Available online at www.sciencedirect.com
ScienceDirect
Fuzzy Sets and Systems 157 (2006) 2833–2857
FUZZY sets and systems
www.elsevier.com/locate/fss

# Partitional fuzzy clustering methods based on adaptive quadratic distances

Francisco de A.T. de Carvalho*, Camilo P. Tenório, Nicomedes L. Cavalcanti Junior

*Centro de Informática, Universidade Federal de Pernambuco, Caixa Postal 7851, CEP 50732-970, Recife (PE), Brazil*

Received 2 August 2005; received in revised form 27 April 2006; accepted 7 June 2006
Available online 30 June 2006

**Abstract**

This paper presents partitional fuzzy clustering methods based on adaptive quadratic distances. The methods presented furnish a fuzzy partition and a prototype for each cluster by optimizing an adequacy criterion based on adaptive quadratic distances. These distances change at each algorithm iteration and can either be the same for all clusters or different from one cluster to another. Moreover, various fuzzy partition and cluster interpretation tools are introduced. Experiments with real and synthetic data sets show the usefulness of these adaptive fuzzy clustering methods and the merit of the fuzzy partition and cluster interpretation tools.


*Keywords:* Clustering analysis; Partitional fuzzy clustering methods; Adaptive quadratic distances; Fuzzy partition interpretation; Fuzzy cluster interpretation

## 1. Introduction

Cluster analysis seeks to organize a set of items (each item usually represented as a vector of quantitative values in a multidimensional space) into clusters such that items (patterns, objects, individuals, etc.) within a given cluster have a high degree of similarity and items belonging to different clusters have a high degree of dissimilarity [5,15]. Cluster analysis techniques can be divided into hierarchical and partitional methods [18,12,10].

Hierarchical methods yield complete hierarchy, i.e., a nested sequence of partitions of the input data. Hierarchical methods can be agglomerative or divisive. Agglomerative methods yield a sequence of nested partitions starting with trivial clustering in which each item is in a unique cluster and ending with trivial clustering in which all items are in the same cluster. A divisive method starts with all items in a single cluster and performs a splitting procedure until a stopping criterion is met (usually upon obtaining a partition of singleton clusters).

Partitioning a set of items into a predefined number of clusters is an important topic in data analysis, pattern recognition and image processing [15]. Partitional methods seek to obtain a single partition of the input data into a fixed number of clusters. Such methods often look for a partition that optimizes (usually locally) an adequacy criterion function. To improve cluster quality, the algorithm is run multiple times at different starting points and the best configuration obtained from the total runs is used as the output clustering.

* Corresponding author. Tel.: +55 81 21268430; fax: +55 81 21268438.
*E-mail addresses:* fatc@cin.ufpe.br (F.de A.T.de Carvalho), cpt@cin.ufpe.br (C.P. Tenório), nlcj@cin.ufpe.br (N.L. Cavalcanti Junior).

0165-0114/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.fss.2006.06.004

Partitional methods can be divided into hard and fuzzy clustering. Hard clustering furnishes a hard partition where each pattern of the data set is assigned to one and only one cluster. Fuzzy clustering [2,3] generates a fuzzy partition that furnishes the degree of membership of each pattern in a given cluster. In real situations, fuzzy clustering is more natural than hard clustering, as objects on the boundaries between several clusters are not forced to belong fully to one of the clusters, but rather are assigned membership degrees between 0 and 1 indicating their partial memberships.

Partitioning dynamic cluster algorithms [8] are iterative two-step relocation hard clustering algorithms involving the construction of clusters at each iteration and the identification of a suitable representative or prototype (means, factorial axes, probability laws, groups of elements, etc.) for each cluster by locally optimizing an adequacy criterion between clusters and their corresponding prototypes. This optimization process begins from a set of prototypes or an initial partition and interactively applies an *allocation step* (the prototypes are fixed) in order to assign patterns to clusters according to their proximity to the prototypes. This is followed by a *representation step* (the partition is fixed) where the prototypes are updated according to the assignment of the patterns in the allocation step until achieving the convergence of the algorithm, when the adequacy criterion reaches a stationary value.

The adaptive dynamic hard clustering algorithm [7] also optimizes a criterion based on a fitting measure between clusters and their prototypes, but the distances used to compare clusters and their prototypes change at each iteration. These distances are not determined absolutely and can be different from one cluster to another. The advantage of these adaptive distances is that the clustering algorithm is able to recognize clusters of different shapes and sizes.

The initialization, allocation step and stopping criterion are nearly the same in the adaptive and non-adaptive dynamic hard clustering algorithms. The main difference between these algorithms lies in the representation step, which has two stages in the adaptive case. The first stage, where the partition and the distances are fixed and the prototypes are updated, is followed by a second one, where the partition and their corresponding prototypes are fixed and the distances are updated.

Fuzzy set theory applied in cluster analysis focuses mainly on fuzzy clustering based on fuzzy relations and on fuzzy clustering based on objective functions. One of the first fuzzy clustering methods, based on an objective function defined by the Euclidean distance, was presented by Dunn [9] and further generalized by Bezdek [2]. Fuzzy clustering algorithms based on the $L_1$ norm have been introduced by Jajuga [16]. Yang [19] presents a comprehensible survey of fuzzy clustering methods. Groenen and Jajuga [13] introduced a fuzzy clustering model based on the root of the squared Minkowsky distance including squared and unsquared Euclidean distances and the $L_1$ distance.

The first adaptive fuzzy clustering algorithm, based on a quadratic distance defined by a fuzzy covariance matrix, was introduced by Gustafson and Kessel [14]. A detailed study of this former algorithm was presented by Krishnapuram and Kim [17]. An improved approach for estimating the fuzzy covariance matrix in the Gustafson–Kessel (GK) algorithm was proposed by Babuska et al. [1]. Later, Frigni and Nasraoui [11] proposed an approach based on the standard fuzzy $c$-means (FCM) algorithm to address the problem of clustering and feature weighting simultaneously.

This paper presents partitional fuzzy clustering methods based on adaptive quadratic distances. These adaptive quadratic distances are defined by positive definite symmetric matrices that must be inverted in order to update the distances in the second stage of the representation step of the algorithm. However, in cases where these matrices are virtually singular, numerical problems can occur and the matrices cannot be inverted. One solution for overcoming this problem is to use the approach proposed by Babuska et al. [1]. In the present paper, we also considered fuzzy clustering methods based on adaptive Euclidean distances defined by positive definite diagonal matrices that are easily inverted. Furthermore, various fuzzy partition and cluster interpretation tools based on the sum of squares (SSQs) are introduced.

Section 2 reviews the standard FCM clustering algorithm. Section 3 presents adaptive versions of the standard FCM clustering algorithm. Section 3.1 introduces a fuzzy clustering method based on a single adaptive quadratic distance defined by a fuzzy pooled covariance matrix (Section 3.1.1) and a special case of this model based on a single adaptive quadratic distance defined by a pooled fuzzy covariance matrix restricted to be diagonal (Section 3.1.2). These models are an extension of the model introduced by Celeux et al. [6] in which a single crisp pooled covariance matrix is used and the adaptive distance is therefore unique. Section 3.2 presents the fuzzy clustering models introduced by Gustafson and Kessel [14] based on a different adaptive quadratic distance for each cluster defined by a fuzzy covariance matrix (Section 3.2.1) and introduces a special case of this method based on a different adaptive quadratic distance for each cluster defined by a fuzzy covariance matrix restricted to be diagonal (Section 3.2.2). In Section 4, we propose various tools for fuzzy Partition and cluster interpretation based on the SSQs: indices for evaluating the quality of a partition, the homogeneity and eccentricity of the individual clusters, and the role of the different variables in the cluster formation

process. To show the usefulness of these adaptive versions of the standard FCM clustering algorithm and the merit of these cluster interpretation tools, experiments with simulated data in the framework of a Monte Carlo schema as well as applications with real data sets are considered (Section 5). Section 6 gives the concluding remarks.

## 2. Standard FCM method

Let $\Omega$ be a set of $n$ patterns described by $p$ quantitative variables $\{y_1, \ldots, y_j, \ldots, y_p\}$. Pattern $k$ ($k = 1, \ldots, n$) is represented by a vector of quantitative feature values $\mathbf{x}_k = (x_k^1, \ldots, x_k^j, \ldots, x_k^p)$.

The standard FCM clustering algorithm [2,3] is a non-hierarchical clustering method the aim of which is to furnish a fuzzy partition of a set of patterns in $c$ clusters $\{P_1, \ldots, P_c\}$ and a corresponding set of prototypes $\{\mathbf{g}_1, \ldots, \mathbf{g}_c\}$ such that a criterion $J1$ measuring the fitting between the clusters and their prototypes is locally minimized. This criterion $J1$ is defined as [2]

$$J1 = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m \phi(\mathbf{x}_k, \mathbf{g}_i) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m \sum_{j=1}^{p}(x_k^j - g_i^j)^2, \tag{1}$$

where

$$\phi(\mathbf{x}_k, \mathbf{g}_i) = \sum_{j=1}^{p}(x_k^j - g_i^j)^2, \tag{2}$$

is the square of the Euclidean distance measuring the dissimilarity between a pair of vectors of quantitative feature values, $\mathbf{x}_k = (x_k^1, \ldots, x_k^p)$ is the quantitative feature vector describing the $k$th pattern, $\mathbf{g}_i = (g_i^1, \ldots, g_i^p)$ is the prototype of cluster $P_i$, $u_{ik}$ is the membership degree of pattern $k$ in cluster $P_i$ and $m \in (1, \infty)$ is a parameter that controls the fuzziness of membership for each pattern $k$.

The algorithm sets an initial membership degree for each pattern $k$ in each cluster $P_i$ and alternates a *representation step* and an *allocation step* until convergence, when the criterion $J1$ reaches a stationary value representing a local minimum.

In the representation step, the membership degree of each pattern $k$ in cluster $P_i$ is fixed and the prototype $\mathbf{g}_i = (g_i^1, \ldots, g_i^p)$ of cluster $P_i$ ($i = 1, \ldots, c$), which minimizes the clustering criterion $J1$, are updated according to the following expression [2,3]:

$$\mathbf{g}_i = \frac{\sum_{k=1}^{n}(u_{ik})^m \mathbf{x}_k}{\sum_{k=1}^{n}(u_{ik})^m}. \tag{3}$$

In the allocation step, each prototype $\mathbf{g}_i$ of cluster $P_i$ ($i = 1, \ldots, c$) is fixed and the membership degree $u_{ik}$ ($k = 1, \ldots, n$) of each pattern $k$ in each cluster $P_i$, minimizing the clustering criterion $J1$ under $u_{ik} \geqslant 0$ and $\sum_{i=1}^{c} u_{ik} = 1$, is updated according to the following expression [2,3]:

$$u_{ik} = \left[\sum_{h=1}^{c}\left\{\frac{\sum_{j=1}^{p}(x_k^j - g_i^j)^2}{\sum_{j=1}^{p}(x_k^j - g_h^j)^2}\right\}^{1/(m-1)}\right]^{-1} \quad \text{for } i = 1, \ldots, c. \tag{4}$$

## 3. Adaptive fuzzy clustering methods

This section presents adaptive versions of the FCM algorithm. The advantage of these methods is that the clustering algorithm is able to find clusters of different shapes and sizes [6–8,14]).

### 3.1. Fuzzy clustering algorithms based on a single adaptive distance

Here, we present fuzzy clustering methods for quantitative data based on a single adaptive distance. The main idea of these methods is that there is a distance to compare clusters and their prototypes that changes at each iteration but that is the same for all clusters [6–8].

These adaptive methods look for a fuzzy partition of a set of patterns in $c$ clusters $\{P_1, \ldots, P_c\}$, the corresponding $c$ prototypes $\{\mathbf{g}_1, \ldots, \mathbf{g}_c\}$ and an adaptive quadratic distance such that a criterion measuring the fitting between the clusters and their representatives (prototypes) is locally minimized.

*3.1.1. Fuzzy clustering method based on a single adaptive quadratic distance defined by a pooled fuzzy covariance matrix (FCM-PFCV)*

The model considered here is an extension of the model introduced by [6, pp. 88–91], which uses a single pooled crisp covariance matrix that defines a single adaptive distance. This method optimizes a criterion $J2$ measuring the fitting between the clusters and their representatives, which is defined as

$$J2 = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m d^2_{\mathbf{M}}(\mathbf{x}_k, \mathbf{g}_i) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m (\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}}\mathbf{M}(\mathbf{x}_k - \mathbf{g}_i), \tag{5}$$

where $\mathbf{x}_k$, $\mathbf{g}_i$, $u_{ik}$ and $m$ are defined as before and

$$d^2_{\mathbf{M}}(\mathbf{x}_k, \mathbf{g}_i) = (\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}}\mathbf{M}(\mathbf{x}_k - \mathbf{g}_i) \tag{6}$$

is a quadratic distance defined by a positive definite symmetric matrix $\mathbf{M}$ that is the same for all the clusters.

The algorithm starts from an initial membership degree for each pattern $k$ in each cluster $P_i$ and alternates a representation step and an allocation step until the convergence of the algorithm, when the criterion $J2$ reaches a stationary value representing a local minimum.

The representation step has now two stages. In the first stage, the membership degree of each pattern $k$ on each cluster $P_i$ and the matrix $\mathbf{M}$ are fixed.

**Proposition 3.1.** *The prototypes* $\mathbf{g}_i = (g_i^1, \ldots, g_i^p)$ *of cluster* $P_i$ $(i = 1, \ldots, c)$, *which minimize the clustering criterion* $J2$, *are updated according to Eq.* (3).

**Proof.** As the membership degree of each pattern $k$ in cluster $P_i$, as well as the parameter $m$ and the matrix $\mathbf{M}$ are fixed, we can rewrite the criterion $J2$ as

$$J2 = \sum_{i=1}^{c} J_i^2 \quad \text{with } J_i^2 = \sum_{k=1}^{n}(u_{ik})^m (\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}}\mathbf{M}(\mathbf{x}_k - \mathbf{g}_i).$$

The criterion $J2$ is additive so that the problem becomes finding $\mathbf{g}_i$ which minimizes $J_i^2$, for $i = 1, \ldots, c$. Let $\mathbf{y} = \sum_{k=1}^{n}(u_{ik})^m \mathbf{x}_k^j / \sum_{k=1}^{n}(u_{ik})^m$. We have

$$\begin{aligned}
J_i^2 &= \sum_{k=1}^{n}(u_{ik})^m (\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}}\mathbf{M}(\mathbf{x}_k - \mathbf{g}_i) \\
&= \sum_{k=1}^{n}(u_{ik})^m [(\mathbf{x}_k - \mathbf{y}) + (\mathbf{y} - \mathbf{g}_i)]^{\mathrm{T}}\mathbf{M}[(\mathbf{x}_k - \mathbf{y}) + (\mathbf{y} - \mathbf{g}_i)] \\
&= \sum_{k=1}^{n}(u_{ik})^m (\mathbf{x}_k - \mathbf{y})^{\mathrm{T}}\mathbf{M}(\mathbf{x}_k - \mathbf{y}) + \left[\sum_{k=1}^{n}(u_{ik})^m (\mathbf{x}_k - \mathbf{y})^{\mathrm{T}}\right]\mathbf{M}(\mathbf{y} - \mathbf{g}_i) \\
&\quad + (\mathbf{y} - \mathbf{g}_i)^{\mathrm{T}}\mathbf{M}\left[\sum_{k=1}^{n}(u_{ik})^m (\mathbf{x}_k - \mathbf{y})\right] + (\mathbf{y} - \mathbf{g}_i)^{\mathrm{T}}\mathbf{M}(\mathbf{y} - \mathbf{g}_i)\sum_{k=1}^{n}(u_{ik})^m \\
&= \sum_{k=1}^{n}(u_{ik})^m (\mathbf{x}_k - \mathbf{y})^{\mathrm{T}}\mathbf{M}(\mathbf{x}_k - \mathbf{y}) + (\mathbf{y} - \mathbf{g}_i)^{\mathrm{T}}\mathbf{M}(\mathbf{y} - \mathbf{g}_i)\sum_{k=1}^{n}(u_{ik})^m
\end{aligned}$$

because $\sum_{k=1}^{n}(u_{ik})^m (\mathbf{x}_k - \mathbf{y})^{\mathrm{T}} = \sum_{k=1}^{n}(u_{ik})^m (\mathbf{x}_k - \mathbf{y}) = \mathbf{0}$. It is clear from the expression above that $J_i^2$ is minimized when $\mathbf{g}_i = \mathbf{y} = \sum_{k=1}^{n}(u_{ik})^m \mathbf{x}_k^j / \sum_{k=1}^{n}(u_{ik})^m$. □

In the second stage, the membership degree $u_{ik}$ of each pattern $k$ in each cluster $P_i$ and the prototypes $\mathbf{g}_i$ of corresponding clusters $P_i$ $(i = 1, \ldots, c)$ are fixed.

**Proposition 3.2.** *The positive definite symmetric matrix* $\mathbf{M}$, *which minimizes the clustering criterion* $J2$ *under* $\det(\mathbf{M}) = 1$, *is updated according to the following expression*:

$$\mathbf{M} = [\det(\mathbf{Q})]^{1/p} (\mathbf{Q})^{-1}, \quad \mathbf{Q} = \sum_{i=1}^{c} \mathbf{C}_i \quad \text{and} \quad \mathbf{C}_i = \sum_{k=1}^{n} (u_{ik})^m (\mathbf{x}_k - \mathbf{g}_i)(\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}}. \tag{7}$$

**Proof.** As the membership degree $u_{ik}$ of each pattern $k$ in each cluster $P_i$, as well as the parameter $m$ and the prototypes $\mathbf{g}_i$ of corresponding clusters $P_i$ $(i = 1, \ldots, c)$ are fixed, we can rewrite the criterion $J2$ as a function of $\mathbf{M}$:

$$J2(\mathbf{M}) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m (\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}} \mathbf{M} (\mathbf{x}_k - \mathbf{g}_i).$$

Let $g(\mathbf{M}) = 1 - \det(\mathbf{M})$. We want to determine the extremes of $J2(\mathbf{M})$ with the restriction $g(\mathbf{M}) = 0$. Let the Lagrangian function be

$$\begin{aligned} L(\mathbf{M}, \beta) &= J2(\mathbf{M}) + \beta\, g(\mathbf{M}) \\ &= \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m (\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}} \mathbf{M} (\mathbf{x}_k - \mathbf{g}_i) + \beta(1 - \det(\mathbf{M})). \end{aligned}$$

Taking the derivative,

$$\frac{\mathrm{d}L(\mathbf{M}, \beta)}{\mathrm{d}\mathbf{M}} = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m (\mathbf{x}_k - \mathbf{g}_i)(\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}} - \beta \det(\mathbf{M})\, \mathbf{M}^{-1} = \mathbf{0}.$$

It follows that, $\mathbf{M}^{-1} = (1/\beta)\mathbf{Q}$, where $\mathbf{Q} = \sum_{i=1}^{c} \mathbf{C}_i$ and $\mathbf{C}_i = \sum_{k=1}^{n} (u_{ik})^m (\mathbf{x}_k - \mathbf{g}_i)(\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}}$, because $\det(\mathbf{M}) = 1$.

As $\det(\mathbf{M}^{-1}) = 1/\det(\mathbf{M}) = 1$, from $\mathbf{M}^{-1} = (1/\beta)\mathbf{Q}$ it follows that $\det(\mathbf{M}^{-1}) = (1/\beta^p)\det(\mathbf{Q}) = 1 \Rightarrow \beta = (\det(\mathbf{Q}))^{1/p}$. Moreover, as $\mathbf{M}^{-1} = (1/\beta)\mathbf{Q} = (1/(\det(\mathbf{Q}))^{1/p})\mathbf{Q}$, it follows also that $\mathbf{M} = (\det(\mathbf{Q}))^{1/p}\, \mathbf{Q}^{-1}$.

We know that $(\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}} \mathbf{M} (\mathbf{x}_k - \mathbf{g}_i) = \mathrm{trace}\left[(\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}} \mathbf{M} (\mathbf{x}_k - \mathbf{g}_i)\right] \Rightarrow (\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}} \mathbf{M} (\mathbf{x}_k - \mathbf{g}_i) = \mathrm{trace}[(\mathbf{x}_k - \mathbf{g}_i)(\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}} \mathbf{M}]$.

It follows that $J2(\mathbf{M}) = \mathrm{trace}[(\sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m (\mathbf{x}_k - \mathbf{g}_i)(\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}})\mathbf{M}] = \mathrm{trace}[\mathbf{Q}\mathbf{M}]$.

An extremum value of $J2$ is reached when $\mathbf{M} = (\det(\mathbf{Q}))^{1/p}\, \mathbf{Q}^{-1}$. This extremum value is $J2((\det(\mathbf{Q}))^{1/p}\, \mathbf{Q}^{-1}) = \mathrm{trace}[\mathbf{Q}(\det(\mathbf{Q}))^{1/p} \mathbf{Q}^{-1}] = p(\det(\mathbf{Q}))^{1/p}$.

On the other hand, $J2(\mathbf{I}) = \mathrm{trace}(\mathbf{Q}\mathbf{I}) = \mathrm{trace}(\mathbf{Q})$. As a positive definite symmetric matrix, $\mathbf{Q} = \mathbf{P}\boldsymbol{\Delta}\mathbf{P}^{\mathrm{T}}$ (according to the singular value decomposition procedure) where: $\mathbf{P}\mathbf{P}^{\mathrm{T}} = \mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I}$, $\boldsymbol{\Delta} = \mathrm{diag}(\delta_1, \ldots, \delta_p)$, and $\delta_j$ $(j = 1, \ldots, p)$ are the eigenvalues of $\mathbf{Q}$. Thus, $J2(\mathbf{I}) = \mathrm{trace}(\mathbf{P}\boldsymbol{\Delta}\mathbf{P}^{\mathrm{T}}) = \mathrm{trace}(\boldsymbol{\Delta}) = \sum_{j=1}^{p} \delta_j$. Moreover, $\det(\mathbf{Q}) = \det(\mathbf{P}\boldsymbol{\Delta}\mathbf{P}^{\mathrm{T}}) = \det(\boldsymbol{\Delta}) = \prod_{j=1}^{p} \delta_j$.

As it is well known that the arithmetic mean is greater than the geometric mean, i.e., $(1/p)\,(\delta_1 + \cdots + \delta_p) > \{\delta_1 \times \cdots \times \delta_p\}^{1/p}$ (the equality holds only if $\delta_1 = \cdots = \delta_p$), it follows that $J2(\mathbf{I}) > J2((\det(\mathbf{Q}))^{1/p}\mathbf{Q}^{-1})$. Thus, we conclude that this extreme is a minimum. □

**Remark.** The matrix $\mathbf{C}_i$ is related to the fuzzy covariance matrix in the $i$th cluster, and therefore the matrix $\mathbf{M}$ is related to the pooled fuzzy covariance matrix.

In the allocation step, the prototypes $\mathbf{g}_i$ of the corresponding clusters $P_i$ $(i = 1, \ldots, c)$ and the matrix $\mathbf{M}$ are fixed.

**Proposition 3.3.** *The membership degree* $u_{ik}$ $(k = 1, \ldots, n)$ *of each pattern* $k$ *in each cluster* $P_i$ $(i = 1, \ldots, c)$, *minimizing the clustering criterion* $J2$ *under* $u_{ik} \geqslant 0$ *and* $\sum_{i=1}^{c} u_{ik} = 1$, *is updated according to the*

*following expression*:

$$u_{ik} = \left[\sum_{h=1}^{c}\left(\frac{d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{g}_i)}{d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{g}_h)}\right)^{1/(m-1)}\right]^{-1} = \left[\sum_{h=1}^{c}\left(\frac{(\mathbf{x}_k-\mathbf{g}_i)^{\mathrm{T}}\mathbf{M}(\mathbf{x}_k-\mathbf{g}_i)}{(\mathbf{x}_k-\mathbf{g}_h)^{\mathrm{T}}\mathbf{M}(\mathbf{x}_k-\mathbf{g}_h)}\right)^{1/(m-1)}\right]^{-1}. \tag{8}$$

**Proof.** As the prototypes $\mathbf{g}_i$ of corresponding clusters $P_i$ $(i = 1,\ldots,c)$ as well as the parameter $m$ and the matrix $\mathbf{M}$ are fixed, we can rewrite $J2$ as

$$J2(\mathbf{u}_1,\ldots,\mathbf{u}_n) = \sum_{k=1}^{n} J_k^2(\mathbf{u}_k), \quad J_k^2(\mathbf{u}_k) = J_k^2(u_{1k},\ldots,u_{ck}) = \sum_{i=1}^{c}(u_{ik})^m d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{g}_i).$$

The criterion $J2$ is additive so the problem becomes minimizing $J_k^2$ $(k = 1,\ldots,n)$. Let $g_k(\mathbf{u}_k) = g_k(u_{1k},\ldots,u_{ck}) = \sum_{i=1}^{c} u_{ik} - 1$. We want to determine the extremum of $J_k^2(u_{1k},\ldots,u_{ck})$ under $g_k(u_{1k},\ldots,u_{ck}) = 0$. To do so, we use the method of Lagrange multipliers. This means solving the equation

$$\operatorname{grad} J_k^2(u_{1k},\ldots,u_{ck}) = \mu \operatorname{grad} g_k(u_{1k},\ldots,u_{ck}).$$

We have, $\operatorname{grad} J_k^2(u_{1k},\ldots,u_{ck}) = (m(u_{1k})^{m-1}d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{g}_1),\ldots,m(u_{ck})^{m-1}d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{g}_c))$ and $\operatorname{grad} g_k(u_{1k},\ldots,u_{ck}) = (1,\ldots,1)$.

Then,

$$\begin{aligned} &m(u_{1k})^{m-1}d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{g}_1) = \mu,\ldots,m(u_{ck})^{m-1}d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{g}_c) = \mu \\ &\Rightarrow u_{1k} = \left(\frac{\mu}{m}\right)^{1/(m-1)}\left(\frac{1}{d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{g}_1)}\right)^{1/(m-1)},\ldots,u_{ck} = \left(\frac{\mu}{m}\right)^{1/(m-1)}\left(\frac{1}{d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{g}_c)}\right)^{1/(m-1)}. \end{aligned}$$

As we know that $\sum_{h=1}^{c} \mu_{hk} = 1$, then:

$$\left(\frac{\mu}{m}\right)^{1/(m-1)} = \frac{1}{\sum_{h=1}^{c}(1/d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{g}_h))^{1/(m-1)}}.$$

It follows that,

$$u_{ik} = \frac{(1/d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{g}_i))^{1/(m-1)}}{\sum_{h=1}^{c}(1/d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{g}_h))^{1/(m-1)}} = \left[\sum_{h=1}^{c}\left(\frac{(\mathbf{x}_k-\mathbf{g}_i)^{\mathrm{T}}\mathbf{M}(\mathbf{x}_k-\mathbf{g}_i)}{(\mathbf{x}_k-\mathbf{g}_h)^{\mathrm{T}}\mathbf{M}(\mathbf{x}_k-\mathbf{g}_h)}\right)^{1/(m-1)}\right]^{-1}.$$

Thus, an extremum of $J_k^2$ is reached when

$$u_{ik} = \left[\sum_{h=1}^{c}\left(\frac{d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{g}_i)}{d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{g}_h)}\right)^{1/(m-1)}\right]^{-1}.$$

We have,

$$\frac{\partial(J_k^2)}{\partial(u_{ik})} = m(u_{ik})^{(m-1)}d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{g}_i) \;\Rightarrow\; \frac{\partial^2(J_k^2)}{\partial(u_{ik})^2} = m\,(m-1)(u_{ik})^{(m-2)}d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{g}_i)$$

and

$$\frac{\partial^2(J_k^2)}{\partial(u_{ik})\partial(u_{jk})} = 0 \quad \forall\, j \neq i.$$

The Hessian matrix of $J_k^2$ evaluated at $\mathbf{u}_k = (u_{1k}, \ldots, u_{ck})$ is

$$H(\mathbf{u}_k) = \begin{pmatrix} \dfrac{m(m-1)d^2_{\mathbf{M}}(\mathbf{x}_k, \mathbf{g}_1)}{\left[\sum_{h=1}^{c}\left(\dfrac{d^2_{\mathbf{M}}(\mathbf{x}_k, \mathbf{g}_1)}{d^2_{\mathbf{M}}(\mathbf{x}_k, \mathbf{g}_h)}\right)^{(m-2)/(m-1)}\right]} & 0 \cdots & 0 \\ \vdots & \vdots \ddots & \vdots \\ 0 & 0 \cdots & \dfrac{m(m-1)d^2_{\mathbf{M}}(\mathbf{x}_k, \mathbf{g}_c)}{\left[\sum_{h=1}^{c}\left(\dfrac{d^2_{\mathbf{M}}(\mathbf{x}_k, \mathbf{g}_c)}{d^2_{\mathbf{M}}(\mathbf{x}_k, \mathbf{g}_h)}\right)^{(m-2)/(m-1)}\right]} \end{pmatrix},$$

where $H(\mathbf{u}_k)$ is positive definite so that we can conclude that this extremum is a minimum. □

*3.1.1.1. Algorithm.* The fuzzy clustering algorithm based on a unique adaptive quadratic distance defined by a pooled fuzzy covariance matrix (FCM-PFCV) is executed in the following steps:

SCHEMA OF THE FCM-PFCV CLUSTERING ALGORITHM

(1) **Initialization**
Fix $c$, $2 \leqslant c < n$; fix $m$,$1 < m < \infty$; fix $T$ (an iteration limit); and fix $\varepsilon > 0$; Initialize $u_{ik}$ ($k = 1, \ldots, n$ and $i = 1, \ldots, c$) of pattern $k$ belonging to cluster $P_i$ such that $u_{ik} > 0$ and $\sum_{i=1}^{c} u_{ik} = 1$
(2) $t = 1$
(3) **Representation step**:
(a) *Stage* 1:
{the membership degree $u_{ik}$ of pattern $k$ belonging to cluster $P_i$ is fixed} Compute the prototypes $\mathbf{g}_i$ of cluster $P_i$ ($i = 1, \ldots, c$) using Eq. (3)
(b) *Stage* 2:
{the membership degree $u_{ik}$ of pattern $k$ belonging to cluster $P_i$ and the prototypes $\mathbf{g}_i$ of the cluster $P_i$ ($i = 1, \ldots, c$) are fixed}
Compute the matrix $\mathbf{M}$ using Eq. (7)
(4) **Allocation step**:
{the prototypes $\mathbf{g}_i$ of cluster $P_i$ ($i = 1, \ldots, c$) and the matrix $\mathbf{M}$ are fixed} Update the fuzzy membership degree $u_{ik}$ of pattern $k$ belonging to cluster $P_i$ ($i = 1, \ldots, c$) using Eq. (8)
(5) **Stopping criterion**
**If** $|J2^{(t+1)} - J2^{(t)}| \leqslant \varepsilon$ **or** $t > T$
stop
**otherwise** $t = t + 1$ and go to step 3

*3.1.2. A fuzzy clustering method based on a single adaptive quadratic distance defined by a pooled diagonal fuzzy covariance matrix (FCM-PDFCV)*

In this section, we present a fuzzy clustering method for quantitative data based on a single adaptive quadratic distance defined by a pooled diagonal fuzzy covariance matrix. For this method, we define a criterion $J3$ measuring the fitting between clusters and their prototypes as

$$J3 = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m d^2_{\mathbf{M}}(\mathbf{x}_k, \mathbf{g}_i) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m (\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}}\mathbf{M}(\mathbf{x}_k - \mathbf{g}_i), \tag{9}$$

where $\mathbf{x}_k$, $\mathbf{g}_i$, $u_{ik}$ and $m$ are defined as before and

$$d^2_{\mathbf{M}}(\mathbf{x}_k, \mathbf{g}_i) = (\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}}\mathbf{M}(\mathbf{x}_k - \mathbf{g}_i) \quad \text{with } \mathbf{M} = \begin{pmatrix} \lambda^1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda^p \end{pmatrix} \tag{10}$$

is now a quadratic distance defined by a positive definite diagonal matrix $\mathbf{M}$ that is the same for all the clusters. Therefore, we can see that this method is a special case of the FCM-PFCV method presented in Section 3.1.1, in which we are restricting the pooled fuzzy covariance matrix to be diagonal.

In the first stage of the representation step, the prototypes $\mathbf{g}_i$ of cluster $P_i$ $(i = 1, \ldots, c)$, which minimizes the clustering criterion $J3$, are updated according to Eq. (3).

In the second stage, the positive definite diagonal matrix $\mathbf{M}$, which minimize the clustering criterion $J3$ under $\det(\mathbf{M}) = 1$, are updated according to the following expression:

$$\mathbf{M} = \begin{pmatrix} \lambda^1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda^p \end{pmatrix} = [\det(\mathbf{Q})]^{1/p}\,\mathbf{Q}^{-1} \quad \text{with } \mathbf{Q} = \sum_{i=1}^{c} \mathbf{C}_i, \tag{11}$$

where now $\mathbf{C}_i = \operatorname{diag}(\sum_{k=1}^{n}(u_{ik})^m(\mathbf{x}_k - \mathbf{g}_i)(\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}})$. Therefore, the diagonal elements of the matrix $\mathbf{M}$ are

$$\lambda^j = \frac{\{\prod_{h=1}^{p}[\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m(x_k^h - g_i^h)^2]\}^{1/p}}{\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m(x_k^j - g_i^j)^2} \quad (j = 1, \ldots, p).$$

In the allocation step, the membership degree $u_{ik}$ $(k = 1, \ldots, n)$ of each pattern $k$ in each cluster $P_i$ $(i = 1, \ldots, c)$, minimizing the clustering criterion $J3$ under $u_{ik} \geqslant 0$ and $\sum_{i=1}^{c} u_{ik} = 1$, is updated according to the following expression:

$$u_{ik} = \left[\sum_{h=1}^{c}\left(\frac{d^2_{\mathbf{M}}(\mathbf{x}_k, \mathbf{g}_i)}{d^2_{\mathbf{M}}(\mathbf{x}_k, \mathbf{g}_h)}\right)^{1/(m-1)}\right]^{-1} = \left[\sum_{h=1}^{c}\left(\frac{\sum_{j=1}^{p}\lambda^j(x_k^j - g_i^j)^2}{\sum_{j=1}^{p}\lambda^j(x_k^j - g_h^j)^2}\right)^{1/(m-1)}\right]^{-1}. \tag{12}$$

### *3.2. Fuzzy clustering methods based on a different adaptive distance for each cluster*

In this section, we present fuzzy clustering methods for quantitative data based on a different adaptive quadratic distance for each cluster. The main idea of this method is to compare clusters and their prototypes using a different distance associated with each cluster that changes at each iteration, i.e., the distance is not determined absolutely and is different from one cluster to another. Again, the advantage of these adaptive distances is that the clustering algorithm is able to find clusters of different shapes and sizes [6–8,14].

This adaptive method looks for a fuzzy partition of a set of patterns in $c$ clusters $\{P_1, \ldots, P_c\}$, the corresponding $c$ prototypes $\{\mathbf{g}_1, \ldots, \mathbf{g}_c\}$ and an adaptive quadratic distance that is different for each cluster such that a criterion measuring the fitting between the clusters and their prototypes is locally minimized.

#### *3.2.1. The GK fuzzy clustering model*

The first and more general adaptive version of the FCM algorithm, based on a quadratic distance for each cluster defined by a fuzzy covariance matrix, was introduced by Gustafson and Kessel [14]. For this method, a criterion $J4$ measuring the fitting between clusters and their representatives is defined as

$$J4 = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m d^2_{\mathbf{M}_i}(\mathbf{x}_k, \mathbf{g}_i) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m(\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}}\mathbf{M}_i(\mathbf{x}_k - \mathbf{g}_i), \tag{13}$$

where $\mathbf{x}_k$, $\mathbf{g}_i$, $u_{ik}$ and $m$ are defined as before and

$$d^2_{\mathbf{M}_i}(\mathbf{x}_k, \mathbf{g}_i) = (\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}}\mathbf{M}_i(\mathbf{x}_k - \mathbf{g}_i) \tag{14}$$

is now a quadratic distance defined by a positive definite symmetric matrix $\mathbf{M}_i$ associated with the cluster $i$ $(i = 1, \ldots, c)$.

The algorithm starts from an initial membership degree for each pattern $k$ in each cluster $P_i$ and alternates a representation step and an allocation step until the convergence of the algorithm, when the criterion $J4$ reaches a stationary value representing a local minimum.

The representation step also has two stages. In the first stage, the membership degree of each pattern $k$ in each cluster $P_i$ and the matrices $\mathbf{M}_i$ are fixed. The prototypes $\mathbf{g}_i$ of cluster $P_i$ $(i = 1, \ldots, c)$, which minimizes the clustering criterion $J4$, are updated according to Eq. (3) [14].

In the second stage, the membership degree of each pattern $k$ in cluster $P_i$ and the prototypes $\mathbf{g}_i$ of the corresponding clusters $P_i$ $(i = 1, \ldots, c)$ are fixed. The positive definite symmetric matrix $\mathbf{M}_i$ $(i = 1, \ldots, c)$, which minimizes the clustering criterion $J4$ under $\det(\mathbf{M}_i) = 1$, is updated according to the following expression [14]:

$$\mathbf{M}_i = [\det(\mathbf{C}_i)]^{1/p}\, \mathbf{C}_i^{-1} \quad \text{with } \mathbf{C}_i = \sum_{k=1}^{n} (u_{ik})^m (\mathbf{x}_k - \mathbf{g}_i)(\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}}. \tag{15}$$

In the allocation step, the prototypes $\mathbf{g}_i$ of corresponding clusters $P_i$ $(i = 1, \ldots, c)$ and the matrices $\mathbf{M}_i$ $(i = 1, \ldots, c)$ are fixed. The membership degree $u_{ik}$ $(k = 1, \ldots, n)$ of each pattern $k$ in each cluster $P_i$ $(i = 1, \ldots, c)$, minimizing the clustering criterion $J4$ under $u_{ik} \geqslant 0$ and $\sum_{i=1}^{c} u_{ik} = 1$, is updated according to the following expression [14]:

$$u_{ik} = \left[ \sum_{h=1}^{c} \left( \frac{d^2_{\mathbf{M}_i}(\mathbf{x}_k, \mathbf{g}_i)}{d^2_{\mathbf{M}_h}(\mathbf{x}_k, \mathbf{g}_h)} \right)^{1/(m-1)} \right]^{-1} = \left[ \sum_{h=1}^{c} \left( \frac{(\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}} \mathbf{M}_i (\mathbf{x}_k - \mathbf{g}_i)}{(\mathbf{x}_k - \mathbf{g}_h)^{\mathrm{T}} \mathbf{M}_h (\mathbf{x}_k - \mathbf{g}_h)} \right)^{1/(m-1)} \right]^{-1}. \tag{16}$$

*3.2.2. A fuzzy clustering method based on an adaptive quadratic distance for each cluster defined by a diagonal fuzzy covariance matrix (FCM-DFCV)*

Here we present a fuzzy clustering method for quantitative data based on an adaptive quadratic distance for each cluster defined by a diagonal fuzzy covariance matrix. For this model, a criterion $J5$ measuring the fitting between the clusters and their prototypes is defined as

$$J5 = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m d^2_{\mathbf{M}_i}(\mathbf{x}_k, \mathbf{g}_i) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m (\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}} \mathbf{M}_i (\mathbf{x}_k - \mathbf{g}_i), \tag{17}$$

where $\mathbf{x}_k$, $\mathbf{g}_i$, $u_{ik}$ and $m$ are defined as before and

$$d^2_{\mathbf{M}_i}(\mathbf{x}_k, \mathbf{g}_i) = (\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}} \mathbf{M}_i (\mathbf{x}_k - \mathbf{g}_i) \quad \text{with } \mathbf{M}_i = \begin{pmatrix} \lambda_i^1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_i^p \end{pmatrix} \tag{18}$$

is now a quadratic distance defined by a positive definite diagonal matrix $\mathbf{M}_i$ associated with each cluster $i = 1, \ldots, c$. Therefore, we can see that this model is a special case of the model introduced by Gustafson and Kessel [14], in which we are restricting the fuzzy covariance matrix to be diagonal.

In the first stage of the representation step, the prototypes $\mathbf{g}_i$ of cluster $P_i$ $(i = 1, \ldots, c)$, which minimizes the clustering criterion $J5$, are updated according to Eq. (3).

In the second stage, the positive definite diagonal matrices $\mathbf{M}_i$ $(i = 1, \ldots, c)$ which minimize the clustering criterion $J5$ under $\det(\mathbf{M}_i) = 1$, are updated according to the following expression:

$$\mathbf{M}_i = \begin{pmatrix} \lambda_i^1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_i^p \end{pmatrix} = [\det(\mathbf{C}_i)]^{1/p} \mathbf{C}_i^{-1}, \quad \mathbf{C}_i = \operatorname{diag}\left( \sum_{k=1}^{n} (u_{ik})^m (\mathbf{x}_k - \mathbf{g}_i)(\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}} \right). \tag{19}$$

Therefore, the diagonal elements of the matrices $\mathbf{M}_i$ are

$$\lambda_i^j = \frac{\{\prod_{h=1}^{p}[\sum_{k=1}^{n}(u_{ik})^m(x_k^h - g_i^h)^2]\}^{1/p}}{\sum_{k=1}^{n}(u_{ik})^m(x_k^j - g_i^j)^2} \quad (j = 1, \ldots, p).$$

In the allocation step, the membership degree $u_{ik}$ $(k = 1, \ldots, n)$ of each pattern $k$ in each cluster $P_i$ $(i = 1, \ldots, c)$, minimizing the clustering criterion $J5$ under $u_{ik} \geqslant 0$ and $\sum_{i=1}^{c} u_{ik} = 1$, is updated according to the following expression:

$$u_{ik} = \left[\sum_{h=1}^{c}\left(\frac{d^2_{\mathbf{M}_i}(\mathbf{x}_k, \mathbf{g}_i)}{d^2_{\mathbf{M}_h}(\mathbf{x}_k, \mathbf{g}_h)}\right)^{1/(m-1)}\right]^{-1} = \left[\sum_{h=1}^{c}\left(\frac{\sum_{j=1}^{p}\lambda_i^j(x_k^j - g_i^j)^2}{\sum_{j=1}^{p}\lambda_h^j(x_k^j - g_h^j)^2}\right)^{1/(m-1)}\right]^{-1}. \tag{20}$$

## 4. Cluster interpretation

Once a clustering of a data set has been obtained, it is necessary to interpret the resulting clusters. The user wants to evaluate the overall data heterogeneity, intra-cluster and between-cluster heterogeneity, the contribution of each variable toward cluster formation, etc.

For quantitative data partitioned by the standard hard $c$-means clustering algorithm, Celeux et al. [6] have introduced a family of indices for cluster interpretation that are based on the SSQs.

In this section, we adapt these indices for quantitative data partitioned by the fuzzy clustering methods presented in this paper.

### *4.1. Interpretation indices for clusters furnished by the standard hard c-means clustering algorithm*

Let a crisp partition $\{P_1, \ldots, P_c\}$ of $\Omega = \{1, \ldots, n\}$ into $c$ clusters of cardinality $n_i$ $(i = 1, \ldots, c)$ obtained from the standard hard $c$-means clustering algorithm and denote by $\mathbf{g}_i = (g_i^1, \ldots, g_i^p\}$, with

$$\mathbf{g}_i = \sum_{k \in P_i} \mathbf{x}_k / n_i, \quad g_i^j = \sum_{k \in P_i} x_k^j / n_i \ \ (j = 1, \ldots, p)$$

the prototype of cluster $P_i$.

Moreover, let

$$\mathbf{z} = (z^1, \ldots, z^p) = \sum_{k=1}^{n} \mathbf{x}_k / n = \sum_{i=1}^{c} n_i \mathbf{g}_i / \sum_{i=1}^{c} n_i$$

with

$$z^j = \frac{\sum_{k=1}^{n} x_k^j}{n} = \frac{\sum_{i=1}^{c}\sum_{k \in P_i} x_k^j}{\sum_{i=1}^{c} n_i} = \frac{\sum_{i=1}^{c}(\sum_{k \in P_i} x_k^j)}{\sum_{i=1}^{c} n_i} = \frac{\sum_{i=1}^{c} n_i g_i^j}{\sum_{i=1}^{c} n_i}$$

$(j = 1, \ldots, p)$ be the overall prototype vector for all $n$ data points.

Three SSQs are defined for this partition:

- the overall SSQ: $T = \sum_{i=1}^{c}\sum_{k \in P_i}\sum_{j=1}^{p}(x_k^j - z^j)^2$;
- the SSQ within clusters: $J = \sum_{i=1}^{c}\sum_{k \in P_i}\sum_{j=1}^{p}(x_k^j - g_i^j)^2$;
- and the SSQ between clusters: $B = \sum_{i=1}^{c} n_i \sum_{j=1}^{p}(g_i^j - z^j)^2$.

The overall SSQ decomposes into the sum of the cluster-specific overall SSQs given by $T = \sum_{i=1}^{c} T_i$ with $T_i = \sum_{k \in P_i}\sum_{j=1}^{p}(x_k^j - z^j)^2$, into the sum of the variable-specific overall SSQ given by $T = \sum_{j=1}^{p} T_j$ with $T_j = \sum_{i=1}^{c}\sum_{k \in P_i}(x_k^j - z^j)^2$ and into the sum of variable–cluster-specific overall SSQ given by $T = \sum_{i=1}^{c} T_i = \sum_{i=1}^{c}(\sum_{j=1}^{p} T_{ij})$ with $T_{ij} = \sum_{k \in P_i}(x_k^j - z^j)^2$.

Similarly, the SSQ within clusters decomposes into the sum of the cluster-specific SSQs within clusters given by $J = \sum_{i=1}^{c} J_i$ with $J_i = \sum_{k\in P_i} \sum_{j=1}^{p}(x_k^j - g_i^j)^2$, into the sum of the variable-specific SSQ within clusters given by $J = \sum_{j=1}^{p} J_j$ with $J_j = \sum_{i=1}^{c} \sum_{k\in P_i}(x_k^j - g_i^j)^2$ and into the sum of variable–cluster-specific SSQ within clusters given by $J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c}(\sum_{j=1}^{p} J_{ij})$ with $J_{ij} = \sum_{k\in P_i}(x_k^j - g_i^j)^2$.

In the same way, the SSQ between clusters decomposes into the sum of the cluster-specific SSQs between clusters given by $B = \sum_{i=1}^{c} B_i$ with $B_i = n_i \sum_{j=1}(g_i^j - z^j)^2$, into the sum of the variable-specific SSQ between clusters given by $B = \sum_{j=1}^{p} B_j$ with $B_j = \sum_{i=1}^{c} n_i(g_i^j - z^j)^2$ and into the sum of variable–cluster-specific SSQ within clusters given by $B = \sum_{i=1}^{c} B_i = \sum_{i=1}^{c} \sum_{j=1}^{p} B_{ij}$ with $B_{ij} = n_i(g_i^j - z^j)^2$.

It is easily verified that the overall SSQs decomposes into the SSQs within clusters plus the SSQs between clusters according to the following expressions: $T = J + B$, $T_i = J_i + B_i$, $T_j = J_j + B_j$ and $T_{ij} = J_{ij} + B_{ij}$. The cluster interpretation indices introduced by Celeux et al. [6] for the hard $c$-means clustering algorithm are based on these expressions. For example, a general index, that measures the ratio between the dispersion of the partition after clustering $B$ (each cluster represented by its prototype) and the overall dispersion without clustering $T$, is defined as

$$R = \frac{B}{T} = \frac{\sum_{i=1}^{c} n_i \sum_{j=1}^{p}(g_i^j - z^j)^2}{\sum_{i=1}^{c} \sum_{k\in P_i} \sum_{j=1}^{p}(x_k^j - z^j)^2}.$$

A greater value of $R$ corresponds to more homogeneous the clusters. Other indices measuring cluster heterogeneity are also proposed by Celeux et al. [6, pp. 154–164].

### 4.2. Interpretation indices for clusters furnished by fuzzy clustering algorithms

We consider a fuzzy partition $\{P_1, \ldots, P_c\}$ of $\Omega = \{1, \ldots, n\}$ in $c$ clusters obtained from one of the methods presented in Sections 2 and 3 and denote by $\mathbf{g}_i = (g_i^1, \ldots, g_i^p\}$ with

$$\mathbf{g}_i = \frac{\sum_{k=1}^{n}(u_{ik})^m \mathbf{x}_k}{\sum_{k=1}^{n}(u_{ik})^m}, \quad g_i^j = \frac{\sum_{k=1}^{n}(u_{ik})^m x_k^j}{\sum_{k=1}^{n}(u_{ik})^m} \ \ (j = 1, \ldots, p)$$

the prototype of cluster $P_i$.

Moreover, the vector

$$\mathbf{z} = (z^1, \ldots, z^p) = \frac{\sum_{i=1}^{c} \sum_{k=1}^{n}(u_{ik})^m \mathbf{x}_k}{\sum_{i=1}^{c} \sum_{k=1}^{n}(u_{ik})^m} = \frac{\sum_{i=1}^{c} \mu_i \mathbf{g}_i}{\sum_{i=1}^{c} \mu_i}, \quad \mu_i = \sum_{k=1}^{n}(u_{ik})^m \ \ (i = 1, \ldots, c)$$

with

$$z^j = \frac{\sum_{i=1}^{c} \sum_{k=1}^{n}(u_{ik})^m x_k^j}{\sum_{i=1}^{c} \sum_{k=1}^{n}(u_{ik})^m} = \frac{\sum_{i=1}^{c}(\sum_{k=1}^{n}(u_{ik})^m x_k^j)}{\sum_{i=1}^{c} \sum_{k=1}^{n}(u_{ik})^m} = \frac{\sum_{i=1}^{c} \mu_i g_i^j}{\sum_{i=1}^{c} \mu_i} \quad (j = 1, \ldots, p)$$

is the overall prototype vector for all $n$ data points.

#### 4.2.1. Measures based on the SSQs

In this section, we define the overall SSQ and the SSQ within and between the clusters for quantitative data in the framework of the presented fuzzy clustering methods, which are the basis for defining the interpretation tools in Section 4.2.2.

*4.2.1.1. Overall SSQs.* The overall heterogeneity of all $n$ data points is measured by the overall SSQs according to the distance function used:

$$T1 = \sum_{i=1}^{c} \sum_{k=1}^{n}(u_{ik})^m \phi(\mathbf{x}_k, \mathbf{z}) = \sum_{i=1}^{c} \sum_{k=1}^{n}(u_{ik})^m \sum_{j=1}^{p}(x_k^j - z^j)^2,$$

$$T2 = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{z}) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m(\mathbf{x}_k-\mathbf{z})^{\mathrm{T}}\mathbf{M}(\mathbf{x}_k-\mathbf{z}),$$

$$T3 = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m \psi(\mathbf{x}_k,\mathbf{z}) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m \sum_{j=1}^{p}\lambda^j(x_k^j - z^j)^2,$$

$$T4 = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m d^2_{\mathbf{M}_i}(\mathbf{x}_k,\mathbf{z}) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m(\mathbf{x}_k-\mathbf{z})^{\mathrm{T}}\mathbf{M}_i(\mathbf{x}_k-\mathbf{z}),$$

$$T5 = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m \psi_i(\mathbf{x}_k,\mathbf{z}) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m \sum_{j=1}^{p}\lambda_i^j(x_k^j - z^j)^2 \tag{21}$$

$Tl\,(l = 1, 2, 3, 4, 5)$ decomposes, on the one hand, into the sum of the cluster-specific SSQs in the clusters $P_i$ given by $Tl = \sum_{i=1}^{c} T_i^l$ $(l = 1, 2, 3, 4, 5)$ with

$$T_i^1 = \sum_{k=1}^{n}(u_{ik})^m \phi(\mathbf{x}_k,\mathbf{z}) = \sum_{k=1}^{n}(u_{ik})^m \sum_{j=1}^{p}(x_k^j - z^j)^2,$$

$$T_i^2 = \sum_{k=1}^{n}(u_{ik})^m d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{z}) = \sum_{k=1}^{n}(u_{ik})^m(\mathbf{x}_k-\mathbf{z})^{\mathrm{T}}\mathbf{M}(\mathbf{x}_k-\mathbf{z}),$$

$$T_i^3 = \sum_{k=1}^{n}(u_{ik})^m \psi(\mathbf{x}_k,\mathbf{z}) = \sum_{k=1}^{n}(u_{ik})^m \sum_{j=1}^{p}\lambda^j(x_k^j - z^j)^2,$$

$$T_i^4 = \sum_{k=1}^{n}(u_{ik})^m d^2_{\mathbf{M}_i}(\mathbf{x}_k,\mathbf{z}) = \sum_{k=1}^{n}(u_{ik})^m(\mathbf{x}_k-\mathbf{z})^{\mathrm{T}}\mathbf{M}_i(\mathbf{x}_k-\mathbf{z}),$$

$$T_i^5 = \sum_{k=1}^{n}(u_{ik})^m \psi_i(\mathbf{x}_k,\mathbf{z}) = \sum_{k=1}^{n}(u_{ik})^m \sum_{j=1}^{p}\lambda_i^j(x_k^j - z^j)^2 \tag{22}$$

and, in the case where the matrix that defines the quadratic distance is diagonal, on the other hand, it decomposes into the sum of the variable-specific overall SSQs for the variables $j = 1, \ldots, p$, given by $Tl = \sum_{j=1}^{p} T_j^l$ $(l = 1, 3, 5)$, with

$$T_j^1 = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m (x_k^j - z^j)^2,$$

$$T_j^3 = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m \lambda^j(x_k^j - z^j)^2,$$

$$T_j^5 = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m \lambda_i^j(x_k^j - z^j)^2. \tag{23}$$

In both cases (when the matrix that defines the quadratic distance is diagonal), $Tl = \sum_{i=1}^{c} T_i^l = \sum_{i=1}^{c}(\sum_{j=1}^{p} T_{ij}^l)$ $(l = 1, 3, 5)$ and

$$T_{ij}^1 = \sum_{k=1}^{n}(u_{ik})^m (x_k^j - z^j)^2,$$

$$T_{ij}^3 = \sum_{k=1}^{n} (u_{ik})^m \lambda^j (x_k^j - z^j)^2,$$

$$T_{ij}^5 = \sum_{k=1}^{n} (u_{ik})^m \lambda_i^j (x_k^j - z^j)^2 \tag{24}$$

denotes the partial SSQ in the cluster $P_i$ relating to the $j$th variable ($j = 1, \ldots, p$; $i = 1, \ldots, c$).

*4.2.1.2. Within-cluster SSQs.* Here we consider the heterogeneity within the clusters $P_i$ and measure it by the within-cluster SSQ:

$$J_i^1 = \sum_{k=1}^{n} (u_{ik})^m \phi(\mathbf{x}_k, \mathbf{g}_i) = \sum_{k=1}^{n} (u_{ik})^m \sum_{j=1}^{p} (x_k^j - g_i^j)^2,$$

$$J_i^2 = \sum_{k=1}^{n} (u_{ik})^m d_{\mathbf{M}}^2(\mathbf{x}_k, \mathbf{g}_i) = \sum_{k=1}^{n} (u_{ik})^m (\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}} \mathbf{M} (\mathbf{x}_k - \mathbf{g}_i),$$

$$J_i^3 = \sum_{k=1}^{n} (u_{ik})^m \psi(\mathbf{x}_k, \mathbf{g}_i) = \sum_{k=1}^{n} (u_{ik})^m \sum_{j=1}^{p} \lambda^j (x_k^j - g_i^j)^2,$$

$$J_i^4 = \sum_{k=1}^{n} (u_{ik})^m d_{\mathbf{M}_i}^2(\mathbf{x}_k, \mathbf{g}_i) = \sum_{k=1}^{n} (u_{ik})^m (\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}} \mathbf{M}_i (\mathbf{x}_k - \mathbf{g}_i),$$

$$J_i^5 = \sum_{k=1}^{n} (u_{ik})^m \psi_i(\mathbf{x}_k, \mathbf{g}_i) = \sum_{k=1}^{n} (u_{ik})^m \sum_{j=1}^{p} \lambda_i^j (x_k^j - g_i^j)^2. \tag{25}$$

Summing up over all clusters, we obtain the overall within-cluster SSQ $Jl = \sum_{i=1}^{c} J_i^l$ ($l = 1, 2, 3, 4, 5$), i.e.,

$$J1 = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \phi(\mathbf{x}_k, \mathbf{g}_i) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \sum_{j=1}^{p} (x_k^j - g_i^j)^2,$$

$$J2 = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m d_{\mathbf{M}}^2(\mathbf{x}_k, \mathbf{g}_i) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m (\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}} \mathbf{M} (\mathbf{x}_k - \mathbf{g}_i),$$

$$J3 = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \psi(\mathbf{x}_k, \mathbf{g}_i) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \sum_{j=1}^{p} \lambda^j (x_k^j - g_i^j)^2,$$

$$J4 = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m d_{\mathbf{M}_i}^2(\mathbf{x}_k, \mathbf{g}_i) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m (\mathbf{x}_k - \mathbf{g}_i)^{\mathrm{T}} \mathbf{M}_i (\mathbf{x}_k - \mathbf{g}_i),$$

$$J5 = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \psi_i(\mathbf{x}_k, \mathbf{g}_i) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \sum_{j=1}^{p} \lambda_i^j (x_k^j - g_i^j)^2. \tag{26}$$

On the other hand, for the case where the matrix that defines the quadratic distance is diagonal, $Jl$ ($l = 1, 3, 5$) decomposes into the sum of the variable-specific overall SSQs for the variables $j = 1, \ldots, p$, given by $Jl = \sum_{j=1}^{p} J_j^l$ ($l = 1, 3, 5$), with

$$J_j^1 = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m (x_k^j - g_i^j)^2,$$

$$J_j^3 = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m \lambda^j (x_k^j - g_i^j)^2,$$

$$J_j^5 = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m \lambda_i^j (x_k^j - g_i^j)^2. \tag{27}$$

Notice that the fuzzy clustering methods FCM-PDFCV and FCM-DFCV, which are based on adaptive quadratic distances defined by positive definite diagonal matrices, impose the same variable-specific within-cluster SSQs $J_j^l$ ($l = 3, 5$) for all variables $j$. Indeed,

$$J_j^3 = \lambda^j \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m (x_k^j - g_i^j)^2 = \left\{\prod_{h=1}^{p}\left[\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m (x_k^h - g_i^h)^2\right]\right\}^{1/p}$$

and

$$J_j^5 = \sum_{i=1}^{c}\lambda_i^j \sum_{k=1}^{n}(u_{ik})^m (x_k^j - g_i^j)^2 = \sum_{i=1}^{c}\left(\left\{\prod_{h=1}^{p}\left[\sum_{k=1}^{n}(u_{ik})^m (x_k^h - g_i^h)^2\right]\right\}^{1/p}\right)$$

and they do not depend on variable $j$.

In both cases (when the matrix that defines the quadratic distance is diagonal), $Jl = \sum_{i=1}^{c} J_i^l = \sum_{i=1}^{c}(\sum_{j=1}^{p} J_{ij}^l)$ ($l = 1, 3, 5$) and

$$J_{ij}^1 = \sum_{k=1}^{n}(u_{ik})^m (x_k^j - g_i^j)^2,$$

$$J_{ij}^3 = \sum_{k=1}^{n}(u_{ik})^m \lambda^j (x_k^j - g_i^j)^2,$$

$$J_{ij}^5 = \sum_{k=1}^{n}(u_{ik})^m \lambda_i^j (x_k^j - g_i^j)^2 \tag{28}$$

denotes the within-cluster SSQ $J_{ij}^l$ ($l = 1, 3, 5$) of the variable $j$ in cluster $P_i$ ($j = 1, \ldots, p$; $i = 1, \ldots, c$).

Here also notice that, for a given cluster $P_i$, the fuzzy clustering method FCM-DFCV, based on an adaptive quadratic distance for each cluster defined by a positive definite diagonal matrix, imposes the same within-cluster SSQ $J_{ij}^5$ for all the variables $j$. Indeed,

$$J_{ij}^5 = \lambda_i^j \sum_{k=1}^{n}(u_{ik})^m (x_k^j - g_i^j)^2 = \left\{\prod_{h=1}^{p}\left[\sum_{k=1}^{n}(u_{ik})^m (x_k^h - g_i^h)^2\right]\right\}^{1/p}$$

which does not depend on variable $j$.

*4.2.1.3. Between-cluster SSQs.* The between-cluster SSQ given by

$$B1 = \sum_{i=1}^{c}\mu_i \phi(\mathbf{g}_i, \mathbf{z}) = \sum_{i=1}^{c}\mu_i \sum_{j=1}^{p}(g_i^j - z^j)^2,$$

$$B2 = \sum_{i=1}^{c}\mu_i d_{\mathbf{M}}^2(\mathbf{g}_i, \mathbf{z}) = \sum_{i=1}^{c}\mu_i (\mathbf{g}_i - \mathbf{z})^{\mathrm{T}}\mathbf{M}(\mathbf{g}_i - \mathbf{z}),$$

$$B3 = \sum_{i=1}^{c}\mu_i \psi(\mathbf{g}_i, \mathbf{z}) = \sum_{i=1}^{c}\mu_i \sum_{j=1}^{p}\lambda^j (g_i^j - z^j)^2,$$

$$B4 = \sum_{i=1}^{c} \mu_i d^2_{\mathbf{M}_i}(\mathbf{g}_i, \mathbf{z}) = \sum_{i=1}^{c} \mu_i (\mathbf{g}_i - \mathbf{z})^{\mathrm{T}} \mathbf{M}_i (\mathbf{g}_i - \mathbf{z}),$$

$$B5 = \sum_{i=1}^{c} \mu_i \psi_i(\mathbf{g}_i, \mathbf{z}) = \sum_{i=1}^{c} \mu_i \sum_{j=1}^{p} \lambda_i^j (g_i^j - z^j)^2 \tag{29}$$

measures the dispersion of the cluster prototypes and, consequently, the distinctness of all clusters. The factor $\mu_i = \sum_{k=1}^{n} (u_{ik})^m$ takes into account the membership degree of each pattern $k$ in cluster $P_i$ (it is the equivalent to the number of individuals $n_i$ belonging to cluster $P_i$ in a hard partition that is used to evaluate the SSQ between clusters). It is decomposed either into the sum of all $c$ cluster-specific SSQs $Bl = \sum_{i=1}^{c} B_i^l$ $(l = 1, 2, 3, 4, 5)$ with

$$B_i^1 = \mu_i \phi(\mathbf{g}_i, \mathbf{z}) = \mu_i \sum_{j=1}^{p} (g_i^j - z^j)^2,$$

$$B_i^2 = \mu_i d^2_{\mathbf{M}}(\mathbf{g}_i, \mathbf{z}) = \mu_i (\mathbf{g}_i - \mathbf{z})^{\mathrm{T}} \mathbf{M} (\mathbf{g}_i - \mathbf{z}),$$

$$B_i^3 = \mu_i \psi(\mathbf{g}_i, \mathbf{z}) = \mu_i \sum_{j=1}^{p} \lambda^j (g_i^j - z^j)^2,$$

$$B_i^4 = \mu_i d^2_{\mathbf{M}_i}(\mathbf{g}_i, \mathbf{z}) = \mu_i (\mathbf{g}_i - \mathbf{z})^{\mathrm{T}} \mathbf{M}_i (\mathbf{g}_i - \mathbf{z}),$$

$$B_i^5 = \mu_i \psi_i(\mathbf{g}_i, \mathbf{z}) = \mu_i \sum_{j=1}^{p} \lambda_i^j (g_i^j - z^j)^2, \tag{30}$$

which measure the heterogeneity in the clusters $P_i$ or (when the matrix that defines the quadratic distance is diagonal) into the sum of the $p$ variable-specific between-cluster SSQs $Bl = \sum_{j=1}^{p} B_j^l$ $(l = 1, 3, 5)$ as given by

$$B_j^1 = \sum_{i=1}^{c} B_{ij}^1 = \sum_{i=1}^{c} \mu_i (g_i^j - z^j)^2,$$

$$B_j^3 = \sum_{i=1}^{c} B_{ij}^3 = \sum_{i=1}^{c} \mu_i \lambda^j (g_i^j - z^j)^2,$$

$$B_j^5 = \sum_{i=1}^{c} B_{ij}^5 = \sum_{i=1}^{c} \mu_i \lambda_i^j (g_i^j - z^j)^2. \tag{31}$$

In all formulas

$$B_{ij}^1 = \mu_i (g_i^j - z^j)^2,$$

$$B_{ij}^2 = \mu_i \lambda^j (g_i^j - z^j)^2,$$

$$B_{ij}^3 = \mu_i \lambda_i^j (g_i^j - z^j)^2 \tag{32}$$

measures the dissimilarity (for the variable $j$) between the cluster prototype $\mathbf{g}_i$ of $P_i$ and the overall prototype $\mathbf{z}$ of all data.

**Proposition 4.1.** *The following relations hold*:

$$Tl = Jl + Bl \quad (l = 1, 2, 3, 4, 5),$$

$$T_i^l = B_i^l + J_i^l \quad (i = 1, \ldots, c \text{ and } l = 1, 2, 3, 4, 5),$$

$$T_j^l = B_j^l + J_j^l \quad (j = 1, \ldots, p \text{ and } l = 1, 3, 5),$$

$$T_{ij}^l = B_{ij}^l + J_{ij}^l \quad (i = 1, \ldots, c,\ j = 1, \ldots, p \text{ and } l = 1, 3, 5).$$

**Proof.** We will start showing that $T1 = J1 + B1$ holds. We have $(x_k^j - z^j)^2 = [(x_k^j - g_i^j) + (g_i^j - z^j)]^2 = (x_k^j - g_i^j)^2 + (g_i^j - z^j)^2 + 2(x_k^j - g_i^j)(g_i^j - z^j)$. Then, from Eq. (21), $T1 = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \sum_{j=1}^{p} (x_k^j - z^j)^2 = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \sum_{j=1}^{p} (x_k^j - g_i^j)^2 + \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \sum_{j=1}^{p} (g_i^j - z^j)^2 + 2 \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \sum_{j=1}^{p} (x_k^j - g_i^j)(g_i^j - z^j) = J1 + B1 + 2 \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \sum_{j=1}^{p} (x_k^j - g_i^j)(g_i^j - z^j)$.

We have also: $(x_k^j - g_i^j)(g_i^j - z^j) = g_i^j (x_k^j - g_i^j) - z^j (x_k^j - g_i^j)$. Then, $\sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \sum_{j=1}^{p} (x_k^j - g_i^j)(g_i^j - z^j) = \sum_{j=1}^{p} \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m [g_i^j (x_k^j - g_i^j) - z^j (x_k^j - g_i^j)] = \sum_{j=1}^{p} \sum_{i=1}^{c} \{g_i^k [\sum_{k=1}^{n} (u_{ik})^m x_k^j - g_i^j \sum_{k=1}^{n} (u_{ik})^m] - z^j [\sum_{k=1}^{n} (u_{ik})^m x_k^j - g_i^j \sum_{k=1}^{n} (u_{ik})^m]\}$.

As $g_i^j \sum_{k=1}^{n} (u_{ik})^m = \sum_{k=1}^{n} (u_{ik})^m x_k^j$, it follows that $\sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \sum_{j=1}^{p} (x_k^j - g_i^j)(g_i^j - z^j) = 0$ and then $T1 = J1 + B1$. The other expressions can be easily obtained in a similar way. □

*4.2.2. Interpretation indices*

The fuzzy clustering interpretation indices presented in this section are a suitable adaptation of the indices presented in [6] for the case of the standard hard $c$-means clustering algorithm in order to evaluate the quality of the partition, the homogeneity and eccentricity of the individual clusters, and also the role of the different variables in the cluster formation process. All these indices range between 0 and 1.

*4.2.2.1. Fuzzy partition interpretation indices.* Interpreting the overall quality of a partition after having applied a fuzzy clustering algorithm to the data is an important problem in clustering analysis.

*Overall heterogeneity index*: The ratio between the dispersion of the partition after clustering $Bl$ ($l = 1, 2, 3, 4, 5$), each cluster represented by its prototype, and the overall dispersion without clustering $Tl$ ($l = 1, 2, 3, 4, 5$), is defined as

$$Rl = \frac{Bl}{Tl} = \frac{Bl}{Bl + Jl} = 1 - \frac{Jl}{Tl} \quad (l = 1, 2, 3, 4, 5). \tag{33}$$

The FCM clustering algorithms are designed so as to (locally) minimize the clustering criterion $Jl$ ($l = 1, 2, 3, 4, 5$). As $Tl = Jl + Bl$, minimizes $Jl$ means that they are designed so as to (locally) maximize $Rl = 1 - Jl/Tl$. A greater value of $Rl$ leads more homogeneous clusters and a better representation of the elements of a cluster $P_i$ by its prototype $\mathbf{g}_i$ ($i = 1, \ldots, c$).

*Overall heterogeneity indices with respect to single variables*: For the case where the matrix that defines the quadratic distance is diagonal, the proportion of the overall dispersion without clustering ($T_j^l,\ l = 1, 3, 5$) concerning the $j$th variable that corresponds to the dispersion of the partition after clustering concerning the $j$th variable, each cluster represented by its prototype ($B_j^l,\ l = 1, 3, 5$), is defined as

$$COR^l(j) = \frac{B_j^l}{T_j^l} = \frac{B_j^l}{B_j^l + J_j^l} \quad (l = 1, 3, 5). \tag{34}$$

By comparing the value of $COR^l(j)$ ($l = 1, 3, 5$) with the value of the general index $Rl$ ($l = 1, 3, 5$), which measures the average discriminant power of all variables, may be evaluated if the discriminant power of the $j$th variable is above or below the average.

The relative contribution of the $j$th variable to the between-cluster SSQs $B$ is given by

$$CTR^l(j) = \frac{B_j^l}{Bl} \quad (l = 1, 3, 5). \tag{35}$$

Notice that $\sum_{j=1}^{p} CTR^l(j) = 1$. A high value of $CTR^l(j)$ ($l = 1, 3, 5$) indicates that the $j$th variable provides an important contribution to the separation of the prototypes of the clusters. An interesting case arises when $COR^l(j)$ ($l =$

1, 3, 5) has a low value and $CTR^l(j)$ $(l = 1, 3, 5)$ is large: this means that the $j$th variable has a low discriminant power, although it makes an important contribution to the SSQs [6].

*4.2.2.2. Cluster interpretation indices.* Another important problem in clustering analysis is evaluating the homogeneity and eccentricity of the individual clusters of a partition after having applied a fuzzy clustering algorithm to the data.

*Cluster heterogeneity indices*: The proportion of the overall SSQs "in cluster $P_i$" is given by

$$T^l(i) := \frac{T_i^l}{Tl} \quad (l = 1, 2, 3, 4, 5). \tag{36}$$

The relative contribution of a cluster $P_i$ to the between-cluster SSQs is measured by the ratio

$$B^l(i) := \frac{B_i^l}{Bl} \quad (l = 1, 2, 3, 4, 5). \tag{37}$$

A high value of $B^l(i)$ $(l = 1, 2, 3, 4, 5)$ indicates that cluster $P_i$ is quite distant from the global centre in comparison to the totality of all clusters.

The relative contribution of cluster $P_i$ to the within-cluster SSQs is given by

$$J^l(i) := \frac{J_i^l}{Jl} \quad (l = 1, 2, 3, 4, 5). \tag{38}$$

A relatively large value of $J^l(i)$ indicates that cluster $P_i$ is relatively heterogeneous in comparison with the other clusters.

Notice that $\sum_{i=1}^{c} T^l(i) = \sum_{i=1}^{c} B^l(i) = \sum_{i=1}^{c} J^l(i) = 1$.

*Cluster heterogeneity indices with respect to single variables*: Again, for the fuzzy clustering algorithms in which the matrix that defines the quadratic distance is diagonal, the heterogeneity of clusters may be different for distinct variables. This can be evaluated by considering the previously proposed indices for a single variable $j$ alone.

The proportion of the discriminant power of the $j$th variable with respect to cluster $P_i$ is given by

$$COR^l(j, i) = \frac{B_{ij}^l}{T_j^l} \quad (l = 1, 3, 5). \tag{39}$$

Notice that $\sum_{i=1}^{c} COR^l(j, i) = COR(j)$. A high value of $COR^l(j, i)$ $(l = 1, 3, 5)$ shows that the $j$th variable has a relatively homogeneous behaviour within the cluster $i$.

The relative contribution of the $j$th variable to the heterogenity in cluster $P_i$ is given by

$$CTR^l(j, i) = \frac{B_{ij}^l}{B_i^l} \quad (l = 1, 3, 5). \tag{40}$$

Finally, we may consider the relative contribution of the $j$th variable and cluster $P_i$ to the between-cluster sum of squares given by

$$CE^l(j, i) = \frac{B_{ij}^l}{Bl} \quad (l = 1, 3, 5). \tag{41}$$

If $CE^l(j, i)$ $(l = 1, 3, 5)$ is close to 1, the $j$th variable has a large contribution to the eccentricity of cluster $P_i$.

## 5. Experimental results

To show the usefulness of these methods, several data sets, available at the UCI Repository [4], and synthetic quantitative data sets in $\Re^2$ are considered.

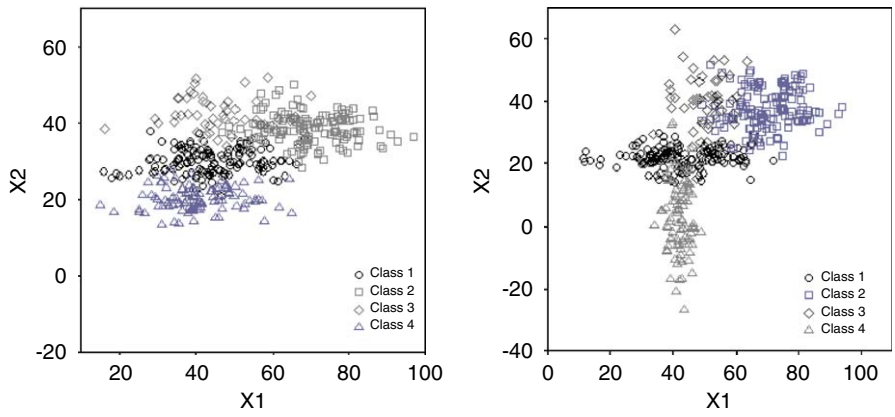


Fig. 1. Clusters drawn from data configurations 1 and 2.

### 5.1. Synthetic quantitative data sets

Each synthetic data set was created having classes with different sizes and shapes. The synthetic data sets have 450 points each, divided into four classes of unequal sizes: two classes of size 150 each, one class with 50 and one with 100. Each class in these data were drawn according to a bi-variate normal distribution with vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ represented by

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

We will consider four different data configurations: (1) the class covariance matrices are diagonal and almost the same; (2) the class covariance matrices are diagonal but unequal; (3) the class covariance matrices are not diagonal but almost the same; (4) the class covariance matrices are not diagonal and are also unequal. It is expected, for example, that the FCM-PDFCV clustering method will perform well if the data are drawn considering configuration 1.

Patterns of each class in data configuration 1 (Fig. 1, left side) were drawn according to the following parameters:

(a) Class 1: $\mu_1 = 45$, $\mu_2 = 30$, $\sigma_1^2 = 100$, $\sigma_2^2 = 9$, $\rho = 0.0$;

(b) Class 2: $\mu_1 = 70$, $\mu_2 = 38$, $\sigma_1^2 = 81$, $\sigma_2^2 = 16$, $\rho = 0.0$;

(c) Class 3: $\mu_1 = 45$, $\mu_2 = 42$, $\sigma_1^2 = 100$, $\sigma_2^2 = 16$, $\rho = 0.0$;

(d) Class 4: $\mu_1 = 42$, $\mu_2 = 20$, $\sigma_1^2 = 81$, $\sigma_2^2 = 9$, $\rho = 0.0$.

Patterns of each cluster in data configuration 2 (Fig. 1, right side) were drawn according to the following parameters:

(a) Class 1: $\mu_1 = 45$, $\mu_2 = 22$, $\sigma_1^2 = 144$, $\sigma_2^2 = 9$, $\rho = 0.0$;

(b) Class 2: $\mu_1 = 70$, $\mu_2 = 38$, $\sigma_1^2 = 81$, $\sigma_2^2 = 36$, $\rho = 0.0$;

(c) Class 3: $\mu_1 = 50$, $\mu_2 = 42$, $\sigma_1^2 = 36$, $\sigma_2^2 = 81$, $\rho = 0.0$;

(d) Class 4: $\mu_1 = 42$, $\mu_2 = 2$, $\sigma_1^2 = 9$, $\sigma_2^2 = 144$, $\rho = 0.0$.

Patterns of each class in data configuration 3 (Fig. 2, left side) were drawn according to the following parameters:

(a) Class 1: $\mu_1 = 45$, $\mu_2 = 30$, $\sigma_1^2 = 100$, $\sigma_2^2 = 9$, $\rho = 0.7$;

(b) Class 2: $\mu_1 = 70$, $\mu_2 = 38$, $\sigma_1^2 = 81$, $\sigma_2^2 = 16$, $\rho = 0.8$;

(c) Class 3: $\mu_1 = 45$, $\mu_2 = 42$, $\sigma_1^2 = 100$, $\sigma_2^2 = 16$, $\rho = 0.7$;

(d) Class 4: $\mu_1 = 42$, $\mu_2 = 20$, $\sigma_1^2 = 81$, $\sigma_2^2 = 9$, $\rho = 0.8$.

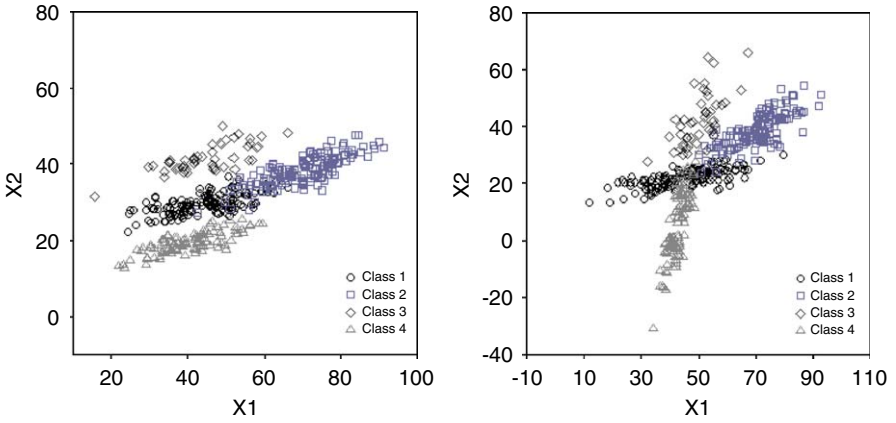


Fig. 2. Clusters drawn from data configurations 3 and 4.

Patterns of each class in data configuration 4 (Fig. 2, right side) were drawn according to the following parameters:

(a) Class 1: $\mu_1 = 45$, $\mu_2 = 22$, $\sigma_1^2 = 144$, $\sigma_2^2 = 9$, $\rho = 0.7$;

(b) Class 2: $\mu_1 = 70$, $\mu_2 = 38$, $\sigma_1^2 = 81$, $\sigma_2^2 = 36$, $\rho = 0.8$;

(c) Class 3: $\mu_1 = 50$, $\mu_2 = 42$, $\sigma_1^2 = 36$, $\sigma_2^2 = 81$, $\rho = 0.7$;

(d) Class 4: $\mu_1 = 42$, $\mu_2 = 2$, $\sigma_1^2 = 9$, $\sigma_2^2 = 144$, $\rho = 0.8$.

The fuzzy clustering algorithms presented in this paper will be applied to the real and synthetic data sets to furnish a fuzzy partition. A hard partition $C = \{C_1, \ldots, C_c\}$ can be obtained from a fuzzy partition defining the cluster $C_i$ $(i = 1, \ldots, c)$ as: $C_i = \{k \in \{1, \ldots, n\} : u_{ik} \geqslant u_{jk}, \forall j \in \{1, \ldots, c\}\}$. The performance of the fuzzy clustering methods considered in the present paper will be evaluated according to the error rate of classification calculated from the hard partition obtained from the corresponding fuzzy partition in the framework of a Monte Carlo simulation with 60 replications for each data configuration.

Each replication furnishes a confusion matrix. In this matrix, the clusters furnished by the clustering algorithm are put in rows and the a priori classes are put in columns. A cluster furnished by the fuzzy clustering algorithm is associated with an a priori class if the number of patterns they share is greater than the number of patterns shared between this cluster and any other class of the a priori partition. In case of equality in the number of patterns shared between two different clusters and an a priori class, the columns are permuted in such a way that the sum of the diagonal elements of this matrix restricted to the rows corresponding to these clusters are a maximum. The cell $(i, j)$ of this matrix contains the number of patterns shared by cluster $i$ and a priori class $j$. The error rate of classification obtained from the confusion matrix is calculated as one minus the sum of the diagonal elements of this matrix divided by the total number of patterns.

The average and standard deviation of the error rate of classification among these 60 replications is calculated. In each replication, a clustering method is run (until the convergence to a stationary value of the adequacy criterion) 60 times and the best result according to the adequacy criterion is selected. Throughout these experiments, parameter $m$ has been set equal to 2.

Table 1 shows the values of the mean and the standard deviation of the error rate of classification for the different methods and data configurations.

As expected, in data configuration 1 (the cluster covariance matrices are diagonal and almost the same), the methods based on a single adaptive distance (FCM-PFCV and FCM-PDFCV) outperform the methods based on adaptive distances for each cluster (GK and FCM-DFCV). For this configuration, the FCM-PFCV method presented the best performance and the FCM method presented the worst.

Data configuration 2 presents cluster covariance matrices that are diagonal but unequal. As a result, the methods based on adaptive distances for each cluster (GK and FCM-DFCV) outperform the methods based on a single adaptive distance (FCM-PFCV and FCM-PDFCV). For this configuration, the best performance was presented by the FCM-DFCV method. The FCM method presented a performance similar to the FCM-PFCV and FCM-PDFCV methods.

Table 1
Mean and standard deviation of the error rate of classification for the different fuzzy clustering methods and data configurations

| Config. | FCM | | FCM-PFCV | | FCM-PDFCV | | GK | | FCM-DFCV | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| 1 | 0.4001 | 0.0536 | 0.1242 | 0.0369 | 0.1259 | 0.0397 | 0.1803 | 0.0763 | 0.1656 | 0.0805 |
| 2 | 0.2207 | 0.0285 | 0.2270 | 0.0368 | 0.2333 | 0.0331 | 0.1501 | 0.0230 | 0.1377 | 0.0217 |
| 3 | 0.4543 | 0.0351 | 0.0723 | 0.0121 | 0.2456 | 0.0819 | 0.0700 | 0.0112 | 0.3179 | 0.1205 |
| 4 | 0.2950 | 0.0264 | 0.3120 | 0.0543 | 0.2977 | 0.0310 | 0.0683 | 0.0252 | 0.2106 | 0.0382 |

In Data configuration 3 (the cluster covariance matrices are not diagonal but almost the same), the methods based on adaptive quadratic distances defined by non-diagonal matrices (FCM-PFCV and GK) outperform the methods based on adaptive quadratic distances defined by diagonal matrices (FCM-PDFCV, FCM-DFCV). The FCM method presented the worst performance once again in this configuration. The best performance was presented by the GK method.

Finally, Data configuration 4 has cluster covariance matrices that are not diagonal and are also unequal. As a result, the methods based on adaptive distances for each cluster (GK and FCM-DFCV) outperform the methods based on single adaptive distances (FCM-PFCV and FCM-PDFCV). For this configuration, the GK method presented the best performance. The FCM method presented a performance similar to FCM-PFCV and FCM-PDFCV methods.

In conclusion, for these data configurations, the methods based on adaptive distances outperform the FCM method. Concerning the adaptive methods, their performance depends on the intra-cluster structure: the methods based on a single adaptive distance perform well when the a priori classes have almost the same dispersions, whereas the methods based on an adaptive distance for each cluster perform well when the a priori classes have unequal dispersions.

### *5.2. UCI machine learning repository data sets*

The FCM, FCM-PFCV, FCM-PDFCV, GK and FCM-DFCV clustering algorithms were applied to data sets on iris plants, the thyroid gland, wine and images [4]. These data sets can be found in the UCI machine learning repository http://www.ics.uci.edu/mlearn/MLRepository.html. The cluster partition obtained with these clustering methods was compared with the known a priori cluster partition. The comparison criterion used is the error rate of classification. Concerning these data sets, each clustering method is run (until the convergence to a stationary value of the adequacy criterion) 60 times and the best result is selected according to the adequacy criterion. The error rate of classification is calculated for the best result. The parameter $m$ has been set equal to 2.

#### *5.2.1. Iris plant data set*

This data set consists of three types (classes) of iris plants: iris setosa, iris versicolour and iris virginica. The three classes have 50 instances each. One class is linearly separable from the other two; the latter are not linearly separable from each other. Each class is described by four real valued attributes. These attributes are: sepal length, sepal width, petal length and petal width.

The FCM, FCM-PFCV, FCM-PDFCV, GK and FCM-DFCV clustering algorithms were applied to this data set. The 3-cluster partition obtained with these clustering algorithms was compared with the known a priori 3-cluster partition. The error rates of classification were 10.67%, 6.0%, 4.67%, 10.0% and 7.33% for these clustering methods, respectively. In conclusion, for this data set, the fuzzy clustering methods based on single adaptive distances (FCM-PFCV and FCM-PDFCV) outperform the other methods (FCM, FCQAD, FCM-DFCV) concerning the identification of the a priori classes of patterns.

#### *5.2.2. Thyroid gland data set*

This data set consists of three classes concerning the state of the thyroid: normal, hyperthyroidism and hypothyroidism. The classes (1, 2 and 3) have, respectively, 150, 35 and 30 instances. Each class is described by five real-valued attributes. These attributes are: T3-resin uptake test, total serum thyroxin, total serum triiodothyronine, basal thyroid-stimulating hormone (TSH) and maximal absolute difference of TSH value.

The clustering algorithms presented in this paper were applied to this data set. The 3-cluster partition obtained with these clustering algorithms were compared with the known a priori 3-cluster partition. The error rates of classification were 20.93%, 45.58%, 11.16%, 13.02% and 3.72% for the FCM, FCM-PFCV, FCM-PDFCV, GK and FCM-DFCV clustering methods, respectively. We can conclude that, for this data set, the fuzzy clustering methods based on quadratic adaptive distances defined by diagonal matrices (FCM-PDFCV and FCM-DFCV) outperform the other methods (FCM, FCM-PFCV, FCQAD) concerning the identification of the a priori classes of patterns.

*5.2.3. Wine data set*

This data set consists of three types (classes) of wines grown in the same region in Italy but derived from three different cultivars. The classes (1, 2 and 3) have, respectively, 59, 71 and 48 instances. Each wine is described by 13 real-valued attributes representing the quantities of 13 constituents found in each of the three types of wines. These attributes are: (1) alcohol; (2) malic acid; (3) ash; (4) alkalinity of ash; (5) magnesium; (6) total phenols; (7) flavonoids; (8) non-flavonoid phenols; (9) proanthocyanins; (10) colour intensity; (11) hue; (12) OD280/OD315 of diluted wines and (13) proline.

The FCM, FCM-PFCV, FCM-PDFCV, GK and FCM-DFCV clustering algorithms were applied to this data set. Again, the 3-cluster partition obtained with these clustering algorithms was compared with the known a priori 3-cluster partition. The error rates of classification were 31.46%, 33.15%, 5.06%, 39.33% and 8.43% for these clustering methods, respectively.

This indicates that, for this data set, the fuzzy clustering methods based on adaptive quadratic distances defined by diagonal matrices (FCM-PDFCV and FCM-DFCV) outperform the other methods (FCM, FCM-PFCV, FCQAD) concerning the identification of the a priori classes of patterns.

The poor results obtained with the FCM-PFCV and GK algorithms on this data set illustrate the difficulties of these methods in managing the inversion of the matrices that define the adaptive quadratic distances in the algorithms.

*5.2.4. Image segmentation data*

This data set consists of images that were drawn randomly from a database of seven outdoor images. The images were segmented by hand to create the seven class labels: sky, cement, window, brick, grass, foliage and path. The seven classes have 330 instances each. Each class is described by 18 real-valued attributes. These attributes are: region-centroid-col; region-centroid-row; short-line-density-5; short-line-density-2; vedge-mean; vegde-sd; hedge-mean; hedge-sd; intensity-mean; rawred-mean; rawblue-mean; rawgreen-mean; exred-mean; exblue-mean; exgreen-mean; value-mean; saturation-mean and hue-mean.

The FCM, FCM-PFCV, FCM-PDFCV, GK and FCM-DFCV clustering algorithms were applied to this data set. The 7-cluster partition obtained with these clustering algorithms was compared with the known a priori 7-cluster partition. The error rates of classification were 60.13%, 83.55%, 37.88% and 83.03% for the FCM, FCM-PFCV, FCM-PDFCV and FCM-DFCV clustering methods, respectively. There are no results from GK algorithm because in the very first iteration of all replications, this algorithm presented numerical problems and the matrix defining the quadratic adaptive distances could not be inverted. This problem with the data set also occurred with the FCM-PFCV method in later iterations of the replications.

In conclusion, for these real data sets, the methods based on adaptive quadratic distances defined by diagonal matrices (FCM-PDFCV and FCM-DFCV) outperform the methods based on adaptive quadratic distances defined by non-diagonal matrices (FCM-PFCV and GK). The methods FCM and FCM-PFCV presented similar performance regarding these real data sets. These experiments stress the difficulties of the methods based on quadratic adaptive distances (FCM-PFCV and GK) in managing real data sets: they either furnish no results, as with the GK method in the case of the image segmentation data set or they present a worse performance than the methods based on adaptive quadratic distances defined by diagonal matrices (FCM-PDFCV, FCM-DFCV), as in the case of wine data set. The methods based on adaptive quadratic distances defined by diagonal matrices have no such difficulties.

*5.3. Fuzzy partition and clusters interpretation: the wine data set*

The interest of the interpretation indices presented in this paper will be highlighted here through the results obtained with application of the FCM and FCM-PDFCV methods to the wine data set.

Table 2
Overall heterogeneity index for FCM and FCM-PDFCV methods

| Method | FCM ($l = 1$) | FCM-PDFCV ($l = 3$) |
|---|---|---|
| $Rl$ ($l = 1, 3$) | 0.874635 | 0.398488 |

Table 3
Overall heterogeneity indices concerning the variables for the FCM method (%)

| Variables | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $COR(j)$ | 39.1 | 5.7 | 5.6 | 20.4 | 18.4 | 27.3 | 31.6 | 11.2 | 11.0 | 11.7 | 9.8 | 14.3 | 87.6 |
| $CTR(j)$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 99.95 |

Table 4
Overall heterogeneity indices concerning the variables for the FCM-PDFCV method (%)

| Variables | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $COR(j)$ | 35.6 | 28.9 | 6.9 | 22.6 | 13.7 | 51.9 | 64.0 | 28.0 | 30.7 | 32.5 | 40.7 | 57.3 | 48.11 |
| $CTR(j)$ | 6.4 | 4.7 | 0.8 | 3.4 | 1.8 | 12.5 | 20.6 | 4.5 | 5.1 | 5.6 | 7.9 | 15.6 | 10.7 |

Table 5
Cluster heterogeneity indices for the FCM clustering method

| Cluster | Cardinal | $T^1(i)$ | $B^1(i)$ | $J^1(i)$ |
|---|---|---|---|---|
| 1 | 71 | 0.353242 | 0.367451 | 0.254105 |
| 2 | 46 | 0.608446 | 0.632297 | 0.442048 |
| 3 | 61 | 0.038312 | 0.000252 | 0.303847 |

*5.3.1. Fuzzy partition interpretation*

Table 2 shows that the proportion of the overall SSQ explained by the partition is higher for the partition obtained with the FCM method. This is due to the fact that the partition furnished by the FCM method is mainly determined by variable 13 (*Proline*) and the clusters obtained are very homogeneous with regard to this variable.

Comparing the values of $COR^1(j)$ with the value of $R1$ (see Tables 2 and 3) for the partition obtained with the FCM clustering method, we may conclude that the discriminant power of variable 13 (*Proline*) is equivalent to the mean value, whereas all other variables have a discriminant power below the mean. From Table 3, we can see that variable 13 (*Proline*) provides a high contribution to the separation of the prototypes of the clusters ($CTR^1(13) = 99.99$).

Comparing the values of $COR^3(j)$ with the values of $R3$ (see Table 2) for the partitions obtained using the FCM-PDFCV clustering method (see Table 4), we may conclude that the discriminant power of the variables 6 (*Total phenols*), 7 (*Flavonoids*), 11 (*Hue*), 12 (*OD*280/*OD*315 *of diluted wines*) and 13 (*Proline*) is above the average, whereas all other variables have a discriminant power slightly below the average. From Table 4, we can see that variables 6, 7, 12 and 13 provide important contributions to the separation of the prototypes of the clusters.

*5.3.2. Cluster interpretation*

From Table 5, we can see that in the partition obtained with the FCM Clustering method, the Cluster 3 mean vector is the closest to the global mean vector ($B^1(3) = 0.000252$), while the Cluster 2 mean vector is the farthest ($B^1(2) = 0.632297$). On the other hand, Cluster 2 is the least homogeneous ($J^1(2) = 0.442048$) and Cluster 1 is the most homogeneous ($J^1(1) = 0.254105$) of the three clusters. The values in Table 6 show that in the partition obtained from the FCM-PDFCV Clustering Method, the Cluster 3 mean vector is the closest to the global mean ($B^3(3) = 0.135392$),

Table 6
Cluster heterogeneity indices for the FCM-PDFCV clustering method

| Cluster | Cardinal | $T^3(i)$ | $B^3(i)$ | $J^3(i)$ |
|---|---|---|---|---|
| 1 | 52 | 0.379111 | 0.466972 | 0.320905 |
| 2 | 64 | 0.361354 | 0.397634 | 0.337320 |
| 3 | 62 | 0.259534 | 0.135392 | 0.341775 |

Table 7
Cluster heterogeneity indices concerning the variables for the FCM method (%)

| | Cluster 1 | | | Cluster 2 | | | Cluster 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | *COR* | *CTR* | *CE* | *COR* | *CTR* | *CE* | *COR* | *CTR* | *CE* |
| 1 | 14.3 | 0.0 | 0.0 | 24.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 2 | 0.2 | 0.0 | 0.0 | 4.1 | 0.0 | 0.0 | 1.3 | 0.0 | 0.0 |
| 3 | 2.7 | 0.0 | 0.0 | 2.6 | 0.0 | 0.0 | 0.3 | 0.0 | 0.0 |
| 4 | 6.5 | 0.0 | 0.0 | 13.8 | 0.0 | 0.0 | 0.07 | 0.03 | 0.0 |
| 5 | 10.5 | 0.06 | 0.02 | 4.3 | 0.01 | 0.01 | 3.6 | 32.1 | 0.01 |
| 6 | 4.9 | 0.0 | 0.0 | 20.6 | 0.0 | 0.0 | 1.8 | 0.03 | 0.0 |
| 7 | 2.5 | 0.0 | 0.0 | 23.7 | 0.0 | 0.0 | 5.3 | 2.3 | 0.0 |
| 8 | 1.4 | 0.0 | 0.0 | 8.5 | 0.0 | 0.0 | 1.2 | 0.0 | 0.0 |
| 9 | 2.4 | 0.0 | 0.0 | 8.1 | 0.0 | 0.0 | 0.4 | 0.0 | 0.0 |
| 10 | 6.6 | 0.0 | 0.0 | 2.7 | 0.0 | 0.0 | 2.3 | 0.54 | 0.0 |
| 11 | 0.16 | 0.0 | 0.0 | 6.6 | 0.0 | 0.0 | 2.0 | 0.0 | 0.0 |
| 12 | 1.1 | 0.0 | 0.0 | 10.7 | 0.0 | 0.0 | 2.5 | 0.0 | 0.0 |
| 13 | 32.2 | 99.9 | 36.7 | 55.4 | 99.9 | 63.2 | 0.01 | 66.9 | 0.01 |

Table 8
Cluster heterogeneity indices concerning the variables, for the FCM-PDFCV method (%)

| | Cluster 1 | | | Cluster 2 | | | Cluster 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | *COR* | *CTR* | *CE* | *COR* | *CTR* | *CE* | *COR* | *CTR* | *CE* |
| 1 | 0.01 | 0.0 | 0.0 | 16.3 | 7.4 | 2.9 | 19.3 | 25.7 | 3.4 |
| 2 | 19.5 | 6.8 | 3.2 | 5.4 | 2.2 | 0.9 | 4.0 | 4.8 | 0.6 |
| 3 | 0.4 | 0.1 | 0.05 | 1.7 | 0.5 | 0.2 | 4.7 | 4.3 | 0.6 |
| 4 | 8.0 | 2.6 | 1.2 | 13.3 | 5.0 | 2.0 | 1.2 | 1.3 | 0.2 |
| 5 | 0.3 | 0.08 | 0.04 | 7.4 | 2.5 | 1.0 | 6.0 | 32.1 | 0.8 |
| 6 | 26.3 | 13.6 | 6.3 | 25.5 | 15.5 | 6.1 | 0.08 | 0.1 | 0.01 |
| 7 | 35.6 | 24.6 | 11.5 | 28.2 | 22.8 | 9.1 | 0.08 | 0.2 | 0.02 |
| 8 | 15.7 | 5.4 | 2.5 | 12.2 | 4.9 | 1.9 | 1.2 | 0.06 | 0.0 |
| 9 | 16.8 | 6.0 | 2.8 | 13.9 | 5.8 | 2.3 | 0.01 | 0.01 | 0.0 |
| 10 | 13.0 | 4.8 | 2.2 | 0.3 | 0.1 | 0.05 | 19.2 | 24.4 | 3.3 |
| 11 | 27.4 | 11.5 | 5.4 | 8.6 | 4.2 | 1.7 | 4.7 | 6.9 | 0.9 |
| 12 | 37.0 | 21.5 | 10.0 | 17.7 | 12.1 | 4.8 | 2.7 | 5.4 | 0.7 |
| 13 | 5.9 | 2.8 | 1.3 | 29.7 | 16.7 | 6.6 | 12.5 | 20.6 | 2.8 |

while the Cluster 1 mean vector is the farthest ($B^3(1) = 0.466972$) and also the most homogeneous ($J^3(1) = 0.320905$) of the three clusters. Table 7 shows the cluster heterogeneity indices concerning the variables for the FCM clustering method. From this table, we can see that variable 13 plays the most important role in the heterogenity of the clusters ($CTR^1(13, 1) = 99.99\%$, $CTR^1(13, 2) = 99.99\%$ and $CTR^1(13, 3) = 66.90\%$) and a homogeneous behaviour within clusters 1 and 2 ($COR^1(13, 1) = 32.2\%$ and $COR^1(13, 2) = 55.4\%$). Finally, variable 13 alone contributes to the separation ($CE(13, 1) = 36.7$ and $CE(13, 2) = 63.2$) of the Cluster 1 mean vector and the Cluster 2 mean vector (the Cluster 3 mean vector is the closest to the global mean vector).

Table 8 shows the cluster heterogeneity indices concerning the variables for the FCM-PDFCV clustering method. From this table, we can see that variables 7 and 12 display a homogeneous behaviour in Cluster 1 and the most important

contribution to the heterogenity of this cluster; variables 6, 7 and 13 display a homogeneous behaviour in Cluster 2 and the most important contribution to the heterogenity of this cluster; and variables 1 and 10 display a homogeneous behaviour in Cluster 3 and the most important contribution to the heterogenity of this cluster. Finally, no variable alone contributes to the separation between the cluster mean vectors (indeed, no variable presents a very high $CE(j,h)$ for any of the clusters). For example, Clusters 1, 2 and 3 mean vectors are a little more eccentric regarding the variables 7 (*Flavonoids*) and 12 ($OD280/OD315$ of diluted wines), variables 6 (*Total phenols*), 7 (*Flavonoids*) and 13 (*Proline*) and variables 1 (Alcohol), 10 (*Colour intensity*) and 13 (*Proline*).

## 6. Conclusion

This paper presented partitional fuzzy clustering methods based on adaptive quadratic distances: methods based on a single adaptive quadratic distance (FCM-PFCV and FCM-PDFCV) and a method based on an adaptive quadratic distance for each class defined by a diagonal matrix (FCM-DFCV), which is a special case of the method introduced by Gustafson and Kessel [14] based on a quadratic adaptive distance of each cluster defined by a fuzzy covariance matrix.

The FCM-PFCV method is based on a single adaptive quadratic distance defined by a pooled fuzzy covariance matrix. The algorithm starts from an initial membership degree for each pattern in each cluster and alternates a representation step and an allocation step until the convergence of the algorithm. The representation has two stages. In the first stage, the membership degree of each pattern in each cluster and the matrix that defines the single adaptive quadratic distance are fixed. The algorithm then looks for the best prototype of each cluster that minimizes the adequacy criterion. The solution for the best prototype is presented in this paper. In the second stage, the membership degree of each pattern in each cluster and the prototypes of the corresponding clusters are fixed. The algorithm then looks for the best matrix that defines the single adaptive quadratic distance and minimizes the adequacy criterion. The solution for the best matrix, as presented in this paper, is a pooled fuzzy covariance matrix. In the allocation step, the prototypes of the clusters and the matrix that defines the adaptive quadratic distance are fixed. The algorithm looks for the best membership degree of each pattern. The solution for the best membership degree is presented in this paper.

The FCM-PDFCV method, which is a special case of the FCM-PFCV method, is based on a single adaptive quadratic distance defined by a pooled fuzzy covariance matrix restricted to be diagonal. As shown in Section 4.2.1.2, this method imposes the same variable-specific within-cluster SSQs for all variables.

The FCM-DFCV method, which is a special case of the method introduced by Gustafson and Kessel [14], is based on an adaptive quadratic distance for each class defined by a fuzzy covariance matrix restricted to be diagonal. As shown in Section 4.2.1.2, this method imposes the same variable-specific within-cluster SSQs for all variables and also imposes the same within-cluster SSQs for all variables in a given cluster.

The problem of interpreting and evaluating the obtained fuzzy partition was addressed. A family of indices for the standard fuzzy $c$-means clustering method (FCM) as well as for the adaptive fuzzy clustering methods based on the notion of SSQs were introduced, constituting a suitable adaptation of indices for interpreting and evaluating partitions furnished by the standard hard $c$-means method [6]. The interest of these indices was highlighted through the results obtained with the application of FCM and FCM-PDFCV methods for a real data set.

The simulation study carried out showed that for the data configurations considered, the methods based on adaptive distances outperform the FCM method. Moreover, the performance of the adaptive methods depend on the intra-cluster structure: the methods based on a single adaptive distance perform well when the a priori classes have almost the same dispersions whereas the methods based on an adaptive distance for each cluster perform well when the a priori classes have unequal dispersions. Concerning the application of these methods to real data sets, the methods based on adaptive quadratic distances defined by diagonal matrices outperform the methods based on adaptive quadratic distances defined by non-diagonal matrices as well as the FCM method, showing the difficulties of the methods based on quadratic adaptive distances defined by non-diagonal matrices in managing real data sets due to numerical problems that can occur with regard to inverting the non-diagonal matrices that define the adaptive quadratic distances.

## Acknowledgements

The authors would like to thank CNPq (Brazilian Agency) for its financial support and the anonymous referees for their helpful comments and suggestions to improve the paper.

## References


[1] R. Babuska, P.J. Van der Veen, U. Kaymak, Improved covariance estimation for Gustafson–Kessel clustering. in: Proc. 2002 IEEE Internat. Conf. on Fuzzy Systems, vol. 2, Honolulu, HI, 2002, pp. 1081–1085.

[2] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

[3] J.C. Bezdek, J. Keller, R. Krisanpuram, N.R. Pal, Fuzzy Models and Algorithms for Pattern Recognition and Image Processing, Kluwer, Norwell (USA), 1999.

[4] C.L. Blake, C.J. Merz, UCI Repository of Machine Learning Databases ⟨http://www.ics.uci.edu/mlearn/MLRepository.html⟩. University of California, Department of Information and Computer Science, Irvine, CA.

[5] H.H. Bock, Classification and clustering: problems for the future, in: E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, B.L. Burtschy (Eds.), New Approaches in Classification and Data Analysis, Springer, Berlin, 1993, pp. 3–24.

[6] G. Celeux, E. Diday, G. Govaert, Y. Lechevallier, H. Ralambondrainy, Classification Automatique des Données, Bordas, Paris, 1989.

[7] E. Diday, G. Govaert, Classification automatique avec distances adaptatives, RAIRO Inform. Comput. Sci. 11 (4) (1976) 329–349.

[8] E. Diday, J.J. Simon, Clustering analysis, in: K.S. Fu (Ed.), Digital Pattern Recognition, Springer, Heidelberg, 1976, pp. 47–94.

[9] J.C. Dunn, A fuzzy relative to the ISODATA process and its use in detecting compact, well-separated clusters, J. Cybernet. 3 (1974) 32–57.

[10] B. Everitt, Cluster Analysis, Halsted, New York, 2001.

[11] H. Frigui, O. Nasraoui, Unsupervised learning of prototypes and attribute weights, Pattern Recognition 37 (2000) 567–581.

[12] A.D. Gordon, Classification, Chapman & Hall/CRC, Boca Raton, FL, 1999.

[13] P.J.F. Groenen, K. Jajuga, Fuzzy clustering with squared Minkowsky distances, Fuzzy Sets and Systems 120 (2001) 227–237.

[14] D.E. Gustafson, W.C. Kessel, Fuzzy clustering with a fuzzy covariance matrix, in: Proc. IEEE Conf. Decision Control, San Diego, CA, pp. 761–766.

[15] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surveys 31 (3) (1999) 264–323.

[16] K. Jajuga, $L_1$-norm based fuzzy clustering, Fuzzy Sets and Systems 39 (2001) 63–83.

[17] R. Krishnapuram, J. Kim, A note on the Gustafson–Kessel and adaptive fuzzy clustering algorithms, IEEE Trans. Fuzzy Systems 7 (4) (1999) 453–461.

[18] H. Spaeth, Cluster Analysis Algorithms, Wiley, New York, 1980.

[19] M.S. Yang, A survey of fuzzy clustering, Math. Comput. Modelling 18 (11) (1993) 1–16.