Hindawi Publishing Corporation
Advances in Decision Sciences
Volume 2016, Article ID 7546963, 7 pages
http://dx.doi.org/10.1155/2016/7546963

Research Article

Classification of the Entities Represented by Samples from Gaussian Distribution

Amar Rebbouh

Université des Sciences et de la Technologie, Houari Boumediene, BP 32, El Alia, 16111 Bab Ezzouar, Alger, Algeria

Correspondence should be addressed to Amar Rebbouh; arebbouh@usthb.dz

Received 29 February 2016; Accepted 12 May 2016

Academic Editor: Panos Pardalos

This paper aims to cluster entities which are described by a data matrix. Under the assumption of normality of observations contained in each table, each entity is represented by samples from Gaussian distribution, that is, a number of measurements in the data matrix, the sample mean vector, and the sample covariance. We propose a new distance based on Mahalanobis's discriminant score to measure the similarity between objects. The present study is thought to be an important and interesting topic of research not only in the quest for an adequate model of the data representation but also in the choice of the distance index between entities that would allow justifying the homogeneity of any observed classes.

1. Introduction

One of the fundamental problems in automatic classification is the development and validation of similarity indices between the objects to be classified. These indices must adapt to classify objects and allow measuring the adequacy between an object and a class of objects. If the objects to be classified are described by matrices comprising such repeated observations of individuals for the variables that describe them over a finite period of time, we present a new distance based on Mahalanobis's discriminant score to measure the similarity between objects.

Usually, for this type of data, before the classification stage, we proceed to a reduction step. We can summarize each table by a vector, or a hyperrectangle, and we can use factorial techniques to reduce each table.

Therefore, these reduction techniques require assumptions that are difficult to achieve in practice. Indeed, the first type of reduction makes sense only if the mean or another central value summarizes perfectly the observations of each individual i, and this reduction does not take into account the variability of the observations. The hyperrectangles are Cartesian products of intervals. The interval estimated depends on the variability of the observations but does not consider the possible relationship between the variables. This type of reduction requires that the variables must be uncorrelated. Several distances between interval objects have been extended to distances between hyperrectangles and remain a subject of research in automatic classification. These include the distance based on city block distance [1], Hausdorff distance between hyperrectangles, Wasserstein based distance [2], and single adaptive distance [3]. Finally, the third type of reduction leads to new uncorrelated variables but poses significant mathematical problems such as the search for compromise space and the number of observations to be used for the reduction of each entry table (see [4, 5]). If the number of observations of each variable is the same for each object, the input data can be considered as a structure of data matrices (see [6]).

This paper aims to cluster entities which are described by a data matrix. Under the assumption of normality of observations contained in each table, each entity is represented by samples from Gaussian distribution, that is, a number of measurements in the data matrix, the sample mean vector, and the sample covariance. We define a new distance based on Mahalanobis's discriminant score to measure the similarity between objects. We propose an extension of the k-means algorithm to this case. The approach can be extended to cluster objects described by variable subjects with errors of measurements.

In analogy to the classical squared-error criterion and the k-means algorithm, clustering is here proceeding by defining and minimizing a joint clustering (heterogeneity) criterion for a partition (with a given number k of classes) and a set of k class prototypes, that is, the sum of the class-specific sums of dissimilarities between class elements and the corresponding class prototype.

The paper is organized as follows. In Section 2, we present the data structure and some references. In Section 3, we introduce the index of distance between objects and the steps of the algorithm. In Section 4, we provide a numerical example and do a comparative study with the classical approach. In Section 5, we explain how the algorithm is applied to cluster the workdays according to the degree of the traffic pollution at the most important roundabout. In combination with six weather conditions parameters measured on the same days, the resulting classes are analyzed and described in terms of six meteorological characteristics. In Section 6, we draw the corresponding conclusions.

2. The Data Structure

Let Ω be a set of n objects described by a set of d quantitative variables $\{V^j;\ j = 1, d\}$.

V^j is a map defined by

$$\begin{aligned} V^j &: \Omega \longrightarrow \mathbb{R}, \\ i &\longrightarrow V^j(i) = v_i^j \in \mathbb{R}, \end{aligned} \tag{1}$$

where v_i^j is the value taken by the individual i for the variable V^j.

We assume that the individual i is described by the matrix $\{T_i;\ i = 1, n\}$

$$i \hookrightarrow T_i = \begin{bmatrix} x_{i1}^{(1)} & \cdots & x_{i1}^{(d)} \\ \vdots & & \vdots \\ x_{iN_i}^{(1)} & \cdots & x_{iN_i}^{(d)} \end{bmatrix}. \tag{2}$$

(i) T_i, for example, represents the medical record of the patient i for d variables made in the daytime; $x_{il}^{(j)}$ represents, in this case, the value taken by the patient i for the variable V^j.

(ii) T_i contains in our study the value of the seven pollution parameters for the day i for the 24 hours of the day.

The input data are

$$X = [[T_1] \ \cdots \ [T_n]]. \tag{3}$$

3. Classical Approach

(i) A standardized principal component analysis on each table T_i leads to the construction of r_i orthogonal factor axes on which we project the N_i observations of the individual i, and we obtain new uncorrelated variables which give n systems of axes

$$\left\{\left\{\Delta_{u_1^{(i)}}, \ldots, \Delta_{u_{r_i}^{(i)}}\right\};\ i = 1, \ldots, n\right\}. \tag{4}$$

In order to compare the objects, we must be in the same reference frame. Thus the basic problem of the search of a compromise axis system is posed. This problem also concerns other disciplines of mathematics, especially in differential geometry [5]. The proposed criteria in literature, for the search of compromise space on which we project the objects to compare them in terms of proximity, are not really justified. The proposed technics are purely heuristics [7], available online for free. Relations between tables are also analyzed with Procrustes analysis and compromise factorial axes in the context of multiple factorial analysis. One important reference can be Gardner et al. [8].

Finally, the conclusion regarding Bouroche's [4] proposal is too reductive of the large domain of research so that this reference could be removed.

(ii) If the matrix T_i has the same dimension (N, d), in [6], an algorithm of k-means type is proposed based on the Hilbert-Schmidt inner product to classify these matrix objects. If T_i does not have the same dimension, we can envisage a step of completion in order to obtain a structure of juxtaposition of data tables of the same dimension.

$\exists i \neq l$ such that $N_i \neq N_l$; we can use the following procedure to complete the tables. We assume that $N_i > 1$; $\forall i = 1, n$. Let N be the least common multiple of N_i:

$$N = \mathrm{LCM}\left(N_i,\ i = 1, n\right). \tag{5}$$

There exists α_i so that

$$N = N_i \times \alpha_i. \tag{6}$$

Now, by duplicating α_i times each table T_i, we obtain a new table $\widehat{T_i}$ of dimension $N \times d$. So, if N_i is a large number, the least common multiple becomes necessarily large and the procedure leads to a structure of large tables. Moreover, this completion removes any chronological order of the data. It seems more reasonable to carry out the classification without processing with this completion step. It seems necessary to study the case where the tables T_i do not have the same dimension and without a reduction stage. If the hypothesis of normality of the observations in each column of table T_i is verified, this matrix T_i can then be considered as regrouping a realization of the normal random vector X whose distribution parameters (μ_i, Σ_i) should be estimated. These parameters will be estimated in an empirical way from the observations in the entry tables. The aim of the present paper is therefore to present a new approach of classification based on the k-means algorithm. This approach uses a new distance index based on the Mahalanobis discrimination scores. The proposed algorithm expands to the tables of different dimensions and is validated on real data of the traffic pollution.

4. Proposed Approach

4.1. Estimating the Distribution Parameters (μ_i, Σ_i). If the d variables are unspecified, for $i = 1, \ldots, n$, the mathematical expectations $\{\mu_i^j;\ j = 1, \ldots, d\}$ and the components of the estimated covariance matrix Σ_i are given by

$$\mu_i = \begin{pmatrix} \mu_i^1 \\ \vdots \\ \mu_i^d \end{pmatrix} \quad \text{with } \mu_i^j = \frac{1}{N_i}\sum_{l=1}^{N_i} x_{il}^j,$$
$$\begin{aligned}(\Sigma_i)_{jk} &= \operatorname{Cov}\left(V_{(i)}^j, V_{(i)}^k\right) = \Sigma_{jk}^i \\ &= \frac{1}{N_i - 1}\sum_{l=1}^{N_i}\left(x_{il}^j - \mu_i^j\right)\left(x_{il}^k - \mu_i^k\right).\end{aligned} \tag{7}$$

These estimators are unbiased, convergent, and consistent and do not depend on the number of observations or trials.

4.2. Classification Algorithm. We wish to gather the n individuals in K homogeneous classes. The heterogeneity of the classes is measured by a criterion of the inertia sum of the classes. This criterion is expressed by

$$\operatorname{Cr}(P, L) = \sum_{k=1}^{K}\left[\sum_{\{l\}\in P_{kj}} \delta^2(x_l, l_k)\right], \tag{8}$$

where l_k is the prototype or the kernel of the class P_k; x_l is the observation of the individual l; and δ is an index of distance between the objects and the prototype or representative elements of the classes. This criterion expresses the adequacy between the individuals with regard to the classes where they are affected.

4.3. Description of Individuals. We suppose that every table T_i groups a sample of size N_i of the Gaussian random vector of parameters (μ_i, Σ_i). For example, in the case of data with errors of measure, the tables data groups the repeated observations about the description of the variables. These observations are the realizations of the Gaussian random vector. It is clear that these observations are not correlated and the estimated variance covariance matrix is complete and thus not singular. Each i is described by $I_i = (N_i, \mu_i, \Sigma_i)$, where

(i) $N_i \in \mathbb{N}$, where N_i is the number of observations of the individual i;

(ii) $\mu_i \in \mathbb{R}^d$, where μ_i is the vector containing the estimated means for each variable;

(iii) $\Sigma_i \in M_d(\mathbb{R})$ is the set of real symmetric positive definite matrices of order d.

4.4. Distance between Individuals. Let i and l be 2 individuals described, respectively, by T_i and T_l. We wish to build an index of distance which takes into account the distribution parameters. To do this, we use the notion of discriminant score. For a realization $O_t^{(i)}$ of the individual i, the discriminant score of Mahalanobis of this observation with regard to the realizations of the individual l is given by

$$\operatorname{sc}^2\left(\frac{O_t^{(i)}}{T_l}\right) = \left\|\mu_i - \mu_l\right\|_{\Sigma_i^{-1}}^2, \tag{9}$$

where μ_i and μ_l are the average vector of the individuals i and l, respectively. It supposes that the tth observation of the individual i is assimilate to the average vector (empirical value of the distribution of its observations), for all the realizations of the individual i:

$$\operatorname{sc}^2\left(\frac{T_i}{T_l}\right) = \sum_{i=1}^{N_i}\left[\left\|\mu_i - \mu_l\right\|_{\Sigma_l^{-1}}^2\right] = N_i \times \left\|\mu_i - \mu_l\right\|_{\Sigma_l^{-1}}^2. \tag{10}$$

Similar arguments lead to

$$\operatorname{sc}^2\left(\frac{T_l}{T_i}\right) = \sum_{l=1}^{N_l}\left[\left\|\mu_i - \mu_l\right\|_{\Sigma_i^{-1}}^2\right] = N_l \times \left\|\mu_i - \mu_l\right\|_{\Sigma_l^{-1}}^2. \tag{11}$$

These scores are positive quantities and perfectly express the similarity between two individuals.

The map δ is defined by

$$\begin{gathered}\left(\mathbb{N}\times\mathbb{R}^d\times M_d(\mathbb{R})\right)\times\left(\mathbb{N}\times\mathbb{R}^d\times M_d(\mathbb{R})\right) \longrightarrow \mathbb{R}, \\ (I_1, I_2) \longrightarrow \delta(I_1, I_2), \\ \delta(I_1, I_2) = \sqrt{\frac{N_1\left\|\mu_1 - \mu_2\right\|_{\Sigma_1^{-1}}^2 + N_2\left\|\mu_1 - \mu_2\right\|_{\Sigma_2^{-1}}^2}{N_1 + N_2}},\end{gathered} \tag{12}$$

where δ is an index of weighted distance.

Without loss of generality, we assume that all objects are observed the same number of times; $N_i = N$ for all $i = 1, \ldots, n$. We assume that

(1) $\delta(I_1, I_2) = 0 \Leftrightarrow i_1 = i_2$;

(2) $2N > d$. This hypothesis implies that the matrices Σ_1 and Σ_2 are nonsingular.

4.5. Criteria and Optimization Problem. Let $\mathbb{P}_K$ be the set of partitions with K clusters and let $\mathbb{L}_K = (\mathbb{N}\times\mathbb{R}^d\times M_d(\mathbb{R}))^K$ be the set of K prototypes of the classes. For $k = 1, \ldots, K$, $l_k = (\widehat{N}_k, \widehat{\mu}_k, \widehat{\Sigma}_k)$. The criterion Cr writes

$$\begin{aligned}\operatorname{Cr}(P, L) &= \sum_{k=1}^{K}\left[\sum_{i\in P_k}\delta^2(I_i, l_k)\right] \\ &= \sum_{k=1}^{K}\sum_{i\in P_k}\left[\frac{N\left\|\mu_i - \widehat{\mu}_k\right\|_{\Sigma_i^{-1}}^2 + \widehat{N}_k\left\|\mu_i - \widehat{\mu}_k\right\|_{\widehat{\Sigma}_k^{-1}}^2}{N + \widehat{N}_k}\right]\end{aligned} \tag{13}$$

for $P = (P_1, \ldots, P_K) \in \mathbb{P}_K$ and $L = (l_1, \ldots, l_K) \in \mathbb{L}_K$.

We search (P^*, L^*) which realizes

$$\min_{\substack{P\in\mathbb{P}_K \\ L\in\mathbb{L}_K}} \operatorname{Cr}(P, L). \tag{14}$$

The algorithms used to solve such problems are of k-means type. These algorithms are based on the definition of the function of representation g and the function of affectation f which will be used alternatively to decrease the criterion. The representation function satisfies the following procedure:

$$\begin{gathered} g : \mathbb{P}_K \longrightarrow \mathbb{L}_K, \\ P = \{P_1, \dots, P_K\} \longrightarrow g(P) = \{l_1, \dots, l_K\}, \\ g \text{ must verify } \min_{L \in \mathbb{L}_K} \operatorname{Cr}(P, L) = \operatorname{Cr}(P, g(P)). \end{gathered} \tag{15}$$

4.6. Characterization of a Class of Individuals. We seek for the kernel from each of the classes generated by the algorithm. Let $\{1, \dots, n_k\}$ be the n_k individuals of the class P_k and let φ be the map defined by, without loss of generality,

$$\begin{aligned} &\varphi : \mathbb{N} \times \mathbb{R}^d \times M_d(\mathbb{R}) \longrightarrow \mathbb{R}_+, \\ &(N, \mu, \Sigma) \longrightarrow \varphi(N, \mu, \Sigma), \\ &\varphi(N, \mu, \Sigma) \\ &\quad = \sum_{l=1}^{n_k} \left[\frac{N_l}{N_l + N} \|\mu_l - \mu\|^2_{\Sigma_l^{-1}} + \frac{N}{N_l + N} \|\mu_l - \mu\|^2_{\Sigma^{-1}} \right]. \end{aligned} \tag{16}$$

Proposition 1. *μ^* and Σ^* which minimize φ are given by*

$$\begin{aligned} &\mu^* = (\Gamma)^{-1} \sum_{l=1}^{n_k} \left(\Sigma_l^{-1}\right)(\mu_l) \quad \textit{with } \Gamma = \sum_{l=1}^{n_k} \left(\Sigma_l^{-1}\right), \\ &\Sigma = 0, \\ &N^* = N. \end{aligned} \tag{17}$$

Proof. We focus on the case where, for all $i = 1, n$, $N_i = N$.

We research μ^* and Σ^* which minimize φ_N. We put $\mu = x$, $\mu_i = x_i$, and $\Sigma = S$. We have

$$\begin{aligned} &\varphi_N(x, S) = \varphi(x, S) \\ &\quad = \sum_{l=1}^{n_k} \left[\frac{N_l}{N_l + N} \|\mu_l - x\|^2_{\Sigma_l^{-1}} + \frac{N}{N_l + N} \|\mu_l - x\|^2_{\Sigma^{-1}} \right] \\ &\quad = \frac{1}{2} \sum_{l=1}^{n_k} \left[\|\mu_i - x\|^2_{\Sigma_i^{-1}} + \|\mu_i - x\|^2_{S^{-1}} \right] \quad \text{if } N_l = N, \end{aligned} \tag{18}$$

$$\min_{'x,S} \varphi(x, S).$$

The necessary condition is written as

$$\begin{aligned} \frac{\partial \varphi}{\partial x} &= 0 \ \ (\text{a}), \\ \frac{\partial \varphi}{\partial S} &= 0 \ \ (\text{b}). \end{aligned} \tag{19}$$

(a) One has $\partial\varphi/\partial x = 0$; then $\sum_{i=1}^{k} [(x_i - x)'(S_i^{-1} + S^{-1})] = 0 \Leftrightarrow$

$$\sum_{i=1}^{k} \left[(x_i - x)' S_i^{-1} \right] + \sum_{i=1}^{k} \left[(x_i - x)' S^{-1} \right] = 0. \tag{20}$$

We note that the first expression of φ does not depend on S. We put

$$\Phi(S) = \sum_{i=1}^{k} \|(x_i - x)\|^2_{S^{-1}}. \tag{21}$$

Then

$$\begin{aligned} &\Phi(S + H) - \Phi(S) \\ &\quad = \sum_{i=1}^{k} \left[(x_i - x)' (S + H)^{-1} (x_i - x) \right] \\ &\qquad - \sum_{i=1}^{k} \left[(x_i - x)' (S)^{-1} (x_i - x) \right] \end{aligned} \tag{22}$$

and also

$$(S + H)^{-1} = \left[S\left(I + S^{-1}H\right) \right]^{-1} = \left(I + S^{-1}H\right)^{-1} S^{-1}. \tag{23}$$

The expansion of $(I + S^{-1}H)^{-1}$ gives

$$\begin{aligned} &\left(I + S^{-1}H\right)^{-1} = I - S^{-1}H + \cdots \Longrightarrow \\ &(S + H)^{-1} = \left[I - S^{-1}H + \cdots\right] S^{-1} \Longrightarrow \\ &\Phi(S + H) - \Phi(S) \\ &\quad = \sum_{i=1}^{k} \left[(x_i - x)' \left(S^{-1} \cdot H \cdot S^{-1}\right) (x_i - x) \right] + \cdots + \\ &\quad = \sum_{i=1}^{K} \left[\left[S^{-1}(x_i - x)\right]' (H) \left[S(x_i - x)\right] \right] + \cdots + \\ &\quad = \sum_{i=1}^{k} \left\| S^{-1}(x_i - x) \right\|^2_H + \cdots. \end{aligned} \tag{24}$$

(b) $\partial\varphi/\partial S = 0$ implies that $\sum_{i=1}^{k} \|S^{-1}(x_i - x)\|^2_H = 0$; then

$$S^{-1}(x_i - x) = 0 \quad \forall i \Longrightarrow S^{-1} = 0 \text{ because } x_i \neq x. \tag{25}$$

Finally (1) $\Leftrightarrow \sum_{i=1}^{k} [(x_i - x)'(S_i^{-1})] = 0 \Rightarrow (x)' \sum_{i=1}^{k} (S_i^{-1}) = (\sum_{i=1}^{k} (x_i)'(S_i^{-1})) \Rightarrow$

$$\begin{aligned} x &= (\Gamma)^{-1} \sum_{i=1}^{k} \left(S_i^{-1}\right)(x_i), \\ \Gamma &= \sum_{i=1}^{k} \left(S_i^{-1}\right). \end{aligned} \tag{26}$$

□

Remark 2. In the case of measuring errors on the obtained classes, characterization is not flawed and is given exactly. This seems quite natural.

4.6.1. The Distance between Individual and a Class. The individual i is described by (μ_i, Σ_i) and the class C_k which contains n_k individuals is characterized by $l_k \in \mathbb{R}^d$ given by

$$l_k = (\Gamma_k)^{-1} \sum_{i=1}^{n_k} \left(\Sigma_i^{-1}\right)(\mu_i) \in \mathbb{R}^d, \tag{27}$$

where $\Gamma_k = \sum_{i=1}^{n_k}(\Sigma_i^{-1})$. The distance between the individual i and the class C_k is given by

$$\delta^2(I_i, C_k) = \frac{1}{2}\left\|\mu_i - l_k\right\|^2_{\Sigma_i^{-1}}. \tag{28}$$

The affectation function is given by

$$\begin{gathered} f : \mathbb{L}_K \longrightarrow \mathbb{P}_K, \\ L = \{l_1, \dots, l_K\} \longrightarrow f(L) = \{P_1, \dots, P_K\}, \\ f \text{ must verify } \min_{P \in \mathbb{P}_K} \operatorname{Cr}(P, L) = \operatorname{Cr}(f(L), L). \end{gathered} \tag{29}$$

The minimum is obtained by

$$\begin{gathered} P_k = \{i \in \Omega \text{ such that } \delta(I_i, l_k) \le \delta(I_i, l_t);\ \forall t \ne k,\ k \\ < t \text{ if equality}\}. \end{gathered} \tag{30}$$

4.6.2. Classification Algorithm. We choose $L^{(0)}$ and in all cases we alternatively use the functions f and g. The algorithm runs as follows:

$$\begin{gathered} L^{(0)} \xrightarrow{f} P^{(1)} \xrightarrow{g} L^{(1)} \xrightarrow{f} P^{(2)} \xrightarrow{g} \cdots \xrightarrow{f} P^{(n)} \\ \xmapsto{g} L^{(n)} \longmapsto \cdots P^{(*)} \xmapsto{g} L^{(*)}. \end{gathered} \tag{31}$$

The algorithm stops as soon as the partition does not change. We build two sequences V_n and U_n.

Proposition 3. *The sequence $U_n = \operatorname{Cr}(P^{(n)}, L^{(n)})$ is decreasing and converges.*

Proof. We have $U_{n+1} = \operatorname{Cr}(P^{(n+1)}, L^{(n+1)}) = \operatorname{Cr}(P^{(n+1)}, g(P^{(n+1)})) \le \operatorname{Cr}(P^{(n+1)}, L^{(n)})$ by definition of g; then $U_{n+1} \le \operatorname{Cr}(P^{(n+1)}, L^{(n)}) = \operatorname{Cr}(f(L^{(n)}), L^{(n)}) \le \operatorname{Cr}(P^{(n)}, L^{(n)}) = U_n$ by definition of f. □

Proposition 4. *The sequence $V_n = (P^{(n)}, L^{(n)})$ is stationary for a given rank.*

Proof. We put $U_n = \operatorname{Cr}(P^{(n)}, L^{(n)})$, $U^* = \operatorname{Cr}(P^*, L^*) \Rightarrow N_\epsilon$ exists; $n \ge N_\epsilon \Rightarrow |U_n - U^*| \le \epsilon \Rightarrow N_\epsilon$ exists; $n \ge N_\epsilon \Rightarrow P^{(n)} \simeq P^*; L^{(n)} \simeq L^* \Rightarrow V_n \simeq V^*$. □

5. Numerical Illustrative Example

We wish to cluster the six objects $\{1, \dots, 6\}$ into 2 clusters. Each object is described by three variables V^1, V^2, and V^3. The three variables are unspecified and we assume that the condition of normality of these observations is verified. The artificial input data is as follows:

Data Input

$$\underbrace{\begin{matrix} 3 & 5 & 6 \\ 7 & 9 & 10 \\ -5 & 2 & 0 \\ 3 & 1 & -1 \\ 3 & 5 & 2 \\ & & \end{matrix}}_{1\longrightarrow T_1} \underbrace{\begin{matrix} 2 & -1 & 0 \\ 3 & 5 & 11 \\ -2 & 3 & 4 \\ 6 & 7 & 8 \\ & & \\ & & \end{matrix}}_{2\longrightarrow T_2} \underbrace{\begin{matrix} 4.1 & 2.6 & -1 \\ 2.4 & 0.5 & 4 \\ 5 & 6 & 7 \\ 2 & 0.1 & 7 \\ & & \\ & & \end{matrix}}_{3\longrightarrow T_3} \underbrace{\begin{matrix} 3 & 3.6 & 4 \\ -1 & 17 & 0 \\ 10 & 3 & 12 \\ 13 & 0 & 11 \\ & & \\ & & \end{matrix}}_{4\longrightarrow T_4} \underbrace{\begin{matrix} 4 & -1 & 2 \\ 3 & 2 & 6 \\ 2 & 3 & 7 \\ 3 & 5 & -11 \\ 1 & 0.2 & 3 \\ 0.4 & 7 & 2 \end{matrix}}_{5\longrightarrow T_5} \underbrace{\begin{matrix} 2 & 7 & 2.5 \\ -1 & 3 & 1.5 \\ 2 & 5 & 0 \\ 1 & 3 & 1 \\ & & \\ & & \end{matrix}}_{6\longrightarrow T_6} \tag{32}$$

In (32) $T_1, \dots, T_6$ are not in the same dimension.

5.1. Classical Approach. Usually, we summarize observations of each individual by a central value that can be the mean. This method of data reduction can lead to erroneous results. This is shown in this numerical example.

(i) The mean value of each variable and the coordinate of final centers of clusters is obtained by using the k-means algorithm:

$$\begin{array}{cc} & \begin{matrix} \mu_i^1 & \mu_i^2 & \mu_i^3 \end{matrix} \\ X = & \begin{bmatrix} 2.6 & 4.2 & 3.4 \\ 2.25 & 3.5 & 5.75 \\ 3.83 & 3.03 & 3.33 \\ 6.25 & 5.9 & 6.75 \\ 2.23 & 2.7 & 1.5 \\ 2.6 & 3.24 & 4.2 \end{bmatrix} \longrightarrow \end{array}$$

$$\operatorname{Ker} = \begin{array}{ccc} & \text{class 1} & \text{class 2} \\ V_1 & 2.7 & 6.25 \\ V_2 & 3.3 & 5.0 \\ V_3 & 3.64 & 6.75 \end{array}. \tag{33}$$

5.2. The Proposed Approach

(i) The means μ_i are given as follows:

$$\begin{array}{ccccccc} & \mu_1 & \mu_2 & \mu_3 & \mu_4 & \mu_5 & \mu_6 \\ V_1 & 2.6 & 2.25 & 3.83 & 6.25 & 2.23 & 2.60 \\ V_2 & 4.2 & 3.50 & 3.03 & 5.90 & 2.70 & 3.24 \\ V_3 & 3.4 & 5.75 & 3.33 & 6.75 & 1.50 & 4.20 \end{array} \tag{34}$$

(ii) The matrix Σ_i^{-1} is given as

$$\begin{pmatrix} .01 & \cdot & \cdot \\ -.06 & 1.07 & \cdot \\ -.02 & -.65 & .48 \end{pmatrix} \begin{pmatrix} .12 & \cdot & \cdot \\ -.06 & .37 & \cdot \\ -.00 & .33 & .17 \end{pmatrix} \begin{pmatrix} 5.66 & \cdot & \cdot \\ -1.77 & 1.68 & \cdot \\ 1.25 & -.33 & .36 \end{pmatrix}, \\ \begin{pmatrix} .19 & \cdot & \cdot \\ .08 & .06 & \cdot \\ -.23 & -.08 & .34 \end{pmatrix} \begin{pmatrix} .94 & \cdot & \cdot \\ .28 & .21 & \cdot \\ .08 & .04 & .03 \end{pmatrix} \begin{pmatrix} 2.09 & \cdot & \cdot \\ -1.47 & 1.36 & \cdot \\ 2.13 & 1.45 & 2.13 \end{pmatrix}. \qquad (35)$$

5.3. The Final Partition Obtained and the Distance between Each Object and the Prototype of the Class Where It Is Assigned in the Case of Reduction and with Proposed Approach. Distances between objects and kernel of the class in the classical approach, after reduction step, are as follows:

$$\begin{bmatrix} \text{num.ind} & \text{cluster} & \text{distance} \\ 1 & 1 & .90305 \\ 2 & 1 & 2.16773 \\ 3 & 1 & 1.20818 \\ 4 & 2 & .00000 \\ 5 & 1 & 2.27794 \\ 6 & 1 & .58051 \end{bmatrix} \qquad (36)$$

Final partition is as follows:

$$C_1 = \{1, 2, 3, 5, 6\}, \\ C_2 = \{4\}. \qquad (37)$$

Distances between objects and kernel of the class in the proposed approach are as follows:

$$\begin{bmatrix} \text{case number} & \text{cluster} & \text{distance} \\ 1 & 1 & 4.201 \\ 2 & 2 & 8.040 \\ 3 & 1 & 3.598 \\ 4 & 2 & 8.040 \\ 5 & 1 & 3.324 \\ 6 & 1 & 3.411 \end{bmatrix} \qquad (38)$$

Final partition is as follows:

$$C_1 = \{2, 4\}, \\ C_2 = \{1, 3, 5, 6\}. \qquad (39)$$

Taking into account the variations that have resulted in errors of measurement and drop of the k-means algorithm with the weighted Mahalanobis distance it appeared that individuals 2 and 4 must be in the same class which has not been reported with the precedent procedure. In the case where the variability of observations plays an important part in the description of the individuals, the classification, made without taking into account these variabilities, leads to incorrect results compared with the reality of the data.

6. Application

Two files were used. The first file contains the observations of the seven parameters measuring the pollution caused by gases emitted by cars at a major intersection center of a city. The seven measured pollutants are *carbon monoxide*, *nitrogen monoxide*, *nitrogen dioxide*, *PM10 dust*, *sulfur dioxide*, *volatile organic compounds,* and *ozone*. These pollutants were measured each hour for each day. These observations concerned 420 days without gaps over the past three years. This file contains 420 tables of dimension 24 × 7 each. For these 420 days, we build up another file by measuring the daily average of 6 meteorological parameters: *temperature*, *rainfall*, *atmospheric pressure*, *humidity*, *wind speed,* and *hours of sunshine*. This table is of dimension 420 × 6. The interest is on the possible relationships between the variables measuring pollution and meteorological variables. We classify the days in three classes according to the degree of pollution and explained them using meteorological variables. Each day i is described by 7 curves corresponding to the 7 pollutants; the proposed algorithm, written in Matlab, brought together the 420 days in 3 classes without reduction step: *class 1* "low-pollution days," *class 2* "days of average pollution," and *class 3* "days of high pollution." The results are convincing; the profile of each class was explained by meteorological variables.

As a result of this, many questions arise, and we want to study the relationship between the pollution variable and the weather conditions variables. We also need to explain the classes in connection with the weather conditions variables and determine the profile of each class in connection with these weather conditions variables.

The first approach to this study has consisted in summarizing the pollution file (420 tables of dimension 24 × 7) in a table of dimension 420 × 7 by measuring the daily average for each pollutant. The variability effect of the measures is removed. We have studied the relationship between the 2 groups of variables "pollution and weather conditions." We are not interested in this approach.

The results are conclusive; the profile of each class has been explained by the weather conditions variables.

Table 1 shows the discriminating variables of each table. It describes the classes of pollution obtained according to the weather conditions parameters.

Characterization of the pollution classes by the weather conditions parameters at the station is as follows.

Class 1 is characterized by low temperatures, an important amount of rain, and strong winds with a minimum of sunshine. This corresponds to the class of weather disturbances.

Class 2 is characterized by the category of days with intermediary weather conditions between the stable situation and the weather disturbances situation.

Class 3 is characterized by high temperatures, light rainfalls, and a lot of sunshine. This represents the anticyclonic situation.

According to Table 1, we notice that the pressure is not a discriminating variable; that is, it does not help us differentiate between classes. Conversely temperature, rainfalls, humidity, wind speed, and sunshine do show the difference

Table 1

Classes		Temp	Rain	Bar-pres	Humd	Wind speed	Shine
Class 1	Mean	18.73	2.99	1015.55	70.82	3.47	5.79
	Standard deviation	5.60	7.46	4.92	10.60	1.62	3.64
Class 2	Mean	19.25	1.32	1015.80	70.99	2.39	5.99
	Standard deviation	4.79	4.33	6.30	11.50	1.04	3.55
Class 3	Mean	20.78	$687\cdot10^{-2}$	1016.00	67.73	1.85	7.60
	Standard deviation	2.97	0.38	4.25	9.76	0.82	3.62

especially between the low-pollution class and the high-pollution class.

According to Şen et al. (2006) [9], the pollution episodes in large cities are often related to high atmospheric pressure situations. The latter represent the ideal conditions for the gathering of pollutants in the air.

Since nowadays we can predict the weather conditions 2 days ahead, we can therefore take action and make such recommendations as to regulate the road traffic at the roundabout cited above and prevent it from falling in the high-pollution class.

7. Conclusion

In several scientific disciplines and particularity in medicine, the variability of the observations plays an important part. The proposed approach permits classifying objects by taking into account variability of the observations. The approach can be extended to the classification of matrix objects even of different dimensions and to functional data and this can integrate variability in the distribution which describes the object for each variable. Thus, each object will be described by a multidimensional distribution. This extension will be developed in future work.

Competing Interests

The author declares that they have no competing interests.

References

[1] F. D. A. T. De Carvalho and Y. Lechevallier, "Dynamic clustering of interval-valued data based on adaptive quadratic distances," *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, vol. 39, no. 6, pp. 1295–1306, 2009.

[2] A. Irpino and R. Verde, "Dynamic clustering of interval data using a Wasserstein-based distance," *Pattern Recognition Letters*, vol. 29, no. 11, pp. 1648–1658, 2008.

[3] F. D. A. T. De Carvalho and Y. Lechevallier, "Partitional clustering algorithms for symbolic interval data based on single adaptive distances," *Pattern Recognition*, vol. 42, no. 7, pp. 1223–1236, 2009.

[4] J. M. Bouroche, *Analyse des données ternaires [Ph.D. thesis]*, DACP Thése de l.Université de Paris IV, Paris, France, 1975.

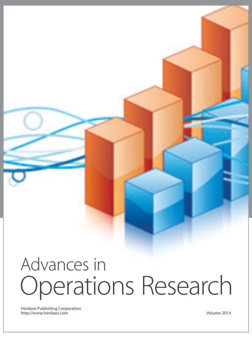
[5] P. Orlick and H. Terao, *Arrangements of Hyperplanes*, Springer, Berlin, Germany, 1992.

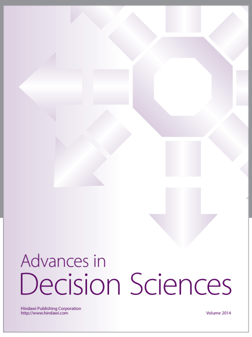
[6] A. Rebbouh, "Clustering the constituent elements of juxtaposition of measuring tables data," *Communications in Statistics*, vol. 32, no. 3, pp. 752–765, 2006.

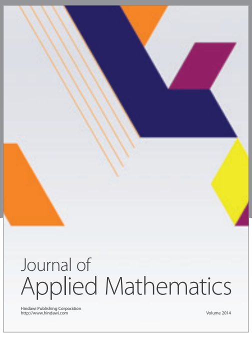
[7] E. Acar and B. Yener, "Unsupervised multiway data analysis: a literature survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 1, pp. 6–20, 2009.

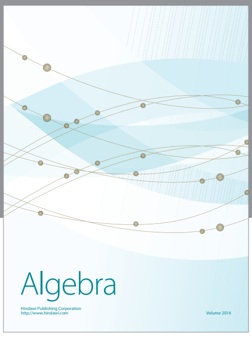
[8] S. Gardner, J. C. Gower, and N. J. le Roux, "A synthesis of canonical variate analysis, generalised canonical correlation and Procrustes analysis," *Computational Statistics & Data Analysis*, vol. 50, no. 1, pp. 107–134, 2006.

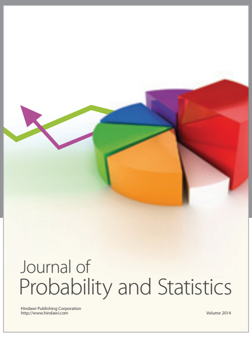
[9] Z. Şen, A. Altunkaynak, and M. Özger, "Space-time interpolation by combining air pollution and meteorologic variables," *Pure and Applied Geophysics*, vol. 163, no. 7, pp. 1435–1451, 2006.


Advances in
Operations Research

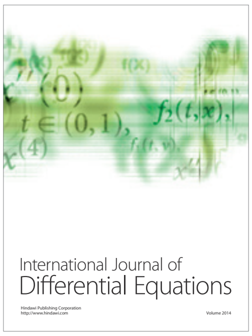

Advances in
Decision Sciences


Journal of
Applied Mathematics

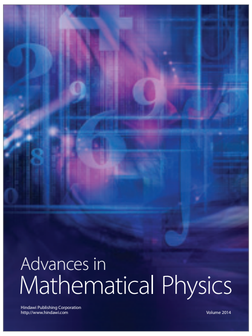

Algebra

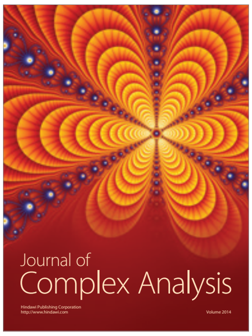

Journal of
Probability and Statistics

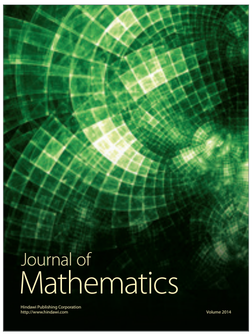
The Scientific
World Journal

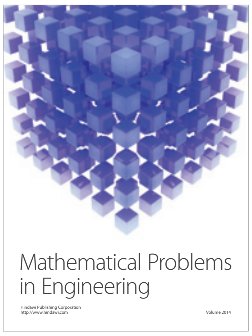

International Journal of
Differential Equations

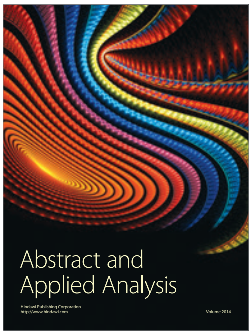
International Journal of
Combinatorics

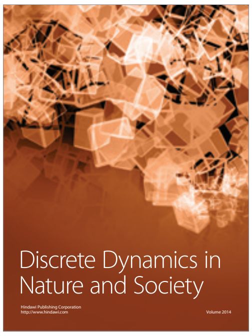

Advances in
Mathematical Physics

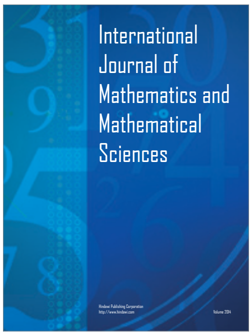

Journal of
Complex Analysis


Journal of
Mathematics


Mathematical Problems
in Engineering

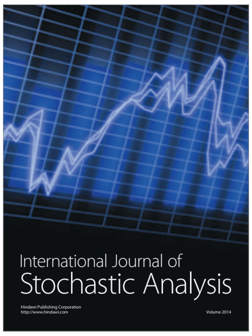

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


International Journal of
Stochastic Analysis

Journal of
Optimization