p$ represents the centroid in the space of the observed variables. In the case of a least-squares estimation, the centroid matrix has the form $\overline{\mathbf{X}}=(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{X}$; and $\overline{\mathbf{x}}_p = \left(\mathbf{u}'_p\mathbf{u}_p\right)\mathbf{u}'_p\mathbf{X}$, where $\mathbf{u}_p$ is the p^{th} column of $\mathbf{U}$;

$\overline{\mathbf{Y}}$ ($P \times Q$) object centroids matrix $[\overline{\mathbf{y}}_1,\ldots,\overline{\mathbf{y}}_p,\ldots,\overline{\mathbf{y}}_P]'$, where $\overline{\mathbf{y}}_p$ represents the centroid in the reduced space. In the case of a least-squares estimation, the centroid matrix has the form $\overline{\mathbf{Y}} = (\mathbf{U}'\mathbf{U})^{-1}\ \mathbf{U}'\mathbf{XA}$; and $\overline{\mathbf{y}}_p = \left(\mathbf{u}'_p\mathbf{u}_p\right)\mathbf{u}'_p\mathbf{XA}$, where $\mathbf{u}_p$ is the p^{th} column of $\mathbf{U}$;

$\mathbf{A}=[a_{jq}]=[\mathbf{a}_q]$ ($J \times Q$) matrix of the coefficients of the linear combination, with $\sum_{j=1}^{J}\left(a_{jq}a_{jr}\right)^2 = 0$, for any q and r ($q \neq r$) and $\sum_{j=1}^{J} a_{jq}^2 = 1$. Note that matrix $\mathbf{A}$ is columnwise orthonormal;

$\mathbf{C}=[c_{jq}]=[\mathbf{c}_q]$ ($J \times Q$) columnwise orthonormal matrix;

$\mathbf{Y}=[y_{iq}=\sum_{j=1}^{J} a_{jq}x_{ij}\ ]$ ($I \times Q$) component score matrix, where y_{iq} is the value of i^{th} object for the q^{th} component y_q; synthesizing the common information of a subset of variables;

The model, proposed in this paper, detects an optimal partition of variables into Q classes. For each class of the partition a component which is the linear combination of the variables in the class with maximal variance is found. Simultaneously, a partition of the units of maximal between variance is detected.

3. Clustering and disjoint PCA model

The model associated to the *clustering and disjoint principal component analysis* can be formally written as follows

$$\mathbf{X} = \mathbf{U}\,\overline{\mathbf{Y}}\,\mathbf{A}' + \mathbf{E}, \tag{1}$$

where $\overline{\mathbf{Y}}$ is the matrix of order ($P \times Q$) of centroids in the reduced space, matrix $\mathbf{U}$ is binary and row stochastic, i.e., formally

$$u_{ip} \in \{0,\ 1\}, \qquad (i=1,\ldots,I;\ p=1,\ldots,P); \tag{2}$$

$$\sum_{p=1}^{P} u_{ip} = 1\ , \qquad (i=1,\ldots,I); \tag{3}$$

and $\mathbf{A}$, with rank($\mathbf{A}$)=$Q \leq J$, satisfies the constraints

$$\sum_{j=1}^{J} a_{jq}^2 = 1 \qquad q=1,\ \ldots,\ Q; \tag{4}$$

$$\sum_{j=1}^{J} (a_{jq}a_{jr})^2 = 0 \qquad q=1,\ldots,Q\text{-}1;\ r=q+1,\ldots,Q. \tag{5}$$

and therefore is orthonormal.

Model (1) specifies a partition of the objects via the membership matrix $\mathbf{U}$ and the centroid matrix $\overline{\mathbf{Y}}$, and simultaneously a dimensionality reduction via the component loading matrix $\mathbf{A}$, which allows for the partition of the variables into classes each summarized by an orthonormal linear combination with constraints (5).

Constraints (2) and (3) specify a membership matrix **U** and consequently a partition for the objects. Constraints (4) and (5) are more restrictive than the orthogonal ones; indeed, they require that for any two different components q^{th} and r^{th}, at least one of the two component loadings a_{jq} and a_{jr} for a variable j^{th} is equal to zero. They imply: *a*) **A** is orthonormal, i.e., $\mathbf{A}'\mathbf{A}=\mathbf{I}_Q$; *b*) each row of **A** (representing a variable) has at most a single positive loading for a component, and therefore, a variable can contribute only to a single component; *c*) from b) a partition of the variables is induced. Constraints (5) are satisfied also if for a variable j^{th} we have $a_{jq} = 0$, for $q = 1,..., Q$, which implies that the variable does not contribute to any component. Thus, this case corresponds to exclude variable j^{th} in the CDPCA and consequently to have in the methodology a variable selection option. However, if we require to use all the observed variables, CDPCA has to satisfy the following additional constraint

$$\sum_{q=1}^{Q} a_{jq}^2 > 0 \qquad q=1,\ldots,Q. \tag{6}$$

We now give an example that helps to understand the constraints (4) and (5), which are crucial in the clustering and disjoint principal component analysis.

Example 1: The following matrix **A** of dimension (J=8 × Q=3),

$$\mathbf{A} = \begin{array}{c} \\ \text{Var. 1} \\ \text{Var. 2} \\ \text{Var. 3} \\ \text{Var. 4} \\ \text{Var. 5} \\ \text{Var. 6} \\ \text{Var. 7} \\ \text{Var. 8} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Comp. 1} & \text{Comp. 2} & \text{Comp. 3} \end{array} \\ \begin{bmatrix} 0.5593 & 0 & 0 \\ 0.6617 & 0 & 0 \\ 0.4994 & 0 & 0 \\ 0 & -0.7993 & 0 \\ 0 & 0.6010 & 0 \\ 0 & 0 & 0.3685 \\ 0 & 0 & 0.6131 \\ 0 & 0 & -0.6988 \end{bmatrix} \end{array}$$

satisfies constraints (4) and (5). **A** is orthonormal. Each row (variable) contributes to a single column (component), i.e., has a positive weight for a single column. **A** can be interpreted as a membership matrix specifying with the positive weights a partition into three classes (class 1: variables 1, 2 and 3; class 2: variables 4 and 5; class 3: variables 6, 7 and 8).

Therefore, for a given matrix **X** of dimension (I×8) written as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6, \mathbf{x}_7, \mathbf{x}_8]$, where $\mathbf{x}_j$ is a I dimensional vector representing a variable, a partition of variables in the following Q=3 classes $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, $\{\mathbf{x}_4, \mathbf{x}_5\}$, $\{\mathbf{x}_6,\mathbf{x}_7, \mathbf{x}_8\}$, is specified. The three components, associated to the three classes have the form:

$$\mathbf{y}_1 = 0.5593\mathbf{x}_1+ 0.6617\mathbf{x}_2 + 0.4994\mathbf{x}_3;$$
$$\mathbf{y}_2 = - 0.7993\mathbf{x}_4 + 0.6010\mathbf{x}_5;$$
$$\mathbf{y}_3 = 0.3685\mathbf{x}_6+ 0.6131\mathbf{x}_7 -0.6988\mathbf{x}_8.♦$$

3. 1 Minimization in CDPCA

From model (1), the least-squares estimators of the CDPCA are the optimal solutions of the following quadratic problem [P1] with respect to unknown **A**, **U** and $\overline{\mathbf{Y}}$,

$$F(\mathbf{U},\ \overline{\mathbf{Y}},\ \mathbf{A}) = \|\ \mathbf{X}\ \text{-}\ \mathbf{U}\overline{\mathbf{Y}}\mathbf{A}'\|^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(x_{ij} - \sum_{p=1}^{P}\sum_{q=1}^{Q} u_{ip}\bar{y}_{pq}a_{jq}\right)^2 \to \min_{\mathbf{A},\mathbf{U},\overline{\mathbf{Y}}} \qquad (7)$$

subject to: **[P1]**

U binary and row stochastic, i.e., satisfying equations (2) and (3);
and **A** satisfying equations (4) and (5).

It can be observed that the following decomposition holds

$$\|\mathbf{X}\|^2 = \|\mathbf{X}\ \text{-}\ \mathbf{U}\,\overline{\mathbf{Y}}\,\mathbf{A}'\|^2 + \|\mathbf{U}\,\overline{\mathbf{Y}}\,\mathbf{A}'\|^2, \qquad (8)$$

where the first term of the right hand side of equation (8) is the *reconstructed* (by **Y**=**XA**) *within deviance* of the partition given by **U** of the observed data **X**, and it is, also, the objective function of CDPCA.
The decomposition can be proved, recalling that $\overline{\mathbf{Y}} = \overline{\mathbf{X}}\,\mathbf{A}$, by showing that

$$\begin{aligned}\|\mathbf{X}\ \text{-}\ \mathbf{U}\,\overline{\mathbf{Y}}\,\mathbf{A}'\|^2 + \|\mathbf{U}\,\overline{\mathbf{Y}}\,\mathbf{A}'\|^2 &= \mathrm{tr}\{[\mathbf{X}\text{-}\mathbf{U}\,\overline{\mathbf{X}}\,\mathbf{AA}'][\ \mathbf{X}\text{-}\mathbf{U}\,\overline{\mathbf{X}}\,\mathbf{AA}']'\} + \mathrm{tr}\{[\mathbf{U}\,\overline{\mathbf{X}}\,\mathbf{AA}'][\mathbf{U}\,\overline{\mathbf{X}}\,\mathbf{AA}']'\}\\ &= \mathrm{tr}\{\mathbf{XX}'\}\ \text{-}\ 2\mathrm{tr}\{\mathbf{U}\,\overline{\mathbf{X}}\,\mathbf{AA}'\mathbf{X}'\} + 2\mathrm{tr}\{\ \mathbf{U}'\mathbf{U}\,\overline{\mathbf{X}}\,\mathbf{AA}'\,\overline{\mathbf{X}}'\ \}\\ &= \mathrm{tr}\{\mathbf{XX}'\}\text{-}2\mathrm{tr}\{\ \mathbf{U}\,\overline{\mathbf{X}}\,\mathbf{AA}'\mathbf{X}'\}+2\mathrm{tr}\{\ \mathbf{U}'\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{X}\ \mathbf{AA}'\,\overline{\mathbf{X}}'\ \}\\ &= \mathrm{tr}\{\mathbf{XX}'\}.\end{aligned}$$

From decomposition (8), we derive that we can minimize equation (7) or alternatively, but equivalently, we can maximize the second term of the right hand side of (8)

$$\|\mathbf{U}\,\overline{\mathbf{Y}}\,\mathbf{A}'\|^2, \qquad (9)$$

that corresponds to the *reconstructed* (by **Y**=**XA**) *between-class deviance* of the partition given by **U** of **X**. Finally, we can prove that

$$\|\mathbf{U}\,\overline{\mathbf{Y}}\,\mathbf{A}'\|^2 = \|\mathbf{U}\,\overline{\mathbf{X}}\,\mathbf{AA}'\|^2 = \mathrm{tr}\{[\mathbf{U}\,\overline{\mathbf{X}}\,\mathbf{AA}'][\mathbf{U}\,\overline{\mathbf{X}}\,\mathbf{AA}']'\} = \mathrm{tr}\{[\mathbf{U}\,\overline{\mathbf{X}}\,\mathbf{A}][\mathbf{U}\,\overline{\mathbf{X}}\,\mathbf{A}]'\} = \|\mathbf{U}\,\overline{\mathbf{X}}\,\mathbf{A}\|^2.$$

Therefore, problem [P1] is equivalent to the maximization of the between-class deviance $\|\mathbf{U}\,\overline{\mathbf{X}}\,\mathbf{A}\|^2$ of the reduced space, subject to constraints (2) – (5).

Remark 1. Model (1) is the joint model associated to *k*-means applied on **X** and the principal component analysis applied on the matrix of centroids.
In fact, *k*-means applied to **X** corresponds to fit the model

$$\mathbf{X} = \mathbf{U}\,\overline{\mathbf{X}} + \mathbf{E}^{(1)} \qquad (10)$$

by

$$\| \mathbf{X} - \mathbf{U}\bar{\mathbf{X}} \|^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(x_{ij} - \sum_{p=1}^{P} u_{ip}\bar{x}_{pj} \right)^2 \rightarrow \min_{\mathbf{U},\bar{\mathbf{X}}} \tag{11}$$

subject to: **U** being a binary and row-stochastic matrix,

where $\mathbf{E}^{(1)}$ is the error term matrix associated to the *k*-means clustering.
The principal component analysis applied on the centroid matrix $\bar{\mathbf{X}}$ corresponds to fit the model

$$\mathbf{U}\bar{\mathbf{X}} = \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}' + \mathbf{E}^{(2)} \tag{12}$$

by

$$\| \mathbf{U}\bar{\mathbf{X}} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}' \|^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\sum_{p=1}^{P} u_{ip}\bar{x}_{pj} - \sum_{q=1}^{Q}\sum_{p=1}^{P} u_{ip}\bar{y}_{pq}a_{jq} \right)^2 \rightarrow \min_{\mathbf{A},\bar{\mathbf{Y}}} \tag{13}$$

subject to:

$$\mathbf{A}'\mathbf{A} = \mathbf{I}_Q, \tag{14}$$

where $\mathbf{E}^{(2)}$ is the error term matrix of PCA applied on $\mathbf{U}\bar{\mathbf{X}}$.
Model (1) can be viewed as the simultaneous version of the procedure based on the application of *k*-means on **X**, followed by PCA applied on $\mathbf{U}\bar{\mathbf{X}}$. In fact, including (12) into (10) and by setting $\mathbf{E} = \mathbf{E}^{(1)} + \mathbf{E}^{(2)}$ equation (1) is derived. Equation (1) subject to **U** binary and row stochastic and orthonormality constraints (14) is the model specified by the *reduced k-means*, De Soete and Carroll, (1994), i.e., the quadratic problem

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\left(x_{ij} - \sum_{p=1}^{P}\sum_{q=1}^{Q} u_{ip}\bar{y}_{pq}a_{jq} \right)^2 \rightarrow \min_{\mathbf{A},\mathbf{U},\bar{\mathbf{Y}}}$$

subject to

U being a binary and row-stochastic matrix.
$\mathbf{A}'\mathbf{A} = \mathbf{I}_Q$.♦

Remark 2. The factorial *k*-means model for simultaneous clustering and PCA is also linked to CDPCA. In fact, factorial *k*-means is mathematically specified as (Vichi & Kiers, 2001),

$$\mathbf{X}\mathbf{A}\mathbf{A}' = \mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}' + \mathbf{E} \tag{15}$$

where **U** is the object membership matrix, $\bar{\mathbf{X}}$ is the centroid matrix and **A** is the component loading matrix of rank(**A**) ≤ *J* and orthonormal, i.e., $\mathbf{A}'\mathbf{A} = \mathbf{I}_Q$. In factorial *k*-means the optimal **U**, $\bar{\mathbf{X}}$, and **A** are obtained by,

$$\| \mathbf{Y}\mathbf{A}' - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}' \|^2 = \| \mathbf{X}\mathbf{A} - \mathbf{U}\bar{\mathbf{X}}\mathbf{A} \|^2 = \sum_{i=1}^{I}\sum_{q=1}^{Q}\left(y_{iq} - \sum_{p=1}^{P}\sum_{j=1}^{J} u_{ip}\bar{x}_{pj}a_{jq} \right)^2 \rightarrow \min_{\mathbf{U},\underline{\bar{\mathbf{Y}}}} \tag{16}$$

subject to

U being a binary and row-stochastic matrix;
$\mathbf{A}'\mathbf{A} = \mathbf{I}_Q$;

Herewith (16) corresponds to the within-class deviance of the component matrix $\mathbf{Y} = \mathbf{XA}$, and $\mathbf{X}$ is column centered, The minimization of (16) subject to constraints (2) - (5) has been discussed by Vichi (2002), who has introduced the specific orthonormal constraints (4) - (5).

The partitioning of the objects in factorial *k*-means induces the following decomposition of the total deviance of the component score matrix $\mathbf{Y}$,

$$||\mathbf{XA}||^2 = ||\mathbf{XA} - \mathbf{U}\overline{\mathbf{X}}\mathbf{A}||^2 + ||\mathbf{U}\overline{\mathbf{X}}\mathbf{A}||^2, \quad (17)$$

where the first term on the right hand side is the *within-class deviance* of the partition given by $\mathbf{U}$ in the reduced space - minimized by the factorial *k*-means- and the second term is the *between-class deviance* of the partition in the reduced space. ♦

Thus, from Remarks 1 and 2 we can conclude that factorial *k*-means minimizes the within-class deviance $||\mathbf{Y} - \mathbf{U}\overline{\mathbf{X}}\mathbf{A}||^2$ of $\mathbf{Y}$, induced by the partition of the objects; while the solution of CDPCA maximizes the between-class deviance $||\mathbf{U}\overline{\mathbf{X}}\mathbf{A}||^2$ of $\mathbf{Y}$ induced by the partition of the objects, as in the reduced *k*-means. Here, in CDPCA, we prefer to maximize the between-class deviance of the reduced space, because we are particularly interested to define factors of maximal variance to be used to specify the classification of the variables, and this is guaranteed only if the between-class deviance is maximized as in reduced *k*-means.

The relationships between the four methods discussed can be summarized as in the table below.

Criterion to optimize	$\mathbf{A}'\mathbf{A}=\mathbf{I}_Q$	Constraints (4) - (5)
Maximizes between cluster deviance of reduced space	Reduced *k*-means (DeSoete Carroll, 1994)	CDPCA
Minimizes within cluster deviance of reduced space	Factorial k-means (Vichi, Kiers 2001)	Vichi (2002)

4. Least-squares estimation of the clustering and disjoint PCA and ALS algorithm

The minimization of (7) or alternatively the maximization of (9) or (9′) (here we consider to maximize (9)) subject to constraints (2), (3), (4) and (5) can be solved by using a constrained alternating least-squares (ALS) approach. Each parameter matrix of CDPCA is updated in turn by maximizing (9) with respect to one of the parameter matrices conditionally upon the others. The loss function (9) increases at each step, or at least never decreases, and the algorithm stops when the loss increment is less than a fixed, arbitrary positive and small threshold. Since function (9) is bounded above, the monotonicity property of the algorithm guarantees that the sequence of function values converges to a stationary point, which usually turns out to be, at least, a local maximum of the problem.

The three basic steps of the algorithm can be described as follows: (*i*) update $\mathbf{U}$, given the current estimate of $\mathbf{A}$ and $\overline{\mathbf{X}}$, subject to the binary and row-stochastic constraints on $\mathbf{U}$; then: (*ii*) update $\overline{\mathbf{X}}$, given current $\mathbf{A}$ and $\mathbf{U}$; and finally (*iii*) update $\mathbf{A}$, given $\mathbf{U}$ and $\overline{\mathbf{X}}$, subject to the constraints (4) and (5) on $\mathbf{A}$.

It can be observed that the orthogonality of $\mathbf{Y}$ is not required. This further constraint can be avoided since components correspond to different subsets of variables. Thus, if two components are highly correlated this would imply that there is no reason to have two clusters of variables and really these should be merged into a unique group.

The estimation of matrices $\mathbf{U}$ and $\overline{\mathbf{X}}$, at steps (*i*) and (*ii*), are obtained by solving an assignment and a regression problem, respectively, as we will describe in the next section. The optimization problem involved at point (*iii*) is not standard and a sequential quadratic program (SQP) algorithm could be used. In fact, SQP is known to solve efficiently non linear problems with continuous variables and non linear constraints. However, the constrained PCA problem described at point (*iii*), only apparently is a problem in continuous variables a_{jq}. In fact, when $a_{jq} \neq 0$, for satisfying constraints (5) any a_{jr} (r=1,…,Q; $r\neq q$) must be zero, thus inducing a combinatorial problem. In the case of mixed problems with continuous and discrete variables, SQP is not the most appropriate optimization algorithm, and therefore, a more efficient algorithm is required.

Before the least-squares estimation is discussed, let us first reformulate the clustering and disjoint PCA in order to point out its combinatorial and continuous nature also for point (*iii*).

Matrix **A** satisfying constraints (4) and (5) is rewritten into a product of two matrices **A**= **BV**, where **V** is the variable membership matrix which specifies the partition of variables and therefore the combinatorial part of our constrained PCA problem, while matrix **B** is a ($J\times J$) diagonal matrix that helps to specify the component loadings and represents the continuous part of the problem. The diagonal matrix **B** has the form

$$\mathbf{B} = \left(\sum_{q=1}^{Q} diag(\mathbf{v}_q) diag(\mathbf{c}_q)\right), \qquad (18)$$

where the notation $diag(\mathbf{a})$ specifies a diagonal matrix with diagonal equal to the vector **a** and $\mathbf{c}_q=[c_{1q}, \ldots, c_{jq}, \ldots, c_{Jq}]'$ is a J dimensional normalized vector used to find the q^{th} component loading.

From equation (18) it can be observed that when the $\mathbf{v}_q$ (q=1,…,Q) are known (i.e., the partition of the variables is known), the loadings $c_{jq} > 0$ are those for variables j such that v_{jq}=1. For each variable j' such that $v_{j'q}$=0, the product $v_{j'q}c_{j'q}$ is always null and therefore, without loss of generality, $c_{j'q}$ can be set equal to zero. Thus, matrix $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_q, \ldots, \mathbf{c}_Q]$ is a ($J \times Q$) orthonormal matrix for construction, that is $\mathbf{C}'\mathbf{C}=\mathbf{I}_Q$ and, furthermore, it satisfies constraints (4) and (5). Finally, $tr(\mathbf{BB})=Q$. An example can help to understand the reparametrization of the matrix **A**.

Example 2: The product of the following matrices

$$\mathbf{BV}=\begin{bmatrix}1&0&0&0&0&0&0&0\\0&1&0&0&0&0&0&0\\0&0&1&0&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\end{bmatrix}\times\begin{bmatrix}c_{11}&0&0&0&0&0&0&0\\0&c_{21}&0&0&0&0&0&0\\0&0&c_{31}&0&0&0&0&0\\0&0&0&c_{41}&0&0&0&0\\0&0&0&0&c_{51}&0&0&0\\0&0&0&0&0&c_{61}&0&0\\0&0&0&0&0&0&c_{71}&0\\0&0&0&0&0&0&0&c_{81}\end{bmatrix}+\begin{bmatrix}0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&1&0&0&0&0\\0&0&0&0&1&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\end{bmatrix}\times\begin{bmatrix}c_{12}&0&0&0&0&0&0&0\\0&c_{22}&0&0&0&0&0&0\\0&0&c_{32}&0&0&0&0&0\\0&0&0&c_{42}&0&0&0&0\\0&0&0&0&c_{52}&0&0&0\\0&0&0&0&0&c_{62}&0&0\\0&0&0&0&0&0&c_{72}&0\\0&0&0&0&0&0&0&c_{82}\end{bmatrix}+\begin{bmatrix}0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&0&1&0&0\\0&0&0&0&0&0&1&0\\0&0&0&0&0&0&0&1\end{bmatrix}\times\begin{bmatrix}c_{13}&0&0&0&0&0&0&0\\0&c_{23}&0&0&0&0&0&0\\0&0&c_{33}&0&0&0&0&0\\0&0&0&c_{43}&0&0&0&0\\0&0&0&0&c_{53}&0&0&0\\0&0&0&0&0&c_{63}&0&0\\0&0&0&0&0&0&c_{73}&0\\0&0&0&0&0&0&0&c_{83}\end{bmatrix}$$

$$=\overset{\mathbf{B}}{\begin{bmatrix}c_{11}&0&0&0&0&0&0&0\\0&c_{21}&0&0&0&0&0&0\\0&0&c_{31}&0&0&0&0&0\\0&0&0&c_{42}&0&0&0&0\\0&0&0&0&c_{52}&0&0&0\\0&0&0&0&0&c_{63}&0&0\\0&0&0&0&0&0&c_{73}&0\\0&0&0&0&0&0&0&c_{83}\end{bmatrix}}\times\overset{\mathbf{V}}{\begin{bmatrix}1&0&0\\1&0&0\\1&0&0\\0&1&0\\0&1&0\\0&0&1\\0&0&1\\0&0&1\end{bmatrix}}$$

gives the matrix

$$\begin{array}{ccc}\mathbf{c}_1&\mathbf{c}_2&\mathbf{c}_3\\ \hline c_{11}&0&0\\c_{21}&0&0\\c_{31}&0&0\\0&c_{42}&0\\0&c_{52}&0\\0&0&c_{63}\\0&0&c_{73}\\0&0&c_{83}\end{array}$$

which is by construction orthonormal, i.e., $\mathbf{C}'\mathbf{C}=\mathbf{I}_Q$. Furthermore, $\mathbf{C}$ satisfies also the constraints (5), i.e., $\sum_{j=1}^{J}(c_{jq}c_{jr})^2=0$,q=1,…,2; r=q+1, …, 3. If c_{11} =0.5593 , c_{21} =0.6617, c_{31} =0.4994, c_{42} =-0.7993, c_{52} =0.6010, c_{63} =0.3685, c_{63} =6131, c_{63} =-0.6988, the above matrix **BV** is equal to matrix **A** in example 1. ♦

The reparameterization (18) splits the estimation of the matrix **A** into two parts: the membership matrix **V** for the partition of the variables and the component loadings $\mathbf{c}_q$ to specify the diagonal matrix **B** as indicated in (18).
Let us now describe the estimation of the parameters of CDPCA.
The sum of squares estimation problem (9′) to be maximized can be rewritten,

$$
\begin{aligned}
F(\mathbf{B},\overline{\mathbf{X}},\mathbf{U},\mathbf{V}) &= \|\mathbf{U}\overline{\mathbf{X}}\mathbf{B}\mathbf{V}\|^2 = tr(\mathbf{V}'\mathbf{B}'\overline{\mathbf{X}}'\mathbf{U}'\mathbf{U}\overline{\mathbf{X}}\mathbf{B}\mathbf{V}) = \\
&= tr[\mathbf{V}'\left(\sum_{q=1}^{Q} diag(\mathbf{v}_q)diag(\mathbf{c}_q)\right)'\overline{\mathbf{X}}'\mathbf{U}'\mathbf{U}\overline{\mathbf{X}}\left(\sum_{q=1}^{Q} diag(\mathbf{v}_q)diag(\mathbf{c}_q)\right)\mathbf{V}] = \\
&= \sum_{p=1}^{P}\sum_{q=1}^{Q}\left(\sum_{j=1}^{J} v_{jq}c_{jq}\overline{x}_{pj}\right)^2\sum_{i=1}^{I}u_{ip}\ , \qquad (19)
\end{aligned}
$$

subject to

$$u_{ip}\in\{0,\ 1\},\qquad (i=1,\ldots,I;\ p=1,\ldots,P); \qquad (20)$$

$$\sum_{p=1}^{P}u_{jp}=1,\qquad (i=1,\ldots,I); \qquad (21)$$

$$v_{jq}\in\{0,\ 1\},\qquad (j=1,\ldots,J;\ q=1,\ldots,Q); \qquad (22)$$

$$\sum_{q=1}^{Q}v_{jq}=1,\qquad (j=1,\ldots,J); \qquad (23)$$

$$\sum_{j=1}^{J}c_{jq}^2=1\qquad q=1,\ \ldots,\ Q; \qquad (24)$$

$$\sum_{j=1}^{J}c_{jq}c_{jr}=0\quad q=1,\ldots,Q\text{-}1;\ r=q+1,\ldots,Q. \qquad (25)$$

Thus, the clustering and disjoint principal component analysis -defined by maximizing (9) or (9′) subject to constraints (2) - (5)- has been reformulated into the equivalent problem to maximize (19) with respect to binary variables u_{ip}, v_{jq} and real variables c_{jq}, subject to (20) – (25).
Equation (19) can be simplified by rewriting,

$$
\begin{aligned}
F(\mathbf{C},\overline{\mathbf{X}},\mathbf{U},\mathbf{V}) = &\sum_{q=1}^{Q} tr\,[\mathbf{V}'diag(\mathbf{c}_q)\ diag(\mathbf{v}_q)\overline{\mathbf{X}}'\mathbf{U}'\mathbf{U}\overline{\mathbf{X}}\,diag(\mathbf{v}_q)\ diag(\mathbf{c}_q)\mathbf{V}] + \\
&+ \sum_{q=1}^{Q}\sum_{\substack{r=1\\ r\neq q}}^{Q} tr\,[\ \mathbf{V}'diag(\mathbf{c}_q)\ diag(\mathbf{v}_q)\overline{\mathbf{X}}'\mathbf{U}'\mathbf{U}\overline{\mathbf{X}}\,diag(\mathbf{v}_r)\ diag(\mathbf{c}_r)\mathbf{V}], \qquad (26)
\end{aligned}
$$

and by observing that the second term of (26) vanishes, because for each r and q with ($r\neq q$),

$$diag(\mathbf{v}_q)\ diag(\mathbf{c}_q)\ \mathbf{V}\mathbf{V}'\ diag(\mathbf{c}_r)\ diag(\mathbf{v}_r) = \mathbf{0},$$

is a ($J \times J$) null matrix, when constraints (22) (23) and (25) are satisfied.

4.1 Update of C and B

The maximization of (26), when $\hat{\bar{\mathbf{X}}}, \hat{\mathbf{U}}, \hat{\mathbf{V}}$ are estimated, involves the maximization with respect to $\mathbf{c}_q$ q=1,…,Q of Q independent sub-problems of PCA. In fact, for each column $\mathbf{c}_q$ of **C**, (q=1,…,Q) it is necessary to solve,

$$F(\mathbf{c}_q, \hat{\bar{\mathbf{X}}}, \hat{\mathbf{U}}, \hat{\mathbf{V}}) = tr[\hat{\mathbf{V}}' \, diag(\mathbf{c}_q) \, diag(\hat{\mathbf{v}}_q) \hat{\bar{\mathbf{X}}}' \hat{\mathbf{U}}'\hat{\mathbf{U}} \hat{\bar{\mathbf{X}}} \, diag(\hat{\mathbf{v}}_q) \, diag(\mathbf{c}_q) \hat{\mathbf{V}}] \tag{27}$$

subject to

$$\sum_{j=1}^{J} c_{jq}^2 = 1 \qquad q=1, \ldots, Q; \tag{28}$$

$$\sum_{j=1}^{J} c_{jq} c_{jr} = 0 \quad q=1,\ldots,Q\text{-}1;\ r=q+1,\ldots,Q. \tag{29}$$

Now, matrix $diag(\mathbf{c}_q)\hat{\mathbf{V}}$ is a ($J \times Q$) matrix with the q^{th} column equal to $\mathbf{c}_q$, while all the other columns are formed by null elements. Therefore, the vector $\mathbf{c}_q$ that maximizes (27) is the eigenvector associated to the largest eigenvalue of the matrix $diag(\hat{\mathbf{v}}_q) \hat{\bar{\mathbf{X}}}' \hat{\mathbf{U}}'\hat{\mathbf{U}} \hat{\bar{\mathbf{X}}} \, diag(\hat{\mathbf{v}}_q)$.

The update of **B** is given by (18), that is, $\mathbf{B} = \left(\sum_{q=1}^{Q} diag(\hat{\mathbf{v}}_q) diag(\hat{\mathbf{c}}_q) \right)$.

4.2. Update of V

The maximization of $F(\hat{\mathbf{C}}, \hat{\bar{\mathbf{X}}}, \hat{\mathbf{U}}, \mathbf{V})$ with respect to **V**, when $\hat{\mathbf{C}}, \hat{\bar{\mathbf{X}}}$ and $\hat{\mathbf{U}}$ are fixed, is obtained for each j (j =1,…,J) by computing

$$v_{jq} = 1 \quad \textit{if} \qquad F(\hat{\mathbf{c}}_q, \hat{\mathbf{U}}, \hat{\bar{\mathbf{X}}}, [v_{jq}]) = \max\{F(\hat{\mathbf{c}}_r, \hat{\mathbf{U}}, \hat{\bar{\mathbf{X}}}, [v_{jr}=1]) \colon r=1,..Q;\ (r \neq q)\} \tag{30}$$

$$v_{jq} = 0 \quad \text{otherwise.}$$

When v_{jr} is fixed equal to 1, $\mathbf{c}_r$ is accordingly updated following the procedure described above, to maximize (26) with respect to $\mathbf{c}_r$. Therefore, the update of **V** induces the consequent update of the columns of **C**.

4.3 Update of $\bar{\mathbf{X}}$

The maximization of (19), with respect to $\bar{\mathbf{X}}$ when $\hat{\mathbf{B}}, \hat{\mathbf{U}}, \hat{\mathbf{V}}$ are fixed is equivalent to the minimization,

$$||\mathbf{X}\hat{\mathbf{B}}\hat{\mathbf{V}} \ \text{-}\ \hat{\mathbf{U}}\bar{\mathbf{X}}\hat{\mathbf{B}}\hat{\mathbf{V}}||^2, \tag{31}$$

as it can be seen from the decomposition (17) by setting **A**=**BV**. The minimization of (31) corresponds to the solution of the multivariate regression problem,

$$\overline{\mathbf{X}} = (\hat{\mathbf{U}}'\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}'\mathbf{X}. \tag{32}$$

4.4. Update of U

The maximization of $F(\hat{\mathbf{B}}, \hat{\overline{\mathbf{X}}}, \mathbf{U}, \hat{\mathbf{V}})$ with respect to **U**, when $\hat{\mathbf{B}}$, $\hat{\overline{\mathbf{X}}}$ and $\hat{\mathbf{V}}$ are estimated, by considering (31), is equivalent to the minimization

$$\|\mathbf{X}\hat{\mathbf{B}}\hat{\mathbf{V}} \ - \mathbf{U}\hat{\overline{\mathbf{X}}}\ \hat{\mathbf{B}}\hat{\mathbf{V}}\|^2 = \sum_{i=1}^{I}\sum_{p=1}^{P}\left\|\hat{\mathbf{V}}'\hat{\mathbf{B}}\mathbf{x}_i - \hat{\mathbf{V}}'\hat{\mathbf{B}}\hat{\overline{\mathbf{x}}}_p\right\|^2 u_{ip}, \tag{33}$$

which involves the minimization of I sub-problems of clustering, (i=1,…,I),

$$\sum_{p=1}^{P}\left\|\hat{\mathbf{V}}'\hat{\mathbf{B}}\mathbf{x}_i - \hat{\mathbf{V}}'\hat{\mathbf{B}}\hat{\overline{\mathbf{x}}}_p\right\|^2 u_{ip} \ , \tag{34}$$

subject to

$$u_{ip} \in \{0, 1\}, \qquad (i=1,\ldots,I;\ p=1,\ldots,P); \tag{35}$$

$$\sum_{p=1}^{P} u_{ip} = 1. \qquad (i=1,\ldots,I). \tag{36}$$

Problem (34) subject to (35) and (36) is an assignment problem solved in linear time by fixing,

$$u_{ip} = 1 \quad \text{if} \quad \left\|\hat{\mathbf{V}}'\hat{\mathbf{B}}\mathbf{x}_i - \hat{\mathbf{V}}'\hat{\mathbf{B}}\hat{\overline{\mathbf{x}}}_p\right\|^2 = \min\{\left\|\hat{\mathbf{V}}'\hat{\mathbf{B}}\mathbf{x}_i - \hat{\mathbf{V}}'\hat{\mathbf{B}}\hat{\overline{\mathbf{x}}}_s\right\|^2 : s=1,\ldots,P;\ s\neq p\}, \tag{37}$$

$$u_{ip} = 0 \quad \text{otherwise.}$$

We are now in position to state an alternating least-squares algorithm.

Remark 3. Let us include the reparameterization of the matrix **A** into model (1) and suppose that the matrix $\mathbf{B}=\mathbf{I}_J$, i.e., **B** is the identity matrix of order J, thus model (1) writes

$$\mathbf{X} = \mathbf{U}\,\overline{\mathbf{Y}}\,\mathbf{V}' + \mathbf{E}, \tag{38}$$

which implies that all the variables have the same loadings equal to 1. The CDPCA degenerates into the double *k*-means (Vichi, 2000), that specifies a partition, both for objects and variables, in *P* classes and *Q* classes, respectively. Thus we can conclude that double *k*-means is a relevant case of the more general CDPCA. In double *k*-means both objects and variables are synthesized by mean profiles of objects belonging to the object class and mean profiles for variables belonging to the variable class. In CDPCA there is an asymmetric treatment of the two modes of the data matrix. Objects are synthesized by mean profiles of clusters, while components are synthesized by linear combinations.♦

4.5. An Alternating Least-Squares algorithm for clustering and disjoint PCA

The constrained problem of maximizing (9) or (9′) can be solved by using an alternating least-squares (ALS) algorithm, which alternates four steps: update **V** (allocation of variables) and **B** (the PCA step) update the centroid matrix $\overline{\mathbf{X}}$ and finally update **U** (the allocation of objects).

Initialization . Initial values are chosen for **U** and **V**. Such values can be chosen randomly or in a rational way (e.g., based on the *k*-means clustering solution applied on matrix **X** and **X**') and, in both cases, they should satisfy the constraints on **U** and **V**. For $\bar{\mathbf{X}}$ formula (32) is used.

Step 1. **B** is updated, given the current $\hat{\bar{\mathbf{X}}}, \hat{\mathbf{U}}, \hat{\mathbf{V}}$ by maximizing $F(\mathbf{B}, \hat{\mathbf{U}}, \hat{\bar{\mathbf{X}}}, \hat{\mathbf{V}})$ over each column $\mathbf{c}_q$ of **C**. The vector $\mathbf{c}_q$ is the eigenvector associated to the largest eigenvalue of the matrix $diag(\mathbf{v}_q)\hat{\bar{\mathbf{X}}}'\hat{\mathbf{U}}'\hat{\mathbf{U}}\hat{\bar{\mathbf{X}}}diag(\mathbf{v}_q)$. In addition, a rotation of this solution can be also chosen, because it does not affect the scalar products. To update **B** formula (18) is used.

Step 2. **V** is updated, given the current estimate of $\hat{\mathbf{B}}, \hat{\bar{\mathbf{X}}}$ and $\hat{\mathbf{U}}$. This problem is sequentially solved for the different rows of **V** by taking

v_{jq}=1, if $F(\hat{\mathbf{c}}_q, \hat{\mathbf{U}}, \hat{\bar{\mathbf{X}}}, [v_{jq}]) = \max\{F(\hat{\mathbf{c}}_r, \hat{\mathbf{U}}, \hat{\bar{\mathbf{X}}}, [v_{jr}=1]): r=1,..Q; (r\neq q)\}$or v_{jq}=0 otherwise.

Step 3. **U** is updated, given the current $\hat{\mathbf{B}}$, $\hat{\bar{\mathbf{X}}}$ and $\hat{\mathbf{V}}$. This problem is solved for the different rows of **U** by taking

u_{ip} = 1, if $\left\|\hat{\mathbf{V}}'\hat{\mathbf{B}}\mathbf{x}_i - \hat{\mathbf{V}}'\hat{\mathbf{B}}\hat{\bar{\mathbf{x}}}_p\right\|^2 = \min\{\left\|\hat{\mathbf{V}}'\hat{\mathbf{B}}\mathbf{x}_i - \hat{\mathbf{V}}'\hat{\mathbf{B}}\hat{\bar{\mathbf{x}}}_s\right\|^2 : s=1,\ldots,P; s\neq p\}$ or u_{ip} = 0 otherwise.

Step 4. $\bar{\mathbf{X}}$ is updated, given $\hat{\mathbf{B}}, \hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ by: $\bar{\mathbf{X}} = (\hat{\mathbf{U}}'\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}'\mathbf{X}$.

Stopping rule. The function value $F(\hat{\mathbf{B}}, \hat{\bar{\mathbf{X}}}, \hat{\mathbf{U}}, \hat{\mathbf{V}})$ is computed for the current values of $\hat{\mathbf{B}}$, $\hat{\bar{\mathbf{X}}}, \hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$. When such updated values have increased considerably (more than an arbitrary small convergence tolerance value) the function value, **B**, $\bar{\mathbf{X}}$, **U** and **V** are updated once more according to Steps 1 - 4. Otherwise, the process is considered to have converged.

The algorithm monotonically increases the objective function and, since function $F(\mathbf{B}, \bar{\mathbf{X}}, \mathbf{U}, \mathbf{V})$ is bounded above by the total variance of **X**, it will converge to a stationary point which can be expected to be at least a local maximum. To increase the chance of finding the global maximum, the algorithm should be run several times, with different initial estimates for **V**. In our experiences we observed that at least 30 runs are necessary.

5. Application of Clustering and Disjoint PCA

The clustering and disjoint PCA has been applied to two real data sets to show the performances of the new methodology. The first data set describes the short-term scenario of the OECD countries analyzed also with factorial *k*-means The second data set takes into account the socio and macroeconomic performances of national economies of 103 countries.

5.1. *SHORT-TERM MACROECONOMIC SCENARIO OF OECD COUNTRIES*

The short-term scenario (September 1999) on macroeconomic performance of national economies of twenty countries, members of the Organization for Economic Co-operation and Development (OECD) has been analyzed in Vichi & Kiers, (2001) to test the ability of the factorial *k*-means analysis in identifying classes of similar economies and help to understand the relationships within the set of observed economic indicators. Six main economic indicators were considered: Gross Domestic Product (GDP), Leading Indicator (LI), Unemployment Rate (UR), Interest Rate (IR), Trade Balance (TB), Net National Savings (NNS). Variables were standardized by unit variance.

A *tandem analysis* was carried out computing the first two principal components and classifying countries on the basis of first two objects scores. The results are shown in Figure 1a. The first two components explain 29% and 24% of the total variance. The first PCA dimension is characterized mainly by net national savings (21%), gross domestic product (18%), unemployment rate (14%), (in fact, the other three variables explain 4%), while 43% is due to the interrelations (sum of the covariances) among the six variables. The second PCA dimension is characterized by interest rate (34%) and trade balance (22%) and unemployment rate (9%), (the remaining three variables explain 4%), while 31% is due to interrelations among the six variables. Note that unemployment rate characterizes both dimensions. The *k*-means algorithm was run on the first two PCA starting from random partitions. It was necessary to run *k*-means for a large number of initial random starts for the presence of several local optima (Running 10000 times *k*-means found the present optimal solution after 4650 runs). Clusters of countries, in Figure 1a, are highlighted by dotted ellipses. The between cluster deviance of the optimal solution was equal to 39.91% of the total deviance.

The results of the CDPCA are reported in Figure 1b, by considering a *biplot* representation (Gabriel, 1971). The three clusters of countries are also highlighted by three dotted ellipses, while the two clusters of variables are represented by the two dotted orthogonal axes. The CDPCA was run 500 times to increase the chance to find the global optimal solution. This optimal solution was found 5 times in the 500 runs. In the best solution the algorithm converged after between 5 and 6 iterations of the steps 1 – 4 (with convergence tolerance value equal to 10^{-5}). The two components of the clustering and disjoint PCA explain almost the same part of variance explained by the PCA (27% and 22%, respectively).

The first component of the disjoint PCA is characterized mainly by net national savings (37%), unemployment rate (15%) and gross domestic product (9%); while 39% contribution is due to interrelations among GDP, UR and NNS. Note that in CDPCA unemployment rate explains the first component only differently from PCA. The second disjoint PCA component is characterized by interest rate (37%), trade balance (35%) and only slightly by leading indicator (4%), while 24% contribution is due to interrelations among IR, LI and TB.

Therefore, the components of the PCA are almost the same of the disjoint PCA both in terms of variance explained and of variables most contributing to specify the components; however, the disjoint PCA clearly shows more homogeneous clusters (between cluster deviance of CDPCA is $||\mathbf{U}\,\bar{\mathbf{Y}}\,||^2/||\mathbf{Y}||^2$=78% of total deviance, while between cluster deviance of the tandem analysis is about 40% of the total deviance only).

Recalling results of Factorial *k*-means (FKM) (Vichi, Kiers, 2001) the first component is specified by the main contribution of TB, IR, NNS, LI; while the second component is mainly characterized by GDP, and in part IR. Finally, UR contributes equally to both components. Comparing these results with CDPCA it can be observed that there are variables such UR and IR for FKM that can contribute to both components, as it happens with LI and UR for PCA (in tandem analysis). This ambiguity in the explanation of components causes difficulties in their interpretation and description due to overlap of the same variables in explaining different factors. Of course, differences of components of FKM with respect to CDPCA and tandem analysis, produces some differences in the classification of countries with respect to FKM and tandem analysis.

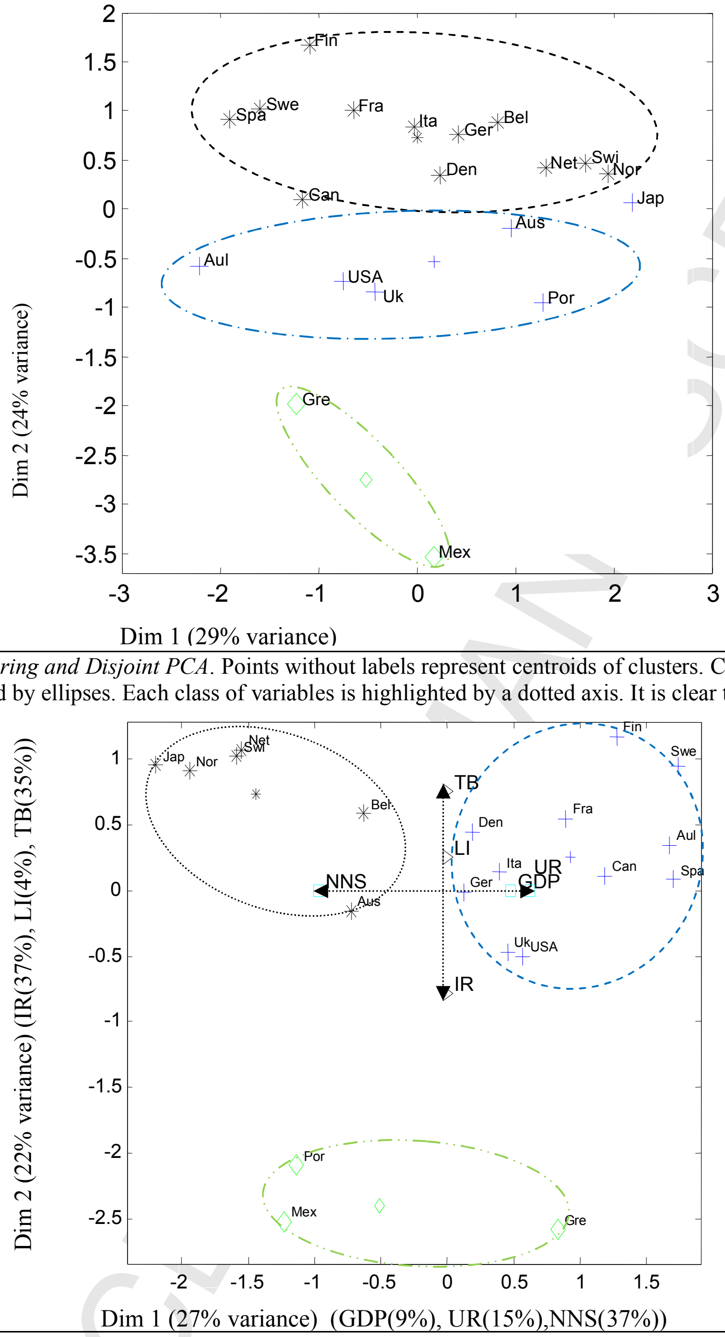
The choice of the number of clusters for objects can be guided by the indices introduced in clustering literature. In a study comparing different indices Milligan and Cooper (1985) suggested to use the ratio of the between and within deviances, each one divided by their respective degree of freedom. For variables the number of clusters should be smaller or equal to the maximum number of eigenvalues since more clusters than latent variables are expected.

By examining the clusters of variables in our example it can be observed that none of the two have more than one eigenvalue greater than one. Therefore, CDPCA has found two clusters of variables each one explained by a subset of variables expressing a component variable only. On the other hand if three clusters of variables are required the best solution of CDPCA (over 500 random starts) divides the first group of variables leaving GDP alone.

It has been already noted that CDPCA produces components that not need to be orthogonal; however when the number of clusters for the variables is correctly chosen the correlation between components is frequently small. In this case study the correlation between component 1 and 2 is -0.0011.

Figure 1: a) *tandem analysis*; b) clustering and disjoint PCA.

a) *Tandem analysis*: *k*-means clustering (*k*=3) computed on the first two principal components (10000 random start for *k*-means). Points without labels represent the centroid of clusters. Clusters of countries are highlighted by ellipses.



b) *Clustering and Disjoint PCA*. Points without labels represent centroids of clusters. Clusters of countries are highlighted by ellipses. Each class of variables is highlighted by a dotted axis. It is clear that axes are orthogonal.

Table 1: Component loadings for PCA and CDPCA

	PCA loadings Component 1	PCA loadings Component 2	CDPCA loadings Component 1	CDPCA loadings Component 2
Gross Domestic Product (GDP)	**0.5669**	0.0650	**0.3831**	0
Interest Rate (IR)	0.1745	**0.6956**	0	**0.6972**
Leading Indicator (LI)	**0.1922**	**0.2289**	0	**-0.2287**
Unemployment Rate (UR)	**0.4893**	**-0.3668**	**0.4978**	0
Net National Savings (NNS)	**-0.6069**	0.0923	**-0.7781**	0
Trade Balance (TB)	-0.0592	**-0.5625**	0	**-0.6794**

5.2 *MACRO ECONOMIC WORLD PERFORMACE IN 2000.*

The socio and macroeconomic performance of national economies of 103 countries has been used to test the ability of the clustering and disjoint PCA to identify clusters of homogeneous countries and classes of variables formed along components of maximum variance. Countries performances were evaluated according to 12 basic indicators of: demography and fertility, health, education, urbanization and economy; they refer to the year 2000. Most indicators are drawn from national sources of regular administrative files and collected together by international institutions. In particular most of these indicators are from World Bank (World development indicators and World development report). For unemployment rate the files of I.L.O. and F.A.O. have been used. Indicators and countries are shown in Tables 2 and 3.

Table 2: List of variables used in the analysis (year 2000)

Demography and Fertility
Crude Birth Rate (per 1.000 people), (CRB)
Crude Death Rate (per 1.000 people), (CDR)
Health
Infant Mortality Rate (per 1.000 live birth), (IMR)
Life Expectancy at Birth (years), (LEB)
Education
Adult Illiteracy Rate (% of people ages 15 and over), (AIR)
Urbanization
Urban Population (% of total population), (UP)
Rural Population Density (people per sq. km. of arable land), (RPD)
Economic (basic),
Gross Domestic Product per Capita (at PPP at 1996 prices US$), (GDP)
Exports Goods and Services (% of GDP), (EXP)
Imports Goods and Services (% of GDP), (IMP)
Unemployment Rate (% of total labor force), (UR)
Economically Active Population Rate (% of population), (EAP)

Table 3: Alphabetical list of 103 countries used in the analysis with the labels used in Figures 2 and 3.

Afghanistan (AFG), Argentina (ARG), Armenia (ARM), Australia (AUS), Austria (AUT), Azerbaijan (AZE),
Bahrain (BHR), Bangladesh (BGD), Barbados (BRB), Belarus (BLR), Belgium (BEL), Bolivia (BOL), Bosnia and Herzegovina (BIH), Botswana (BWA), Brazil (BRA), Bulgaria (BGR), Burkina Faso (BFA), Burundi (BDI),
Cambodia (KHM), Cameroon (CMR), Canada (CAN), Central African Rep. (CAF), Chile (CHL), China (CHN), Colombia (COL), Costa Rica (CRI), Croatia (HRV), Cuba (CUB), Czech Repub. (CZE),
Denmark (DNK), Dominican R. (DOM),
Ecuador (ECU), Egypt, Arab R. (EGY), El Salvador (SLV),Estonia (EST), Ethiopia (ETH),
Finland (FIN), France (FRA),
Gabon (GAB), Georgia (GEO), Germany (DEU), Greece (GRC), Guatemala (GTM),
Haiti (HTI), Honduras (HND), Hong Kong (HKG), Hungary (HUN),
Iceland (ISL), India (IND), Indonesia (IDN), Iran, Islam Re. (IRN), Ireland (IRL), Israel (ISR), Italy (ITA),
Japan (JPN), Jordan (JOR),
Kenya (KEN), korea, Dem. Rep (PRK), Korea, Rep. (KOR), Kuwait (KWT),
Latvia (LVA), Lebanon (LBN), Liberia (LBR), Libya (LBY), Lithuania (LTU), Malaysia (MYS), Mexico (MEX), Morocco (MAR),
Netherlands (NLD), New Zealand (NZL), Nicaragua (NIC), Nigeria (NGA), Norway (NOR),
Pakistan (PAK), Panama (PAN), Paraguay (PRY), Peru (PER), Philippines (PHL), Poland (POL), Portugal (PRT),
Romania (ROM), Russia (RUS),
Saudi Arabia (SAU), Senegal (SEN), Singapore (SGP), South Africa (ZAF), Spain (ESP), Sweden (SWE), Switzerland (CHE), Syrian Arab R. (SYR),
Tanzania (TZA), Thailand (THA), Turkey (TUR),
Uganda (UGA), Ukraine (UKR), United Arab Emirates (ARE), United Kingdom (GBR), United States (USA), Uruguay (URY), Uzbekistan (UZB),
Venezuela (VEN), Vietnam (VNM),
Zambia (ZMB),

Variables have been column standardized to take into account different scale measurements. First, PCA has been applied on the (103 × 12) matrix of standardized variables. Three components have been retained corresponding to eigenvalues larger than 1, which account for 73.94% of the total variance. The component loading matrix, after varimax rotation is reported in Table 4 (a). It can be observed that even after rotation some variables have loadings larger than 0.1. The highest values are highlighted in black, while values larger than 0.1 in grey.

Table 4 (a), (b) Component loadings of the PCA and CDPCA on the 103 countries × 12 macroeconomic variables

	a)	PCA		b)	CDPCA	
Variables	Comp. 1	Comp. 2	Comp. 3	Comp. 1	Comp. 2	Comp. 3
	44.44%	14.99%	12.36%	34.50%	14.00%	9.49%
Crude Birth Rate (CBR)	**-0.3814**	-0.0182	0.2242	0.4304	0	0
Crude Death Rate (CDR)	-0.2952	0.1121	**-0.4409**	0.3293	0	0
Infant Mortality Rate (IMR)	**-0.4258**	0.0318	-0.0233	0.4596	0	0
Life Expectancy at Birth (LEB)	**0.4351**	-0.0054	0.1051	-0.4619	0	0
Adult Illiteracy Rate (AIR)	**-0.3720**	-0.0589	0.1623	0.3933	0	0
Urban Population (UP)	**0.3788**	0.0049	0.1552	-0.3564	0	0
Rural Population Density (RPD)	-0.1264	**-0.4803**	-0.0019	0	0.3655	0
Gross Domestic Product per Capita (DGP)	**0.3051**	-0.0002	-0.1917	0	0	0.8040
Unemployment Rate (UR)	-0.0322	0.0562	**0.3988**	0	0	-0.3160
Economically Active Population Rate (EAP)	0.0377	-0.0344	**-0.7051**	0	0	0.5038
Imports Goods and Services (IMP)	0.0076	**-0.6233**	-0.0098	0	0.6473	0
Exports Goods and Services (EXP)	0.0577	**-0.5992**	-0.0165	0	0.6689	0

The analysis of clustering and disjoint PCA has been repeated by fixing the number of clusters for variables equal to Q=3, to compare this solution with the one given by PCA, while the number of clusters of units have been set to vary from 5 to 12. The results are reported in Table 5. Each time the analysis was repeated 30 times by starting with different random initial partitions both for variables and clusters and retaining the best solution. This multistart strategy is necessary to avoid that the algorithm stops into a local maximum of the problem. It can be observed that the larger increase of the accounted variance can be observed between 6 and 7 clusters, then stabilizes between 0.59 and 0.60, for a larger number of clusters of countries. Therefore, the solution with 7 clusters for the objects has been considered the most parsimonious, even accounting for a relevant part of the total variance. This strategy of choice has been introduced by Cattell, (1966) for choosing the number of factors, in a factorial analysis.

Table 5: Explained between variance (10) for different values of P = 5, 6, 7, 8, 9, 10, 11, 12 and Q=3.

# classes Objects	# classes variables	Explained Variance
5	3	0.5389
6	3	0.5586
7	3	0.5799
8	3	0.5874
9	3	0.5929
10	3	0.5988
11	3	0.6016
12	3	0.6049

The component loading matrix of the clustering and disjoint PCA is reported in Table 4 (b). The first component explains 34.50% of the total variance and it is characterized by demographic, health, education and urbanization variables. It represents a socio-demographic dimension. The second and third components represent two economic dimensions.

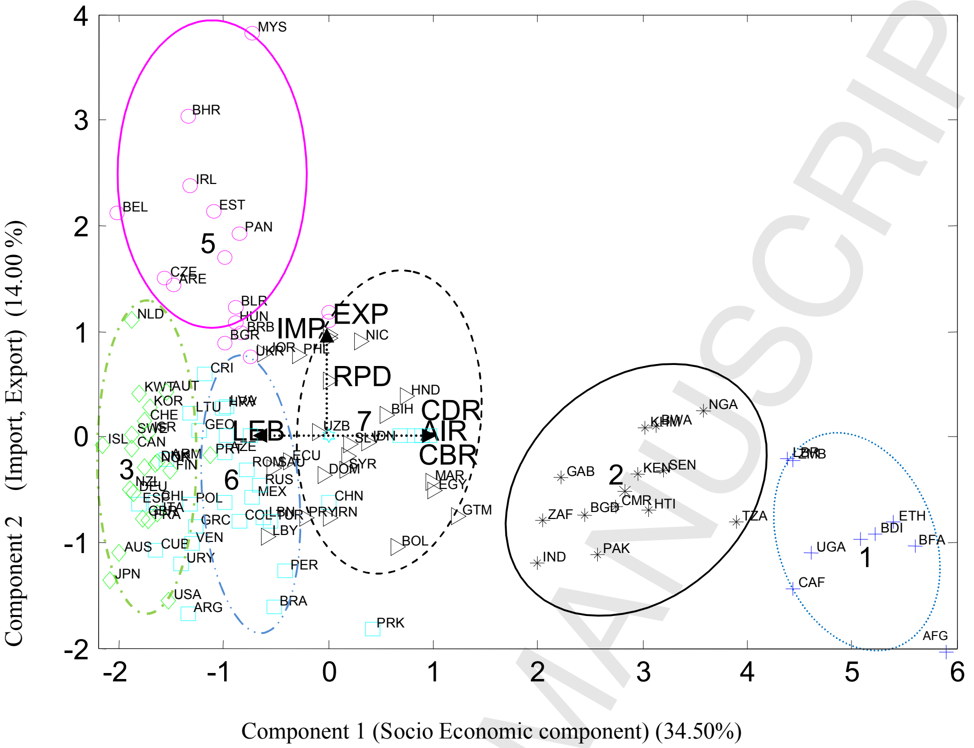
The classification of the countries is shown in table 6

Table 6: Partition of the 103 countries in 7 clusters according còustering and disjoint PCA.

Cluster 1	8	Afghanistan, Burkina Faso, Burundi, Central African Rep., Ethiopia, Liberia, Uganda, Zambia
Cluster 2	13	Bangladesh, Botswana, Cambodia, Cameroon, Gabon, Haiti, Kenya, India, Nigeria, Pakistan, Senegal, South Africa, Tanzania
Cluster 3	21	Australia, Austria, Canada, Denmark, Finland, France, Germany, Kuwait, Korea, Rep.,Iceland, Israel, Italy, Japan, Netherlands, New Zealand, Norway, Portugal, Sweden, Switzerland, United Kingdom, United States
Cluster 4	2	Hong Kong, Singapore
Cluster 5	15	Bahrain, Belgium, Barbados, Belarus, Bulgaria, Czech Repub., Etonia, Hungary, Ireland, Malaysia, Panama, Thailand, Ukraine, United Arab Emirates, Vietnam
Cluster 6	25	Argentina, Chile, Costa Rica, Lebanon, Mexico, Peru, Venezuela, China, Croatia, Georgia, Latvia, Lithuania, Poland, Russia, Greece, Spain, Armenia, Azerbaijan, Brazil, Colombia, Cuba, Korea Dem. Rep, Romania, Turkey, Uruguay
Cluster 7	19	Arab R., Bolivia, Bosnia and Herzegovina, Dominican R., Ecuador, Egypt, El Salvador, Guatemala, Honduras, Indonesia, Iran, Islam Re., Jordan, Libya, Morocco, Nicaragua, Paraguay, Philippines, Saudi Arabia, Syrian Arab R., Uzbekistan

The plot of the countries in the first two components plane is shown in Figure 2; while in Figure 3 the plot of the first and third component is displayed.

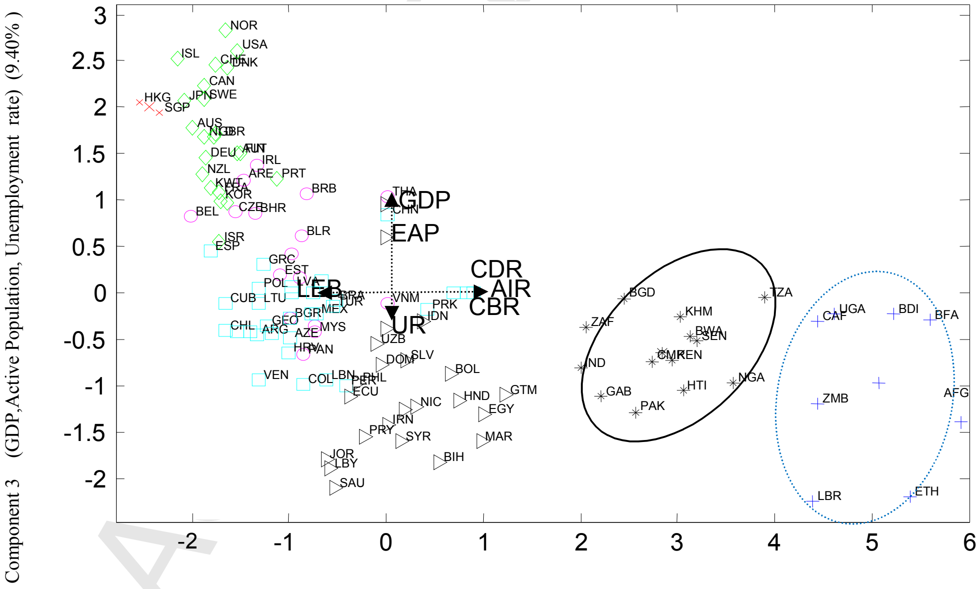
Figure 2: Macroeconomic Performance of World 2000. Plot of the component 1 (Socio-Economic) Vs component 2 (Import-Export) of CDPCA. Points with the number represent centroids of clusters. Clusters of countries are highlighted by ellipses.



Component 1 (Socio Economic component) (34.50%)

NB Cluster 4 formed by Singapore and Hong Kong, with centroid (-2.4; 7.2), for readability of the plot was not included in the Fig.

Figure 3: Plot of the component 1 (Socio-Economic) vs component 3 (GDP, Active population & Unemployment rate). Points with the number represent the centroids of the clusters. Clusters of countries are highlighted by ellipses.



Component 1 (Socio Economic component) (34.50%)

The correlation coefficient between component 1 and component 2 is equal to 0.5430, while the correlation coefficients between components 1 and 3 and 2 and 3 are equal to -0.3187 and -0.2713, respectively. Requiring two components with always 7 clusters for objects the correlation coefficient reduces to 0.3581, confirming that when there is some correlation between factors, as for components 1 and 2, it is possible to reduce it by reducing the number of clusters of variables.

This time, the first component explains 40% of total variance and it is equal to the previous first component plus Gross Domestic Product, while the second component (explaining 18%) is the second and the third component (unless GDP) of the previous case.

The classification remains almost the same so as the graphical representation and therefore is not reported. Of course by increasing the number of clusters for the countries considerably, the explained variance also increases; e.g., for P=30 and P=40 the explained variance is 0.6566, 0.6662, respectively; while for P=103 (corresponding to the case of partitioning only variables) the explained variance is 0.6836. However, in such cases the interpretability of the clusters becomes a more serious task.

6. Conclusions

In this paper a new methodology for partitioning simultaneously both objects around a set of centroids and variables along a set of components of maximum variance is presented. Since clustering is required for objects, the between cluster variance of the component scores is maximized. The new methodology is named Clustering and Disjoint Principal Component Analysis (CDPCA).

CDPCA is particularly appropriate when the researcher has to reduce both objects and variables for interpretability reasons; for this purpose, principal component analysis followed by a clustering methodology is frequently applied. This *tandem analysis* can be shown to produce ambiguous results, because PCA finds components with maximum variance, while only between variance has to be explained for the following partitioning methodology, such as k-means. In other terms the components of the PCA can explain variance that masks the clustering structure in the data (Vichi & Kiers, 2001). Two alternative techniques have been proposed in the literature to simultaneously classify objects and obtain a dimensionality reduction of variables. Factorial k-means (Vichi & Kiers, 2001), by minimizing the within cluster deviance in the reduced space, is particularly useful when clustering of objects is the major objective of analysis. However, in Factorial k-means the dimensionality reduction may not be guaranteed to explain the largest variance. On the other hand, reduced k- means by maximizing the between deviance of the reduced space is useful when dimensions explaining maximal variance are needed. However, in reduced k-means clustering with minimal variance may not be guaranteed. Both techniques do not classify variables and components may be explained by the same variables, thus inducing possible complications in the interpretation of components. CDPCA, as reduced k-means, maximizes the between deviance of the reduced space because it includes the classification of variables where each class is specified by a component. Hence, it is relevant to find components of maximal variance to obtain stable classes of the partition of variables. In synthesis, with CDPCA a double advantage is obtained: first, to identify a classification of variables and a classification of objects; second, to obtain a dimensionality reduction of the data matrix via a reduced set of centroids for objects and a reduced set of components (linear combinations) for variables.

Another important advantage of CDPCA is the easy interpretability of the components since each one is characterized by a disjoint set of variables. Therefore, with CDPCA it is not possible to have one observed variable characterizing two different components as it frequently happens for PCA. It follows that for CDPCA it is not necessary to use a rotation method to improve the interpretability of results.

Components of CDPCA are generally not orthogonal. However, for the authors this has not to be considered a problem because if two components of CDPCA are highly correlated this means that only one of the two is needed, and simply it is necessary to reduce the number of clusters for variables in order to allow the merge of the two components.

In the case all observed variables are highly correlated CDPCA should find one component only and therefore there is no classification for variables.

In the case the classification of variables only is required (i.e., *P*=*I*) the CDPCA can be seen as an alternative method to the Sparse Principal Component Analysis (Zou, Hastie and Tibshirani, 2006) that uses the *lasso* (*elastic net*) to produce modified principal components with sparse loadings. However, differently from this last technique, CDPCA produces a more simplified interpretation of the components by specifying a partition of variables so that, as mention above, these can explain a single component only.

CDPCA maximizes the between cluster deviance of the reduced space, but unconditionally to the within variance. Thus, it would be interesting to find the optimal solution of CDPCA for a fixed within cluster deviance or as a convex combination of the between and within cluster deviances as in Vichi, Rocci and Kiers (2007); however, this will be subject of another paper.

References

Cattell, R.B., 1966. The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-276.

DeSarbo, W. S., Jedidi, K., Cool, K. And Schendel, D., 1990. Simultaneous Multidimensional Unfolding and Cluster Analysis: An Investigation of Strategic Groups, *Marketing Letters,* **2** 129-146.

DeSarbo, W., S., Howard, D., J., and Jedidi, K., 1991. MULTICLUS: A New Method for Simultaneous Performing Multidimensional Scaling and Clustering, *Psychometrika*, **56**, 121-136

De Soete G., and Carroll, J. D., 1994. K-means Clustering in a Low-dimensional Euclidean Space, in: E. Diday et al. (Eds), *New Approaches in Classification and Data Analysis*, Springer, Heidelberg, 212-219.

De Soete G., and Heiser, W.J., 1993. A latent class unfolding model for analyzing single stimulus preference ratings, *Psychometrika,* **58** 545-565.

Gabriel, K. R., 1971. The biplot graphic display of matrices with application to principal component analysis. *Biometrika,* **58***,* 453-67.

Heiser, W.J., 1993. Clustering in low-dimensional space, in: O. Opitz, B. Lausen, and R. Klar (Eds.), *Information and classification* Springer-Verlag, Berlin, 162-173.

Heiser, W.J., & Groenen, P.J.F., 1997. Cluster differences scaling with a within-clusters loss component and a fuzzy successive approximation strategy to avoid local minima, *Psychometrika*, *62*, 63-83.

Kaiser, H.F., 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrika,* **23** 187-200.

Milligan, G. W. & Cooper, M., 1985. An Extimation of Procedures for Determining the Number of Clusters in a Data Set, *Psychometrika*, 50, 159-179.

Vichi M. and Kiers H. A. L., 2001. Factorial *k*-means analysis for two way data (2001), *Computational Statistics and Data Analysis,* 37, 49-64.

Vichi M., 2000. Double *k*-means Clustering for simultaneous classification of Objects and Variables. In Borra et al. (eds): *Advances in Classification and Data Analysis*. (Eds), *series Studies in Classification, Data Analysis, and Knowledge Organization*, Springer-Verlag, Heidelberg, ISBN 3-540-41488-6, 43-52..

Vichi M., 2002. Discrete and Continuous models for two way data (2002). *Classification Automation and new Media,* in: Studies in Classification, Data Analysis and Knowledge Organization, Springer, Heidelberg, Eds W. Gaul, G. Ritter, 139-147.

Vichi M., Rocci R, H.A.L. Kiers, 2007. Simultaneous Component and Clustering models for three-way data: Within and Between Approaches. *Journal of Classification*, vol. 24 n1, 71-98.

Vigneau, E. and Qannari E. M., 2004. Clustering of variables around latent component - application to sensory analysis., *Communications in Statistics, Simulation and Computation,* 32(4), 1131-1150

Widaman, K.F., 2007. Common factors versus components: Principals and principles, errors and misconceptions. In: R. Cudeck & R.C. MacCallum (Eds). *Factor analysis at 100. Historical developments and future directions,* Mahwah: Lawrence Erlbaum Associates, 177-203.

Zou, H., Hastie, T. and Tibshirani R., 2006. Sparse principal component analysis, *Journal of Computational and Graphical Statistics*, 15(2), 262-286.