# Factorial PD-Clustering

Cristina Tortora, Francesco Palumbo, Mireille Gettler Summa

# Factorial PD-Clustering

Cristina Tortora[1,2], Francesco Palumbo[1], Mireille Gettler Summa[2]
[1]Università di Napoli Federico II
[2]Université Paris Dauphine CEREMADE CNRS

## Abstract

Factorial clustering methods have been developed in recent years thanks to the improving of computational power. These methods perform a linear transformation of data and a clustering on transformed data optimizing a common criterion. Factorial PD-clustering is based on Probabilistic Distance clustering (PD-clustering). PD-clustering is an iterative, distribution free, probabilistic, clustering method. Factorial PD-clustering make a linear transformation of original variables into a reduced number of orthogonal ones using a common criterion with PD-Clustering. It is demonstrated that Tucker 3 decomposition allows to obtain this transformation. Factorial PD-clustering makes alternatively a Tucker 3 decomposition and a PD-clustering on transformed data until convergence. This method could significantly improve the algorithm performance and allows to work with large dataset, to improve the stability and the robustness of the method.

## 1 Introduction

In a wide definition Cluster Analysis is a multivariate analysis technique that seeks to organize information about variables in order to discover homogeneous groups, or "clusters" into data. The presence of groups in data depends on the association structure over the data. Clustering algorithms aim at finding homogeneous groups with respect to their association structure among variables. Proximity measures or distances can be properly used to separate homogeneous groups. A measure of the homogeneity of a group is the variance. Dealing with numerical linearly independent variables, clustering problem consists in minimizing the sum of the squared Euclidean distances within classes: within groups deviance.

"*The term cluster analysis refers to an entire process where clustering maybe only a step*" [Gordon, 1999]. According to Gordon's definition cluster analysis can be sketched in three main stages:

- transformation of data into a similarity/dissimilarity matrix;
- clustering;
- validation.

Transformation of data into similarity/dissimilarity measures depends on data type. On this transformed matrix a clustering method can be applied. Clustering methods can be divided into four main types: hierarchical, non hierarchical, probabilistic and mixture model. Non hierarchical clustering methods are considered in this paper. Among them the most known and used method is k-means. It is an iterative method that starts with a random initial partition of units and keeps reassigning the units into clusters based on the squared distances between the unit and the cluster

centers until the convergence is reached. Interested readers can refer to [Gordon, 1999]. Major k-means issues are that clusters can be sensitive to the choice of the initial centers and that the algorithm could converge to local minima.

A well known problem of non hierarchical method is the choice of the number of clusters, this problem will be not dealt in this paper where the number of clusters is assumed as a priori known. Non hierarchical clustering methods performance can be strongly affected by the dimensionality. Let us consider an $n \times J$ data matrix $X$, with $n$ number of units and $J$ number of variables. Non hierarchical methods easily deal with large $n$, however they can fail when $J$ becomes large or very large and when the variables are correlated. They do not converge or they converge into a different solution at each iteration. To cope with these issues, French school of *Analyse de Données* proposed the *tandem analysis* algorithm to improve the overall quality of clustering, Arbie and Hubert in 1996 [Arabie et al., 1996] formalized the method that consists in two phases: variables transformation through a factorial method and clustering method on transformed variables. The choice of the factorial method is an important and tricky phase because it will affect the results. Principal factorial methods are:

- quantitative data;
  - Principal Component Analysis (PCA);
- binary data;
  - Principal Component Analysis (PCA);
  - Multiple Correspondence Analysis (MCA);
- nominal data;
  - Multiple Correspondence Analysis (MCA);

The second phase of the *tandem analysis* consists in applying clustering methods.
Tandem analysis exploits the factor analysis capabilities that consist in obtaining a reduced number of uncorrelated variables which are linear transformation of original variables. This method gives more stability to the results and makes the procedure faster. Although tandem analysis minimizes two different functions that can be in contrast and the first factorial step can in part obscure or mask the clustering structure.
This technique has the advantage that it works on a reduced number of variables that are orthogonal and ordered with respect to the borrowed information.
To cope with these issues Vichi and Kiers [Vichi and Kiers, 2001] proposed Factorial k-means analysis for two-way data. The aim of this method is to identify the best partition of the objects and to find a subset of factors that best describe the classification according to the least squares criterion. An alternating least squares algorithm (ALS) based on two steps solves this problem. The advantage of Factorial k-means is that the two steps optimize a single objective function.
Although the k-means algorithm itself, and as consequence the tandem analysis and Factorial k-means, is based on the arithmetic mean that gives rise to unsatisfactory solutions when clusters have not spherical shape.

Probabilistic clustering methods may allows us to obtain better results under this condition because they assign a statistical unit to a cluster according to a probability function that can be independently defined with respect to the arithmetic mean.
Probabilistic D-clustering [Ben-Israel and Iyigun, 2008] is an iterative, distribution free, probabilistic, clustering method. PD-clustering assigns units to cluster according to their probability of belonging to a cluster, under the constraint that the ratio between the probability and the distance

of each point to any cluster center is a constant.
When the number of variables is large and variables are correlated, PD-Clustering becomes very unstable and the correlation between variables can hide the real number of clusters. A linear transformation of original variables into a reduced number of orthogonal ones using common criteria with PD-Clustering can significantly improve the algorithm performance. The objective of this paper is to introduce an improved version of Probabilistic D-clustering called Factorial PD-Clustering.

The paper has the following structure: section 1: short presentation of factorial clustering methods; section 2: detailed presentation of Probabilistic D-clustering method; section 3: presentation of our proposed Factorial PD-clustering method; section 4: application of Factorial PD-clustering on a real dataset and comparison with Factorial k-means.

# 2 Related methods

Factorial clustering methods have been proposed to cope with large dataset and to obtain stable clusters. These methods perform a factorial step and a clustering step iteratively, optimizing a common criterion.
Iterative methods are used when the direct solution is unfeasible or cannot be computed. These methods attempt to solve a problem by finding successive approximations to the solution, starting from an initial guess.
Two-step clustering methods are the combination of a factorial method with an iterative clustering procedure. Two quantities have to be computed: linear transformation of the dataset and clustering partition; they cannot be computed at the same time. Two strategies are faceable: two-step tandem analysis or iterative two-step methods. Two-step tandem analysis minimizes two different functions that can be in contrast and the first factorial step can in part obscure or mask the clustering structure. This issue can be overcame. De Soete and Carroll in 1994 proposed an alternative method [De Soete and Carroll, 1994]. This method is an alternative k-means procedure, after a clustering phase, it represents centroids in a lower dimensional space chosen such that the distances between centroids and points belonging to the cluster are minimized. After all the points are projected in this low-dimensional space obtaining a low-dimensional representation of points and clusters. This method can fail in finding the real clustering structure when the data have much variance in directions orthogonal to the one capturing the interesting clustering. Iterative two-step methods overcome these issues. Among two-step iterative methods a first contribute was given by Vichi and Kiers [Vichi and Kiers, 2001] with Factorial k-means analysis for two-way data. Related methods have been developed to cope with categorical and binary data. Some relevant methods are: multiple correspondence analysis for identifying heterogeneous subgroups of respondents [Hwang et al., 2006] and Iterative Factorial Clustering of Binary Data [Iodice D'Enza and Palumbo, 2010].

## 2.1 Factorial k-means

The aim of this method is to identify the best partition of the objects and to find a subset of factors that best describe the classification according to the least squares criterion. An alternating least squares two-step algorithm solves this problem.
Defined with:

- $X$ data matrix with $n$ units and $J$ variables;
- $A$ columnwise $J \times R$ orthonormal matrix, the elements are a linear combination of observed variables, $R$ is the number of factor to be used;

- $U$ $n \times K$ of general element $u_{ij}$ binary membership to each cluster matrix, with $K$ number of clusters;
- $E$ error components matrix;
- $\bar{Y}$ centroids matrix.

Factorial k-means model is defined as follow:

$$XAA' = U\bar{Y}A' + E \tag{1}$$

The model is an orthogonal projection of units on a subspace spanned by the columns of the columnwise orthonormal matrix $A$. The objective is to minimize the squared error:

$$\begin{aligned} &\min ||XA - U\bar{Y}||^2 \\ &\quad \text{subject to} \qquad A'A = I \\ &\qquad\qquad\qquad\quad U \text{ is a binary matrix and } \sum_{j=1}^{J} u_{ij} = 1 \forall i \end{aligned} \tag{2}$$

Centroids matrix $\bar{Y}$ can be expressed as: $\bar{Y} = (U'U)^{-1}U'XA$ so the expression minimized in the (2) becomes: $||XA - U(U'U)^{-1}U'XA||^2$. This quantity can be decomposed in two parts and the function to be minimized becomes:

$$\min[tr(A'X'XA) - tr(A'X'U(U'U)^{-1}U'XA)] \tag{3}$$

The first component of the (3) is the total deviance of $XA$, the second is the between classes deviance. An ALS algorithm can be used to solve the (3).
The advantage of Factorial k-means is that the two steps optimize a single objective function. Tandem analysis instead minimizes two different functions that can be in contrast and the first factorial step can in part obscure or mask the cluster structure.
Another advantage is that clustering method is applied on factorial axes. If the number of chosen factors is lower than the number of variables the method is external robust.

# 3 Probabilistic D-Clustering

Probabilistic D-clustering is a non hierarchical algorithm that assigns units to clusters according to their belonging probability to the cluster. Given some random centers, the probability of any point to belong to each class is assumed inversely proportional to the distance from the centers of clusters. Given an $X$ data matrix with $n$ units and $J$ variables, given $K$ clusters that are assumed not empty, Probabilistic D-Clustering is based on two quantities: the distance of each data point $x_i$ from the $K$ cluster centers $c_k$, $d(x_i, c_k)$, and the probabilities of each point to belong to a cluster, $p(x_i, c_k)$ with $k = 1, \ldots, K$ and $i = 1, \ldots, n$. The relation between them is the basic assumption of the method. Let us consider the general term $x_{ij}$ of $X$ and a center matrix $C$, of elements $c_{kj}$ with $k = 1, \ldots, K$, $i = 1, \ldots, n$ and $j = 1, \ldots, J$, their distance can be computed according to different criteria, the squared norm is one of the most commonly used. The generic distance $d(x_i, c_k)$ represents the distance of the generic point $i$ to the generic center $k$. The probability $p(x_i, c_k)$ of each point to belong to a cluster can be computed according to this assumption: the product between the distances and the probabilities is a constant depending on $x$: $D(x)$.
For short we use $p_k(x_i) = p(x_i, c_k)$ and $d_k(x_i) = d(x_i, c_k)$; PD-clustering basic assumption is expressed as:

$$p_k(x)d_k(x) = D(x) \tag{4}$$

for a given value of $x$ and for all $k = 1, \ldots, K$.

At the decreasing of the point closeness form the cluster center the belonging probability of the point to the cluster decreases. The constant depends only on the point and does not depend on the cluster $k$.

Starting from the (4) it is possible to compute $p_k(x)$:

$$p_t(x)d_t(x) = p_k(x)d_k(x);\ p_t(x) = \frac{p_k(x)d_k(x)}{d_t(x)};\ \forall t = 1, \ldots, K \tag{5}$$

The term $p(x)$ is a probability so, under the constraint $\sum_{t=1}^{K} p_t(x) = 1$, the sum over $t$ of (5) becomes:

$$p_k(x) \sum_{t=1}^{K} \left( \frac{d_k(x)}{d_t(x)} \right) = 1$$

$$p_k(x) = \frac{1}{\sum_{t=1}^{K} \left( \frac{d_k(x)}{d_t(x)} \right)} = \frac{\prod_{j \neq j} d_j(x)}{\sum_{t=1}^{K} \prod_{j \neq t} d_j(x)}, k = 1, \ldots K. \tag{6}$$

Starting from the (4) and using (6) it is possible to define the value of the constant $D(x)$:

$$D(x) = p_k(x)d_k(x), k = 1, \ldots K$$

$$D(x) = \frac{\prod_{k=1}^{K} d_k(x)}{\sum_{t=1}^{K} \prod_{j \neq t} d_j(x)}. \tag{7}$$

The quantity $D(x)$, also called *Joint Distance Function* (JDF), is a measure of the closeness of $x$ from all clusters centers. The JDF measures the classificability of the point $x$ with respect to the centers $c_k$ with $k = 1, \ldots, K$. If it is equal to zero, the point coincides with one of the clusters centers, in this case the point belongs to the class with probability 1. If all the distances between the point and the centers of the classes are equal to $d$, $D(X) = d/k$ and all the belonging probabilities to each class are equal: $p(x) = 1/K$. Smaller is the value of the JDF higher is the probability of the point to belong to one cluster.

The whole clustering problem consists in the identification of the centers that minimizes the JDF. Without loss of generality the PD-Clustering optimality criterium can be demonstrated according to $k = 2$.

$$\begin{aligned} &\min d_1(x)p_1^2 + d_2(x)p_2^2 \\ &\text{subject to} \qquad p_1 + p_2 = 1 \\ &\qquad\qquad\quad\ \ p_1, p_2 \geq 0 \end{aligned} \tag{8}$$

The probabilities are squared because it is a smoothed version of the original function. The Lagrangian of this problem is:

$$L(p_1, p_2, \lambda) = d_1(x)p_1^2 + d_2(x)p_2^2 - \lambda(p_1 + p_2 - 1) \tag{9}$$

Setting to zero the partial derivates with respect to $p_1$ and $p_2$, substituting the probabilities (6) and considering the principle $p_1(x)d_1(x) = p_2(x)d_2(x)$ we obtain the optimal value of the Lagrangian.

$$L(p_1, p_2, \lambda) = \frac{d_1(x)d_2(x)}{d_1(x) + d_2(x)}. \tag{10}$$

This value coincides with the JDF, the matrix of centers that minimizes this principle minimizes the JDF too. Substituting the generic value $d_k(x)$ with $\|x - c_k\|$, we can find the equations of the centers that minimize the JDF (and maximize the probability of each point to belong to only one cluster).

$$c_k = \sum_{i=1,\dots,N} \left( \frac{u_k(x_i)}{\sum_{j=1,\dots,N} u_k(x_j)} \right) x_i, \tag{11}$$

where

$$u_k(x_i) = \frac{p_k(x_i)^2}{d_k(x_i)}. \tag{12}$$

As showed before, the value of JDF at all centers $k$ is equal to zero and it is necessary positive elsewhere. So the centers are the global minimizer of the JDF. There exist other stationary points, because the function is not convex neither quasi-convex, but they are saddle points.

There are alternative ways for modeling the relation between probabilities and distances, for example the probabilities can decay exponentially as distances increase. In this case the probabilities $p_k(x)$ and the distances $d_k(x)$ are related by:

$$p_k(x) e^{d_k(x)} = E(x), \tag{13}$$

where $E(x)$ is a constant depending on $x$.
Many results of the previous case can be extended to this case by replacing the distance $d_k(x)$ with $e^{d_k(x)}$.
Interested readers are referred to [Ben-Israel and Iyigun, 2008]

### 3.1 Choice of the distance

The optimization problem presented in (8) is the original version proposed by Ben-Israel and Iyigun. Notice that in the optimization problem the probabilities $p_k$ are considered in squared form. The authors affirm that it is possible to consider $d_k$ as well $d_k^2$. Both choices have some advantages and drawbacks. Squared distances offer analytical advantages due to linear derivates. Using simple distances endures robustness results and the optimization problem can be reconducted to a Fermat-Weber location problem. The Fermat-Weber location problem aims at finding a point that minimizes the sum of the Euclidean distances from m given points. This problem can be solved with the Weiszfeld method [Weiszfeld, 1937]. Convergence of this method was established by modifying the gradient so that it is always defined [Khun, 1973]. The modification is not carried out in practice. The global solution is guaranteed only in case of one cluster. Dealing with more then one cluster, in practice, the method converges only for a limited number of centers depending on the data.

### 3.2 Probabilistic D-clustering algorithm

The solution of Probabilistic D-clustering problem can be obtained through an iterative algorithm. Given a data set $X$ and a set of center $C$:

step 0 random initialization of center matrix;

step 1 distances $d_k(x)$ for all $x \in X$ with $k = 1, \dots K$;

step 2 update the center matrix $C^*$;

step 3 if $\sum_{k=1}^{K} \left\| c_k^* - c_k \right\|_2 < \varepsilon$ stop, else return to 1.

Where $c_k$ is the generic center and $d_k(x)$ is defined in (4).
Cluster centers and JDF change at each iteration, the objective function decreases and the algorithm converges.
Points are assigned to the $k^{th}$ cluster according to the higher probability that are computed a posteriori according to (6)

# 4 Factorial PD-Clustering

When the number of variables is large and variables are correlated PD-Clustering becomes very unstable and the correlation between variables can hide the real number of clusters. A linear transformation of original variables into a reduced number of orthogonal ones can significantly improve the algorithm performance. Combination of PD-Clustering and variables linear transformation implies a common criterion. This section shows how the Tucker 3 method [Kroonenberg, 2008] can be properly adopted for the transformation in the Factorial PD-Clustering.

Factorial PD-clustering is an iterative procedure that consists of two main step that are:

- Linear transformation of original data;
- PD-Clustering on transformed data.

Center matrix $C$ of elements $c_{kj}$ with $k = 1,\ldots,K$ and $j = 1,\ldots,J$ is pseudorandomly defined before starting the algorithm. So that probabilities and distance matrices can be computed. The distance between a generic observation $x_i$ and a generic center $c_k$ is the absolute difference between each coordinate of the point and each coordinate of the center:

$$d(x_i, c_k) = g_{ijk} = |x_{ij} - c_{kj}| \tag{14}$$

with $i = 1,\ldots,n$, $k = 1,\ldots,K$, for a given value of $j$.
The matrix $G$ of elements $g_{ijk}$ is a 3-way matrix $n \times J \times K$ where $n$ is the number of units, $J$ the number of variables and $K$ the clusters.
The aim is to transform variables in new ones obtained as liner transformation of original variables minimizing JDF.
The probabilities $p_{ik}$ are arranged in $K$ diagonal $n \times n$ matrices where the general term is given by: (6). The problem is to transform original variable minimizing:

$$\min \left( \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{k=1}^{K} (x_{ij} - c_{kj})^2 p_{ik}^2 \right) \tag{15}$$

In the section 4.2 we demonstrate that the minimization problem in (15) corresponds to the Tucker3 decomposition of distance matrix $g$. It is demonstrated that this solution minimizes JDF too. Consequently it is demonstrated that applying a Tucker 3 decomposition on the distance matrix $d$ it is obtained the space that better represents the data according to PD-Clustering criteria.

For any $c_k$ with $k = 1,\ldots,K$, it is defined a $G_k$ $n \times J$ distances matrices. Tucker3 method decomposes the matrix $G$ in three components, one for each mode, in a full core array $\Lambda$ and in an error term $E$.

$$g_{ijk} = \sum_{r=1}^{R} \sum_{q=1}^{Q} \sum_{s=1}^{S} \lambda_{rqs} (u_{ir} b_{jq} v_{ks}) + e_{ijk}$$

with $i = 1,\ldots,n$, $j = 1,\ldots,J$, $k = 1,\ldots,K$.
Defining with: $R$ number of components of $U$, $Q$ number of components of $B$ and $S$ number of

components of $V$.
The units coordinates into the space of variables obtained by Tucker3 decomposition are obtained by the following expression:

$$x^{\star}_{iq} = \sum_{j=1}^{J} x_{ij} b_{jq}. \tag{16}$$

On the projection of units on these first $R$ factorial axes PD-clustering algorithm can be applied. Using PD-clustering a new center matrix $C$ is obtained and the distance matrix is updated. Basing on the new distances, the algorithm is reiterated. The whole algorithm is reiterated until convergence.

As in all factorial methods, factorial axes in Tucker 3 model are sorted according to explained variability. The first factorial axes explain the great part of the variability, latest factors represent outliers or ground noise. For this reason the choice of a number of factors lower than the number of variables makes the method externally robust. Consequently in order to reduce the dimensionality and to make the result robust a number of factor $R \leq J$ can be chosen.

### 4.1 Factorial PD-clustering algorithm

Given a dataset $X$ and a center matrix $C$, the iterative algorithm can be summarized as follow.

step 0 PD-cluster algorithm: compute center matrix $C$;

step 1 distances $G_k$ for all $x \in X$ with $k = 1, \dots K$;

step 2 Tucker3 decomposition and computation of $X^{\star}$ matrix;

step 3 PD-clustering of reduced data and update of the center matrix;

step 4 if $\sum_{k=1}^{K} \left\| c_k^{J*} - c_k^{J} \right\| < \varepsilon$ stop, else return to 1.

### 4.2 Factorial PD-clustering method

In this section it is demonstrated that the space that better represents the data according to PD-Clustering criteria is the one obtained with Tucker3 decomposition on distance matrix.
The probability matrix can be written as a diagonal matrix $P = diag(vec(p))$ of generic element $p_h$ with $h = 1, \dots, n \times K$. Each element of this matrix represents the probability $p_{ik}$ that the point $x_i$ belongs to the cluster $k$, with $i = 1, \dots, n$ and $k = 1, \dots, K$. As we have seen before, the objective is to minimize the JDF that is equivalent to minimize the product between $G$ and $P$ that is still a 3-way matrix. Factorial PD-Clustering is a soft modeling method that aims at finding a unique solution that at the same time minimizes JDF and that makes a linear transformation of data. We can prove that using Tucker 3 transformation we obtain the space that minimizes the JDF, consequently we can obtain a unique solution obtained optimizing the same criteria in the two steps:

$$\min \left( \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{k=1}^{K} (x_{ij} - c_{kj})^2 p_{ik}^2 \right).$$

The unknown quantities of this formula are $c_{kj}$ and $p_{ik}$.
*THEOREM 1*: To maximize $-[\sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{k=1}^{K} (x_{ij} - c_{kj})^2 p_{ik}^2]$ is equivalent to the decompose $(x_{ij} - c_{kj}) = \sum_{r=1}^{R} \sum_{q=1}^{Q} \sum_{s=1}^{S} \lambda_{rqs} (u_{ir} b_{jq} v_{ks}) + e_{ijk}$ (Tucker3 decomposition).

*PROOF*: Objective function is:

$$\max\left[-\left(\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ij}-c_{kj})^2 p_{ik}^2\right)\right]$$

$$\text{sub constraint} \qquad \sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}^2 \leq n$$

Replacing $(x_{ij}-c_{kj})$ with $g_{ijk}$ (14) we obtain:

$$\max\left(-\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K} g_{ijk}^2 p_{ik}^2\right)$$

$$\text{sub constraint} \qquad \sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}^2 \leq n$$

The Lagrangian is:

$$L = -\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K} g_{ijk}^2 p_{ik}^2 + \lambda\left(\sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}^2 - n\right)$$

where $\lambda \geq 0$ is the Lagrangian multiplier. In order to maximize the Lagrangian we have to compute the first derivate.

$$\frac{\delta L}{\delta p} = 0$$

$$-2\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K} g_{ijk}^2 p_{ik} + 2\lambda\sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik} = 0$$

$$\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K} g_{ijk}^2 p_{ik} = \lambda\sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik} \tag{17}$$

It can be demonstrated that the second derivate is not positive. The value of $\lambda$ that optimize the (17) can be find trough the research of the eigenvalues of the matrix $G$.

$$g_{ijk} = \sum_{r=1}^{R}\sum_{q=1}^{Q}\sum_{s=1}^{S}\lambda_{rqs}(u_{ir}b_{jq}v_{ks}) + e_{ijk}$$

with $i = 1,\ldots,n,\ j = 1,\ldots,J,\ k = 1,\ldots,K$.
Defining with: $R$ number of components of $U$, $Q$ number of components of $B$ and $S$ number of components of $V$.
In matrix notation:

$$G = U\Lambda(V' \otimes B') + E \tag{18}$$

That is the Tucker3 decomposition of matrix G.

■

PD-clusering objective function is: to maximize $-[\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ij}-c_{kj})^2 p_{ik}^2]$. It can be said that:
*LEMMA 1*:The space that better represents the data according to PD-Clustering criteria is the one obtained with Tucker3 decomposition on distance matrix $g$.

Starting from the (16) the proof is the same as proof of Theorem 1.

# 5 Application on a real dataset

Factorial PD-clustering has been applied on the dataset used in [Vichi and Kiers, 2001]. The dataset *latest short-term indicators and economic performance indicators*[1] contains 6 macroeconomic variables measured on 20 countries members of the OECD. Variables are the six main

[1]OECD, Paris 1999

economic indicators: Gross Domestic Product (GDP), Leading Indicator (LI), Unemployment Rate (UR), Interest Rate (IR), Trade Balance (TB), Net National Savings (NNS). Table 1 contains the dataset.

The first step of Factorial PD-clustering is the choice of the number of clusters $K$ that has been fixed equal to 3.

The choice of the number of factors is a ticklish well known issue; however it will be not dealt

| Country | Label | GDP | LI | UR | IR | TB | NNS |
|---|---|---|---|---|---|---|---|
| Australia | A-lia | 4.80 | 8.40 | 8.10 | 5.32 | 0.70 | 4.70 |
| Canada | Can | 3.20 | 2.50 | 8.40 | 5.02 | 1.60 | 5.20 |
| Finland | Fin | 3.90 | -1.00 | 11.80 | 3.60 | 8.80 | 7.70 |
| France | Fra | 2.30 | 0.70 | 11.70 | 3.69 | 3.90 | 7.30 |
| Spain | Spa | 3.60 | 2.50 | 19.00 | 4.83 | 1.20 | 9.60 |
| Sweden | Swe | 4.10 | 1.10 | 8.90 | 4.20 | 7.00 | 4.00 |
| United States | USA | 4.10 | 1.40 | 4.50 | 5.59 | -1.40 | 7.00 |
| Netherlands | Net | 2.90 | 1.60 | 4.20 | 3.69 | 7.00 | 15.80 |
| Greece | Gre | 3.20 | 0.60 | 10.30 | 11.70 | -8.30 | 8.00 |
| Mexico | Mex | 2.30 | 5.60 | 3.20 | 20.99 | 0.00 | 12.70 |
| Portugal | Por | 2.80 | -7.50 | 4.90 | 4.84 | -8.70 | 14.00 |
| Austria | A-tria | 1.10 | 0.60 | 4.70 | 3.84 | -0.60 | 9.40 |
| Belgium | Bel | 1.40 | -0.10 | 9.60 | 3.64 | 4.50 | 12.40 |
| Denmark | Den | 1.00 | 1.50 | 5.30 | 4.08 | 3.30 | 5.00 |
| Germany | Ger | 0.80 | -2.00 | 9.50 | 3.74 | 1.50 | 7.70 |
| Italy | Ita | 0.90 | -0.40 | 12.30 | 6.08 | 4.30 | 8.20 |
| Japan | Jap | 0.10 | 5.40 | 4.20 | 0.74 | 1.20 | 15.10 |
| Norway | Nor | 1.40 | 0.90 | 3.30 | 4.47 | 7.10 | 15.10 |
| Switzerland | Swi | 1.10 | 2.10 | 3.80 | 1.84 | 4.40 | 13.20 |
| United Kingdom | UK | 1.20 | 4.90 | 6.40 | 7.70 | -0.50 | 4.80 |

Table 1: Six macroeconomic performance indicators of twenty OECD countries (percentage change from the previous year, September 1999)

in this context. In this case the number of factors that have been chosen are: 4 factors for the variables, 4 factors for the units and 2 factors for the clusters. The factors correspond to the values of $R$, $Q$ and $S$ respectively in the (18).

The factors used in the analysis explain the 88% of the variability.

The method has been iterated 50 times, to test the results stability. The JDF index has been measured at each iteration, fig. 6 represents the JDF value. In the best case the value of JDF is 15.12, it occurs in 4% of cases. The modal value 16.19 occurs in 35% of cases.

Table 3 displays Factorial PD-clustering iterations summary statistic.

The partition of units in clusters is:

cluster 1 Mexico, Austria, Denmark, Japan, Norway, Switzerland, United Kingdom;

cluster 2 Finland, France, Spain, Netherlands, Portugal, Belgium, Germany, Italy;

cluster 3 Australia, Canada, Sweden, United States, Greece.

To describe the separating power of our variables in the following are shortly described results illustrated in figures 2 to 5.

The differences between the medians have been evaluated for each cluster and for each variable, to understand which are the variables that mostly contribute to the class separability. The results can be better understood looking at the box-plots in fig. 2. The variable Net National savings presents the highest difference between the medians. NNS variable separates the second cluster from the

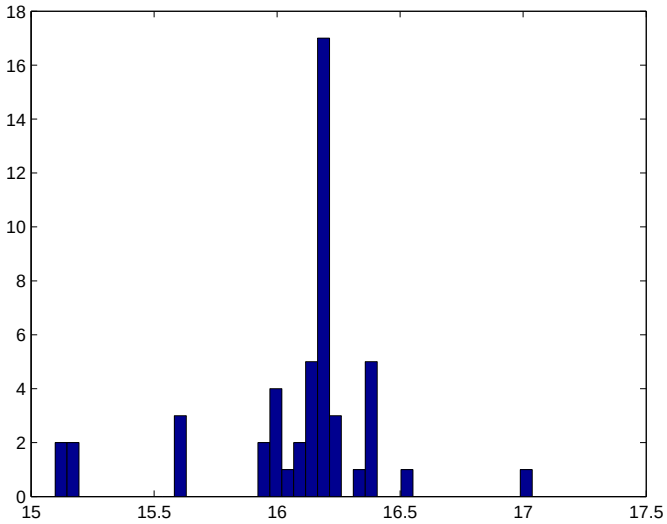


Figure 1: Frequency of JDF on 200 iterations of the Factorial PD-clustering

| JDF | Frequency | mean execution time | mean number of iterations |
|---|---|---|---|
| 15,12 | 4,08% | 0,39 | 8 |
| 15,17 | 4,08% | 0,46 | 12 |
| 15,61 | 6,12% | 0,34 | 6 |
| 15,95 | 4,08% | 0,18 | 5 |
| 15,99 | 8,16% | 0,67 | 17,75 |
| 16,04 | 2,04% | 0,3 | 8 |
| 16,09 | 4,08% | 0,33 | 8 |
| 16,14 | 10,20% | 0,33 | 8,6 |
| 16,19 | 34,69% | 0,45 | 10,23 |
| 16,24 | 6,12% | 0,29 | 6,33 |
| 16,33 | 2,04% | 0,32 | 7 |
| 16,38 | 10,20% | 0,24 | 5,6 |
| 16,53 | 2,04% | 0,38 | 5 |
| 17,01 | 2,04% | 0,12 | 3 |

Table 2: Summary statistics on 50 iterations of Factorial PD-clustering

others, where the values of the variable are lower, in cluster 1 the variable has high variability. The variable Unemployment Rate presents high difference between clusters medians: in cluster 1 the values of this variables are smaller than in the other clusters; in cluster 2 UR presents a small variability. The variable Gross Domestic Product presents high values in cluster 2 and small values in cluster 1. The variable Trade Balance is low in cluster 2. The medians of the variables Interest Rate and Leading Indicator are not significantly different among clusters.

Fig. 3, 4 and 5 represent scatter-plots of units on the variables ordered according to discriminating power.

Scatter-plots in figures 3 to 5 show that:

Mexico, Austria, Denmark, Japan, Norway, Switzerland and United Kingdom have a low variation of Gross Domestic Product but they have a low Unemployment Rate and hight Trade Balance;

Australia, Canada, Sweden, United States and Greece have a high Gross Domestic Product variations, low Unemployment Rate but they have low values of the Net National Savings variable;

Finland, France, Spain, Netherlands, Portugal, Belgium, Germany and Italy have average values

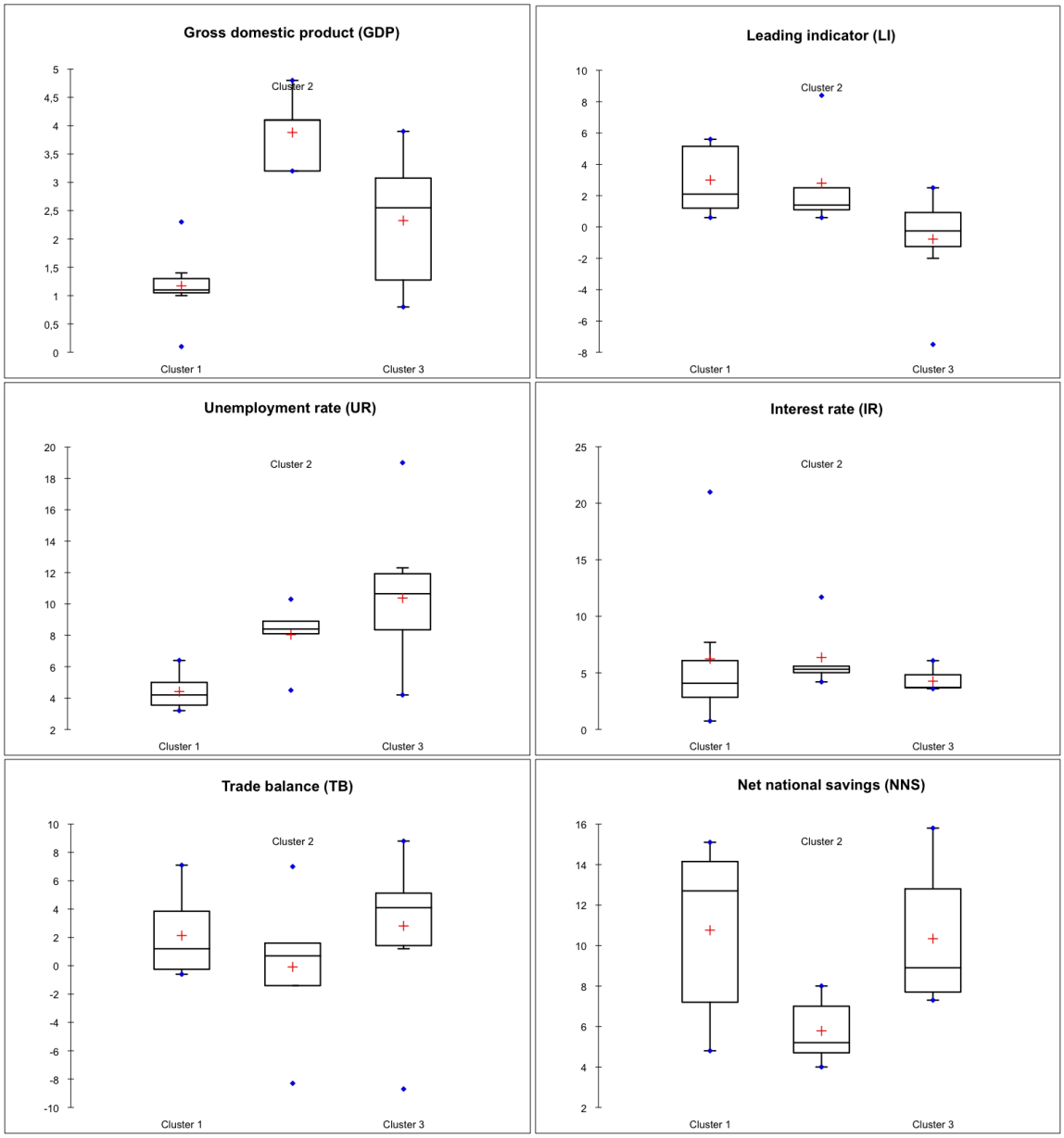


Figure 2: Box-plot of variables for each cluster obtained with Factorial PD-clustering

of Gross Domestic Product and Unemployment Rate and high values of Net National Savings.

The clusters description, presented in the foregoing, corresponds to the most frequent case of the JDF, which is in the interval $[5.6; 5.7]$. Taking into account all iterations, it is worth noticing that three stable groups of countries are always together in the same cluster:

- Australia, Canada, United States;
- Finland, France, Spain;
- Belgium, Germany, Italy.

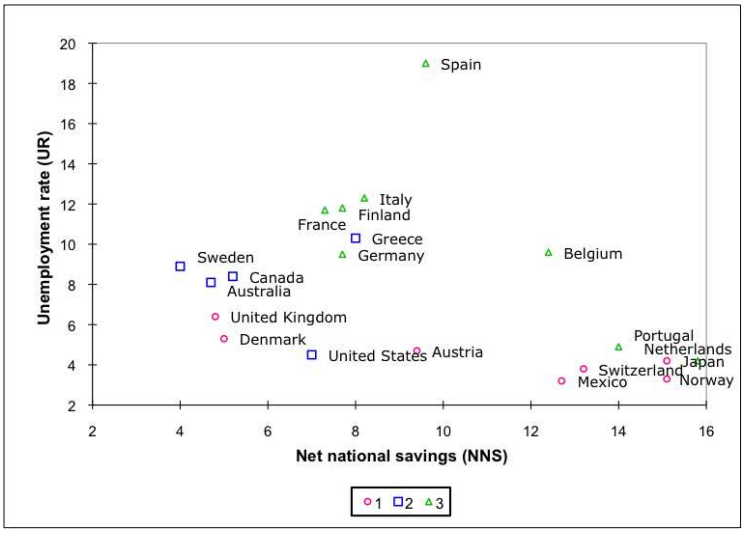


Figure 3: Scatter-plot of units divided in clusters obtained with Factorial PD-clustering

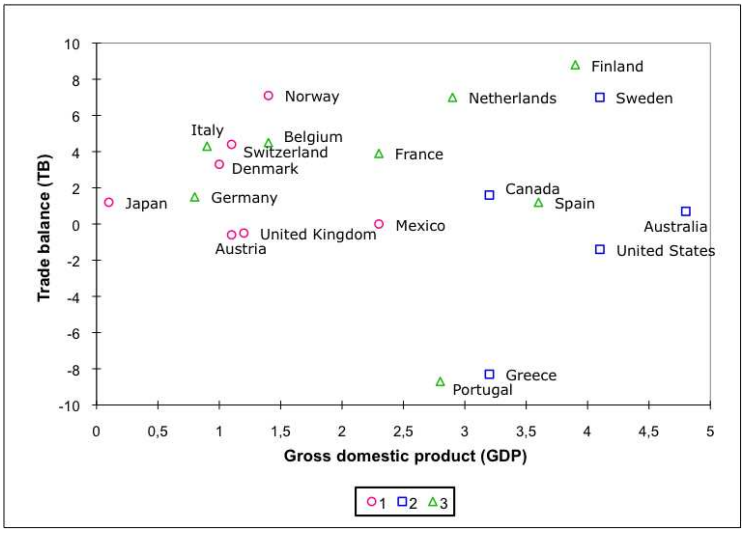


Figure 4: Scatter-plot of units divided in clusters obtained with Factorial PD-clustering

## 5.1 A comparison with Factorial k-means

In order to evaluate Factorial PD-clustering results, the method has been compared with Factorial k-means 2.1. The method have been iterated 200 times.
Using the same scheme adopted for Factorial PD-Clustering result description, in the following Factorial k-means results are shortly described. The method has been applied on the same dataset; the results shown are consistent with those presented in the original Vichi and Kiers' paper.
To identify the most discriminating variables, in this case, the differences between the mean values of the clusters have been evaluated. This because the k-means maximizes the differences between cluster centroids.
In order to easily compare the results the same graphics are represented: box-plots in fig. 7; Scatter-plots indicating the cluster membership are represented in fig.8, 9, 10.
It is important to notice that Factorial k-means results are consistent with the Factorial PD-clustering ones: in particular the method identifies the same stable groups of countries. The most significative

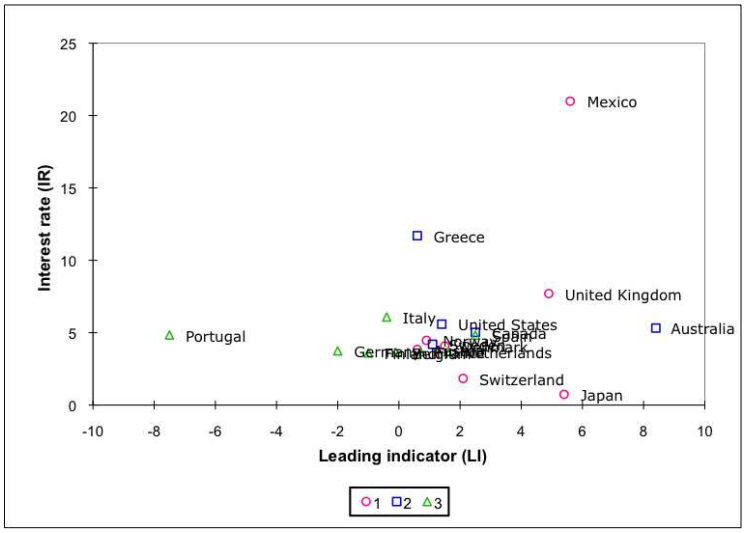


Figure 5: Scatter-plot of units divided in clusters obtained with Factorial PD-clustering

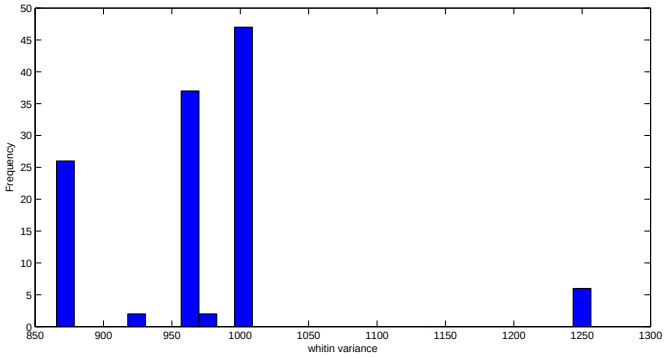


Figure 6: Frequency of JDF on 200 iterations of Factorial k-means

difference is in cluster 2: Factorial k-means identifies a cluster composed by Portugal, Greece and Mexico; Factorial PD-clustering assigns Portugal, Greece and Mexico in three different clusters. Looking at the scatter plot in fig. 8, where statistical units are represented according to the two most separating variables, these three countries appear as the most different from the global mean. It is not surprising that the k-means algorithm separates this three points from the others, because it minimizes the variance within the clusters, and as a consequence it gives maximum importance to the variables Trade Balance and Interest Rate.

To summarize: Factorial k-means has found a cluster of few elements and large variability and two clusters having a larger number of elements (7 and 10 elements, respectively) and a small variability. This partition emphasizes the differences between the variables: Interest Rate, Trade Balance and Net National Savings. Differently Factorial PD-clustering have divided the space in three regions defining three clusters having almost the same variability and almost the same number of elements (7,8 and 5 elements, respectively) The cluster emphasize the differences between the variables: Net National Savings, Unemployment Rate and Gross Domestic Product.

Results obtained with Factorial k-means are different in terms of discriminating variables; the two methods emphasize different aspects of the same phenomenon.

Results presented in this section have been produced with Matlab.
The toolbox *N-way* have been integrated in Matlab code to obtain Tucker3 decomposition. The

| Whitin variance | Frequency | mean execution time | mean number of iterations |
|---|---|---|---|
| 5865,59 | 16,53% | 0,004 | 2,63 |
| 961,92 | 23,14% | 0,004 | 2,64 |
| 970,9 | 3,31% | 0,004 | 2,75 |
| 999,52 | 54,55% | 0,004 | 2,27 |
| 1256,42 | 2,48% | 0,004 | 2 |

Table 3: Summary statistics on 200 iterations of Factorial k-means

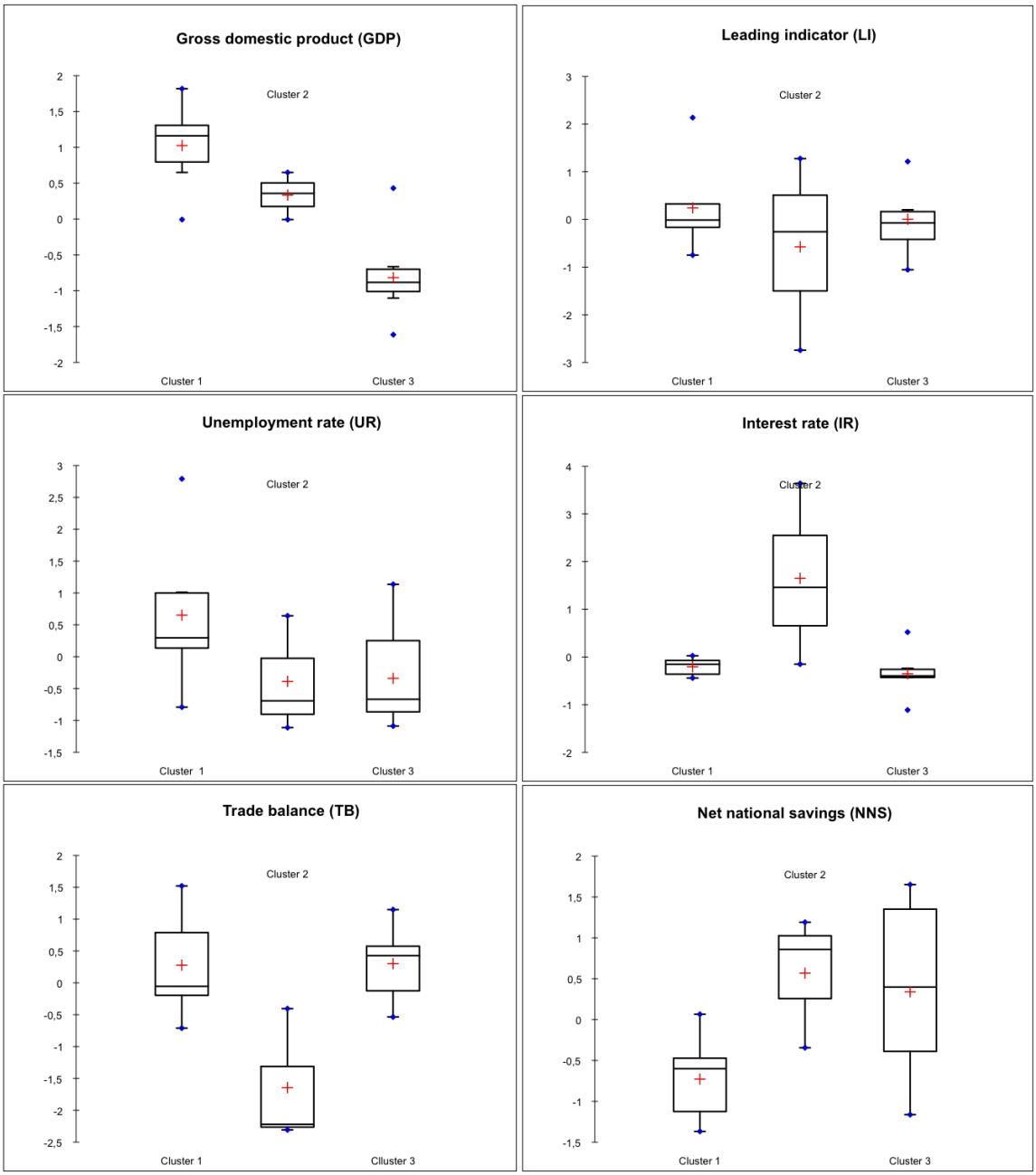


Figure 7: Box-plot of variables for each cluster obtained with Factorial k-means

toolbox is freely available at Matlab Central web site.[2].

[2]http://www.mathworks.com/matlabcentral/

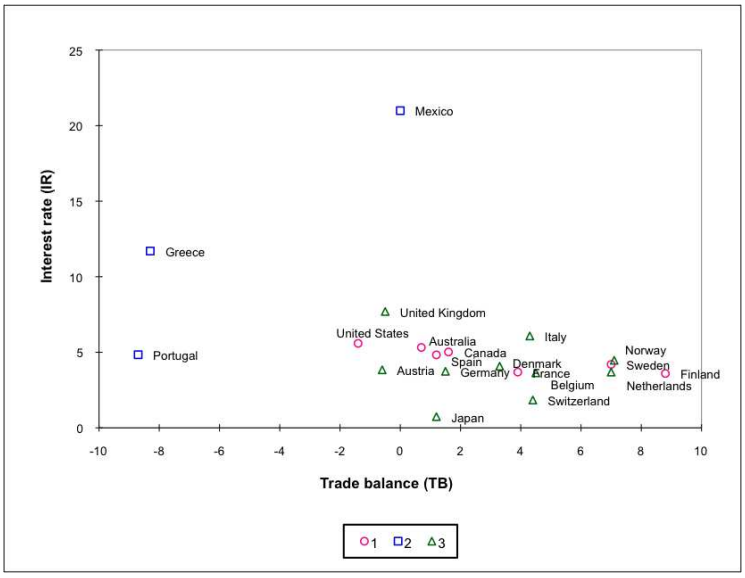


Figure 8: Scatter-plot of units divided in clusters obtained with Factorial k-means

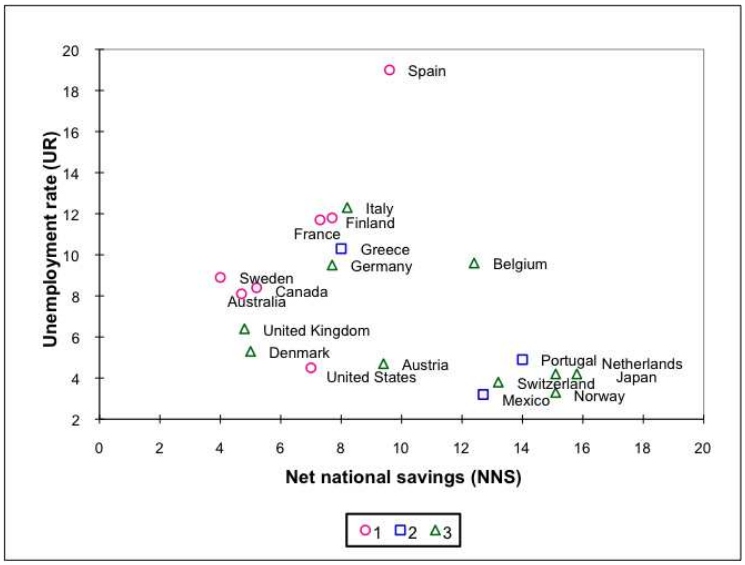


Figure 9: Scatter-plot of units divided in clusters obtained with Factorial k-means

# 6 Conclusions

In this paper a new factorial two-step clustering method has been proposed: Factorial PD-clustering. This method can be inlaid into a new field of clustering technique which has been developed in recent years: iterative clustering methods. Two-step clustering methods was proposed by French school of Analyse de Données in order to cope with some clustering issues. Thanks to computer developing in recent years iterative clustering methods have been introduced. These methods iteratively perform a linear transformation of data and a clustering optimizing a common criterion. Factorial PD-clustering perform a linear transformation of data and Probabilistic D-clustering iteratively. Probabilistic D-clustering is an iterative, distribution free, probabilistic, clustering method. When the number of variables is large and variables are correlated PD-Clustering becomes very unstable and the correlation between variables can hide the real number of clusters. A linear trans-

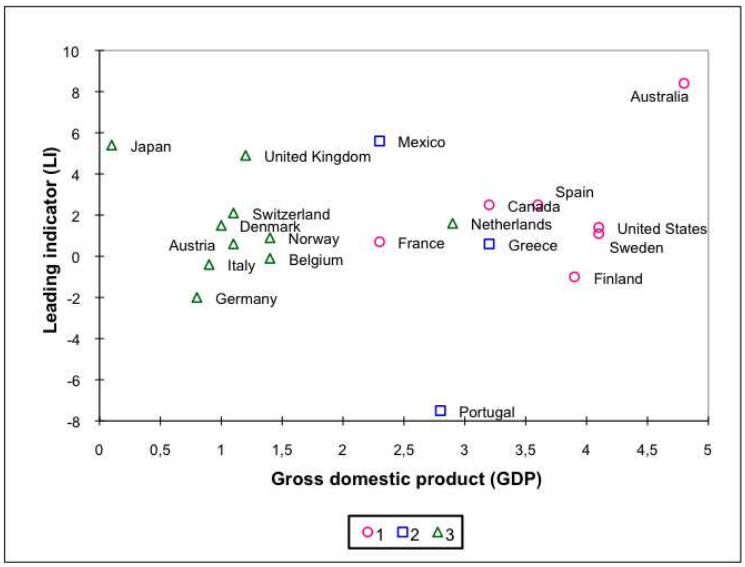


Figure 10: Scatter-plot of units divided in clusters obtained with Factorial k-means

formation of original variables into a reduced number of orthogonal ones using common criteria with PD-Clustering can significantly improve the algorithm performance. Factorial PD-clustering allows to work with large dataset, to improve the stability and the robustness of the method.

# References