Int. J. Fuzzy Syst. (2021) 23(7):2182–2199
https://doi.org/10.1007/s40815-021-01088-9

Comparing Different Approaches to Archetypal Analysis as a Fuzzy Clustering Tool

Abdul Suleman[1]

Received: 3 August 2020 / Revised: 19 February 2021 / Accepted: 19 March 2021 / Published online: 10 June 2021

Abstract We summarize the results of an intensive simulation study carried out to compare the performance of three approaches to archetypal analysis regarded as a fuzzy clustering tool: the original approach, namely that of Cutler and Breiman (Technometrics 36(4):338–347, 1994), the Ding et al. (IEEE Trans Pattern Anal Mach Intell 32(1):45–55, 2010) proposal, and the factorized fuzzy c-means algorithm. The artificial data we use in our experiment are generated from polytopes in low-dimensional $\mathbb{R}^n$ spaces $(2 \leq n \leq 7)$, and comprise a diversity of cluster contexts. The simulation results show that the original proposal is generally a more accurate method to uncover the cluster structure hidden in the data and to reproduce the data themselves. However, this supremacy, if any, is not clear for the data generated from real life problems, and devoted to unsupervised clustering problems.

Keywords Fuzzy clustering · Matrix factorization · Archetypal analysis · Simulation

1 Introduction

The application of the matrix factorization approach to data analysis, notably in fuzzy clustering, appeared in the literature long before the seminal work on nonnegative matrix factorization (NMF) by Lee and Seung [26], following that of Paatero and Tapper [35]. Woodbury and Clive [47] devise a method to estimate fuzzy partitions, hypothetically underlying high-dimensional clinical categorical data, for medical diagnostic and prognostic purposes. It is based on the so-called grade of membership (GoM) model, and has since also been successfully used beyond the medical contexts for which it was primarily designed (e.g. [28, 38, 41, 43]). It expresses the position of every individual in a structure set out by $c \geq 2$ pure types or prototypes, as a convex combination of these pure types. Independently, Mirkin and Satarov [30] propose an extension of the GoM model to real-valued data, and it reflects the use of matrix factorization for fuzzy cluster analysis as we understand it nowadays. However, the underlying model does not per se keep the prototypes close to the observations [34]. A refinement of this model, that potentially overcomes the referred drawback, is provided by Cutler and Breiman [12], and called archetypal analysis (AA). In this case, the prototypes, now archetypes, are themselves convex combinations of data points, thus entailing a representation of meaningful cluster centroids. An alternative approach to the original proposal [30] is provided by Nascimento in [33], and regarded as its smooth version; it is known as fuzzy clustering with proportional membership (FCPM), and is more resonant its FCPM-2 version (see also [32]).

The AA enjoyed some popularity and following in the literature, and again attracted researchers' attention, e.g. [4, 7, 16–18, 31, 39, 44]. The work by Ding et al. [13] is another exemplary application of the matrix factorization approach to cluster analysis, notably the classical k-means. It provides a reliable algorithm to estimate the archetypes, and it is explored in the present study. Also of note is the work by Thurau et al. [42], especially when it comes to addressing massive data sets.

At this point, it is important to emphasize that the data are analyzed differently in a matrix factorization

✉ Abdul Suleman
abdul.suleman@iscte-iul.pt

[1] Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal

framework than in a common fuzzy cluster analysis. The popular fuzzy c-means (FCM) algorithm [3], for example, seeks the central properties of the data and the prototypes are therefore cluster centers. In the alternative approach, the prototypes are instead the extreme points of a bounded convex polyhedron or, simply, a polytope that, by assumption, is the *population* from which the data are sampled. In other words, the population is the convex hull of the prototypes. A bridge between these two ways of looking at the data structure is given in [37] and is called factorized fuzzy c-means (FFCM) algorithm.

Despite the recent growing interest in AA or, more generally, in the matrix factorization approach to fuzzy clustering,[1] the literature lacks a systematic study on the behavior of different methods used to operationalize this analytical tool. We will show that, under an alternating optimization scheme, different approaches to AA differ from each other only in the way the archetypes are estimated. The estimation of the partition matrix can be reduced to a set of independent constrained least squares (CLS) problems and solved using a flat common solver. This study therefore concerns the estimation of archetypes. We must mention here the related work carried out by Mendes and Nascimento in [29]. While these authors explore two different ways of tackling the extremal approach to fuzzy clustering, namely FCPM and an AA, our study fits within the AA framework. This explains why we do not include FCPM. Our research question can therefore be formulated as follows: how reliable is the matrix factorization approach to fuzzy clustering, provided that it is an AA? In an attempt to answer this question, the following empirical study is conducted.

We consider three different approaches to estimate the archetypes in AA: the original proposal [12] which results in another CLS problem, the one proposed in [13] developed under the framework of semi-nonnegative matrix factorization (semi-NMF), and the FFCM [37]. We do not subject the archetypes to constraints other than the one entailed in its definition. For example, we do not require them to be archetypoids, i.e. observed data, as in [44]. However, when this restriction is mandatory, the outcome of our research work can be used upstream for seeding purposes. We subject the three referred approaches to AA to a test with synthetic data, by means of an intensive Monte Carlo simulation, and data from real life problems and devoted to clustering. As the simulation is computationally highly demanding, we opt to draw the synthetic data from polytopes in low-dimensional spaces (≤ 7). In this case we notice a better performance of the algorithm proposed in [12] when compared to those in [13, 37]. However, the same does not hold when it comes to clustered data. Here, the latter two algorithms outperfom the former one. Moreover, the results are promising when we compare their outcomes to that of the FCM algorithm. So the output of our research work make practitioners aware of the behavior of different methods of performing AA and, therefore, can guide their choice for a particular method according to the specific nature of the problem at hand. Additionally, they can look at AA more confidently as a credible alternative to FCM for data hypothetically organized in clusters.

Our manuscript develops as follows. In Sect. 2 we briefly describe the theoretical framework of the matrix factorization approach to fuzzy clustering and explain how different forms of AA are operationalized; then we give a detailed account of our experimental design and the results obtained in Sect. 3; finally, Sect. 4 provides guidelines for future work and some concluding remarks.

2 Matrix Factorization Framework

2.1 A Brief Review on NMF

An elegant form of introducing the archetypal analysis (AA) is to frame it in a more general setting, namely under a matrix factorization approach to data decomposition. For pedagogical purposes, we start from the nonnegative matrix factorization (NMF). In mathematical terms, this approach can be formulated as follows.

Let $\mathbf{X} = [\mathbf{X}_{jk}]$ be an $n \times N$ real sample data matrix, and suppose there are two matrices $\mathbf{U} = [\mathbf{U}_{ik}]$ of size $c \times N$, and $\mathbf{V} = [\mathbf{V}_{ji}]$ of size $n \times c$, where $n \geq 2$ is the dimension of the feature space, $N > n$ is the sample size, and $c \geq 2$, such that $\mathbf{X}_{jk}$, $\mathbf{U}_{ik}$, $\mathbf{V}_{ji} \in \mathbb{R}_0^+$, and

$$\mathbf{X} \approx \mathbf{VU}. \tag{1}$$

This can also be written suggestively as

$$\mathbf{X}_+ \approx \mathbf{V}_+\mathbf{U}_+, \tag{2}$$

where the plus sign in subscript highlights the nonnegative restrictions. The product $\mathbf{VU}$ is called an NMF of $\mathbf{X}$, and the notation (1) emphasizes that we are seeking an approximate factorization for the data matrix. For example, $\mathbf{X}$ may be corrupted by additive noise and the product $\mathbf{VU}$ is a representative of noise-free observations.

The column-wise representation of $\mathbf{X}$, $\mathbf{U}$ and $\mathbf{V}$ are: $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2 \ldots \mathbf{x}_N]$, $\mathbf{U} = [\mathbf{U}_1\ \mathbf{U}_2 \ldots \mathbf{U}_N]$, and $\mathbf{V} = [\mathbf{v}_1\ \mathbf{v}_2 \ldots \mathbf{v}_c]$, respectively. The columns of $\mathbf{V}$ can be interpreted as basis (generator) or component vectors whereas the matrix $\mathbf{U}$

[1] A special session entitled 'SS_37: Matrix Factorization for Fuzzy Clustering and Related Approaches' took place at 2017 IEEE International Conference on Fuzzy Systems, in Naples, Italy: https://www.fuzzieee2017.org/specialSessions.php.

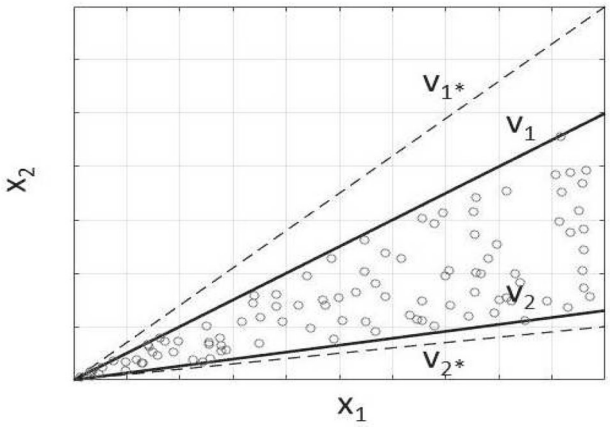

Fig. 1 Illustration of NMF in 2-dimensional space, for the ideal setting $\mathbf{X} = \mathbf{VU}$. The circles represent data points

gathers up the representations of the data points with regard to these vectors [24]. So we alternatively write (1) as

$$\mathbf{x}_k \approx \sum_{i=1}^{c} \mathbf{U}_{ik}\mathbf{v}_i, \quad k = 1, 2, ..., N, \tag{3}$$

to highlight each observation $\mathbf{x}_k$ as a nonnegative linear combination of the component vectors. This analytical representation tells that the data points lie (approximately) in a (polyhedral) cone generated by the component vectors $\mathbf{v}_1, \mathbf{v}_2, ...,$ and $\mathbf{v}_c$ [14]. Figure 1 exemplifies it for $n = 2$ and $c = 2$. In general, the representation of data points (2) is not unique; if a cone, say Γ, contains the data any other cone Γ^*, such that $\Gamma \subset \Gamma^*$, also contains the data (in Fig. 1, the data (x_1, x_2) represented by circles are contained in both cones: one with solid line $[\mathbf{v}_1\ \mathbf{v}_2]$, Γ in our nomenclature, and Γ^* with a dashed line $[\mathbf{v}_{1^*}\ \mathbf{v}_{2^*}]$). The issue of non-uniqueness of NMF is beyond the scope of this research and therefore will not be discussed here; interested reader may wish to consult [14], where the subject is treated in greater detail.

Given a pre-specified value of c, the factors $\mathbf{U}$ and $\mathbf{V}$ can be estimated by minimizing the generic criterion, loss or objective function

$$D_c(\mathbf{X}\|\mathbf{VU}), \tag{4}$$

subject to the nonnegative constraints referred to above. In (4), $D_c(.)$ is a divergence measure that accounts for the difference or discrepancy between $\mathbf{X}$ and the product $\mathbf{VU}$. Three broad classes of divergence measures emerge in developing an NMF [8]: the Bregman divergences, the Amaris α-divergences and the Csiszár divergences. The aforementioned paper explores this latter measure of divergence; examples of the application of the two former measures are [1] and [9], respectively. So there is great flexibility in tackling a data decomposition problem using an NMF approach, thus also making it possible to tailor the divergence measure to specific purposes. For example, Cichocki et al. [9] explore the α-divergences ($\alpha = 0.5; 1; 2$) in EEG data classification. Nevertheless, the least squares or Frobenius norm, symbolically $\|.\|_F$, which is a particular case of the Bregman divergence, is perhaps the most popular measure used in solving NMF problems. In particular, it is optimal for additive Gaussian noise [10] (quoted in [11]). The underlying optimization problem aims to minimize the objective function

$$J_c(\mathbf{U}, \mathbf{V}|\mathbf{X}) = \frac{1}{2}\|\mathbf{X} - \mathbf{VU}\|_F^2, \tag{5}$$

and it is adopted in this study.

It is known that any divergence measure is individually convex in $\mathbf{U}$ and $\mathbf{V}$, but not necessarily in the product $\mathbf{VU}$. Therefore, an alternating optimization scheme is suitable for estimation purposes [2, 50]. Furthermore, this optimization technique allows parallelization [25] and can be very fast [2]. The convergence to a local minimum is another issue of NMF approaches, regardless of the way the matrices $\mathbf{U}$ and $\mathbf{V}$ are estimated. An appropriate initialization of the estimation algorithms can help mitigate this drawback; see [5] for details on this subject. Readers interested in a comprehensive review on NMF may refer to [2, 50], from which we derive most of this section.

2.2 Archetypal Analysis

More recently, Ding et al. [13] extend the scope of the application of NMF ideas, as expressed in (2), allowing the data matrix to have mixed signs, that is $\mathbf{X}_{jk} \in \mathbb{R}$. However, this new perspective of the matrix factorization approach to data analysis only entails the entries of the matrix $\mathbf{V}$, i.e. $\mathbf{V}_{ji}$, to have mixed signs; the matrix $\mathbf{U}$ remains nonnegative. Formally, this modifies (2) to

$$\mathbf{X}_{\pm} \approx \mathbf{V}_{\pm}\mathbf{U}_{+}, \tag{6}$$

and is called semi-NMF. The above cited authors recast the classical objective function for k-means clustering in a matrix factorization form, provided $\mathbf{U}$ is an indicator matrix, which leads its entries $\mathbf{U}_{ik} \in \{0, 1\}$ and, of course,

$$\sum_{i=1}^{c} \mathbf{U}_{ik} = 1. \tag{7}$$

Subsequently, the columns of $\mathbf{V}$ are referred to as prototypes. Relaxing the condition $\mathbf{U}_{ik} \in \{0, 1\}$ to $\mathbf{U}_{ik} \in [0, 1]$, while keeping the restriction (7), potentially turns (6) into a soft or fuzzy clustering of $\mathbf{X}$.

This approach to fuzzy clustering was first proposed by Mirkin and Satarov in [30]. Each observation (3) is now a(n) (approximate) convex combination of c prototypes,

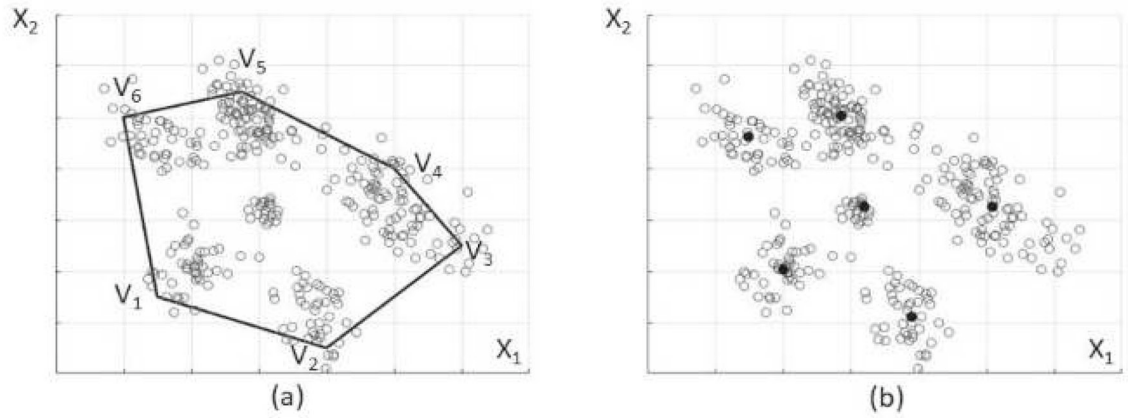

Fig. 2 (a) An approximate convex hull of data points (grey circles) in 2-dimensional space, with $c = 6$ vertices; (b) output of fuzzy c-means algorithm for 6-cluster solution (the prototypes are represented by filled black circles)

which configure a polytope and are its extreme points or vertices (see Fig. 2a for an example of $n = 2$ and $c = 6$). In fact, this polytope translates a fuzzy c-partition of the data matrix $\mathbf{X}$. From now on we refer to $\mathbf{U}$ as partition matrix and $\mathbf{U}_{ik}$, usually written μ_{ik}, as the membership degree of the observation $\mathbf{x}_k$ in fuzzy cluster i. In Fig. 2b we present an output of the fuzzy c-means algorithm to highlight the differences between two approaches to fuzzy clustering; here the prototypes are represented by filled black circles.

It can be shown that, given $\mathbf{V}$, the optimization of (5) reduces to N independent constrained least squares (CLS) problems [12], which can be solved easily and in parallel. Specifically, minimizing $J_c(\mathbf{U}|\mathbf{X}, \mathbf{V})$ is equivalent to solving

$$\begin{aligned} &\min_{\mathbf{U}_k}\left(\frac{1}{2}\|\mathbf{x}_k - \mathbf{V}\mathbf{U}_k\|^2\right), \quad 1 \le k \le N, \\ &\text{s.t.} \sum_{\mathrm{i}=1}^{\mathrm{c}} \mathbf{U}_{\mathrm{ik}} = 1, \quad 0 \le \mathbf{U}_{\mathrm{ik}} \le 1, \ 1 \le \mathrm{i} \le \mathrm{c}, \end{aligned} \tag{8}$$

regardless of the way the matrix $\mathbf{V}$ is obtained. Here, $\|.\|$ symbolizes the Euclidean L_2-norm. Our concern is therefore the estimation of $\mathbf{V}$, given $\mathbf{U}$, i.e. the minimization of the objective function $J_c(\mathbf{V}|\mathbf{X}, \mathbf{U})$.

If there is no restriction on $\mathbf{V}$, we can estimate this matrix by setting the derivative of the objective function (5), with respect to $\mathbf{V}$, equal to zero, i.e.

$$\frac{\partial J_c}{\partial \mathbf{V}} = -2(\mathbf{X} - \mathbf{V}\mathbf{U})\mathbf{U}^T = 0,$$

which gives

$$\mathbf{V} = \mathbf{X}\mathbf{U}^T\left(\mathbf{U}\mathbf{U}^T\right)^{\dagger}, \tag{9}$$

where $\mathbf{A}^{\dagger}$ symbolically denotes the pseudo-inverse matrix of $\mathbf{A}$. We call this approach unrestricted least squares (ULSQ) solution, and will see that it does not necessarily keep the prototypes close to the data points, as with the proposal by Mirkin and Satarov [30].

An alternative approach to the estimation of $\mathbf{V}$ is proposed by Cutler and Breiman in [12], and referred to as archetypal analysis (AA). Accordingly, the prototypes, now termed archetypes, are constrained to lie in the data space by being convex combinations of the data points, i.e.

$$\mathbf{v}_i = \sum_{k=1}^{N} \beta_{ki}\, \mathbf{x}_k, \quad 1 \le i \le c, \tag{10}$$

or, in matricial form,

$$\mathbf{V} = \mathbf{X}\mathbf{B}, \tag{11}$$

where $\mathbf{B} = [\beta_{ki}] \equiv [\beta_1\ \beta_2 \ldots \beta_c]$, such that $\beta_{ki} \ge 0$ and $\sum_{k=1}^{N} \beta_{ki} = 1$. This confers the status of cluster centroids on prototypes and potentially improves their interpretability. The equation (6) can subsequently be rewritten as

$$\mathbf{X}_{\pm} \approx \mathbf{X}_{\pm}\mathbf{B}_{+}\mathbf{U}_{+},$$

which fits in the framework of convex-NMF [13]. As a consequence, in AA the estimation of the matrix of prototypes $\mathbf{V}$ converts into the estimation of $(N-1) \times c$ of β coefficients. This new perspective of the prototypes strengthens the possibility of performing fuzzy clustering via matrix factorization (see also [4]). In sum: given the matrix $\mathbf{B}$, $\mathbf{V}$ is updated using the relation (11). The optimization process therefore alternates between the estimation of $\mathbf{U}$ and $\mathbf{B}$, and subsequently of $\mathbf{V}$, until convergence. Our study addresses the estimation of β coefficients, and aims to examine how reliable an AA is as a fuzzy clustering tool. We insist, a flat procedure, i.e. (8), can be used to estimate $\mathbf{U}$, regardless of the algorithm adopted for obtaining the matrix $\mathbf{B}$.

2.3 Estimation of β coefficients

In the original work [12], Cutler and Breiman estimate the β coefficients using the so-called archetype algorithm,

given the partition matrix $\mathbf{U}=[\mathbf{U}_{ik}]$. It can be briefly described as follows. Suppose $\mathbf{v}_a$ is the archetype of interest, where $1 \leq a \leq c$; let

$$\mathbf{z}_k = \frac{\mathbf{x}_k - \sum_{i=1, i\neq a}^{c} \mathbf{U}_{ik}\ \mathbf{v}_i}{\mathbf{U}_{ak}}$$

and the *intermediate* archetype

$$\tilde{\mathbf{v}}_a = \frac{\sum_{k=1}^{N} \mathbf{U}_{ak}^2\ \mathbf{z}_k}{\sum_{k=1}^{N} \mathbf{U}_{ak}^2}.$$

A little algebra shows that the minimization of $J_c(\mathbf{V}|\mathbf{X},\mathbf{U})$ is equivalent to finding a solution for each of the following c CLS problems:

$$\begin{aligned} &\min_{\beta_a}\left(\frac{1}{2}\|\tilde{\mathbf{v}}_a - \mathbf{X}\beta_a\|^2\right), \quad 1 \leq a \leq c, \\ &\text{s.t. } \beta_{\mathrm{ka}} \geq 0 \text{ and } \sum_{\mathrm{k}=1}^{\mathrm{N}} \beta_{\mathrm{ka}} = 1, \end{aligned} \tag{12}$$

which are formally similar to (8). After solving (12) for each c archetypes, the $\mathbf{V}$ matrix is updated using formula (11). This approach to the estimation of $\mathbf{U}$ and $\mathbf{V}$ iteratively alternates between two least squares steps, namely (8) and (12), and is referred to in the literature as alternating least squares (ALS) solution [2]. In the same vein, Eugster and Leish [17] propose a slightly different way to calculate the intermediate archetypes $\tilde{\mathbf{v}}_a$; later, we will illustrate how close the two approaches can be.

The proposal by Ding et al. [13] arises from an attempt to recast the classical k-means algorithm in the form of a matrix factorization. The authors claim that this approach generally works better if $\mathbf{U}_{ik}$ are allowed to range over values in $(0,1)$ instead of $\{0,1\}$, which potentially favors fuzzy cluster analysis. To extend from an NMF approach to semi-NMF, i.e. from $\mathbf{X}_{jk} \in \mathbb{R}_0^+$ to $\mathbf{X}_{jk} \in \mathbb{R}$, they consider the positive and negative parts of a given matrix $\mathbf{A}=[\mathbf{A}_{pq}]$, respectively, $\mathbf{A}_{pq}^+$ and $\mathbf{A}_{pq}^-$, where

$$\mathbf{A}_{pq}^+ = \left(|\mathbf{A}_{pq}| + \mathbf{A}_{pq}\right)/2 \text{ and } \mathbf{A}_{\mathrm{pq}}^- = \left(|\mathbf{A}_{\mathrm{pq}}| - \mathbf{A}_{\mathrm{pq}}\right)/2,$$

and deduct the following update rule for β coefficients:

$$\beta_{ki} \leftarrow \beta_{ki}\sqrt{\frac{\left[\left(\mathbf{X}^T\mathbf{X}\right)^+\mathbf{U}^T\right]_{ki} + \left[\left(\mathbf{X}^T\mathbf{X}\right)^-\mathbf{B}\mathbf{U}\mathbf{U}^T\right]_{ki}}{\left[\left(\mathbf{X}^T\mathbf{X}\right)^-\mathbf{U}^T\right]_{ki} + \left[\left(\mathbf{X}^T\mathbf{X}\right)^+\mathbf{B}\mathbf{U}\mathbf{U}^T\right]_{ki}}}, \tag{13}$$

where $\mathbf{A}^T$ is the transpose of the matrix $\mathbf{A}$. This rule can be easily written in matrix form and efficiently implemented in most computers, as the authors claim. We note that the matrix product $\mathbf{X}^T\mathbf{X}$ and the operations involved in calculating its positive and negative parts are performed only once.

A special case of AA is proposed in [37] and is called factorized fuzzy c-means (FFCM) algorithm. In general terms, it aims to perform an FCM clustering using a matrix factorization approach. Here, the archetypes are calculated in a similar way to the prototypes in FCM and, technically, it corresponds to setting the weighting exponent parameter of the latter algorithm to $m = 1$. In formal terms, the β coefficients in FFCM clustering are calculated as follows:

$$\beta_{ki} = \frac{\mathbf{U}_{ik}}{\sum_{k=1}^{N} \mathbf{U}_{ik}}, \quad 1 \leq i \leq c, \tag{14}$$

and consequently the archetypes are obtained by the formula (10). We stress that (14) does not result from an optimization procedure but rather is an effort to bridge the two algorithms and provide practitioners with an alternative way to perform an FCM-like clustering. It is also worth noting that the FFCM algorithm is computationally less demanding than any of the previously reported AA methods since it essentially only requires the estimation of the partition matrix $\mathbf{U}$, i.e. solving the common CLS problems (8).

We may therefore ask about the differences in the way the various methods of performing AA tackle a given data set. There is no simple or single answer to this question. Anyhow, we use a toy data set and illustrate in Fig. 3 what kind of data decomposition practitioners could expect from these methods, considering a $c = 4$ cluster solution. For reasons that will soon become clear, we omit the solution provided by the method proposed by Eugster and Leish [17]; however we include the ULSQ solution (Fig. 3a) for comparison purposes. A snapshot shows that the latter solution takes the concept of convex hull too literally, and elucidates what is meant by the ideal types not being "close enough to the observed data points" [34]. The restriction on prototypes (10) that gives rise to AA results in a smoother version of the unrestricted solution (Fig. 3b), and a simple visual inspection allows us to conclude that it provides more insightful data decomposition. At the other extreme, the proposal by Ding et al. [13] and the FFCM algorithm (Figs. 3c, 3d) attempt to place the archetypes within data clouds, according to the methodological principle behind them, i.e. reproducing k-means or FCM, respectively. Apparently, the polytope generated by FFCM is more stretched.

In continuing our preliminary study, we also notice the similar behavior of the proposals by Cutler and Breimen [12] and by Eugster and Leish [17]. Figure 4 shows how close the solutions provided by the two approaches to fuzzy clustering are. We also display graduated axes to highlight their quantitative similarity. In our massive empirical analysis, we therefore decided to omit the latter one to save computational time.

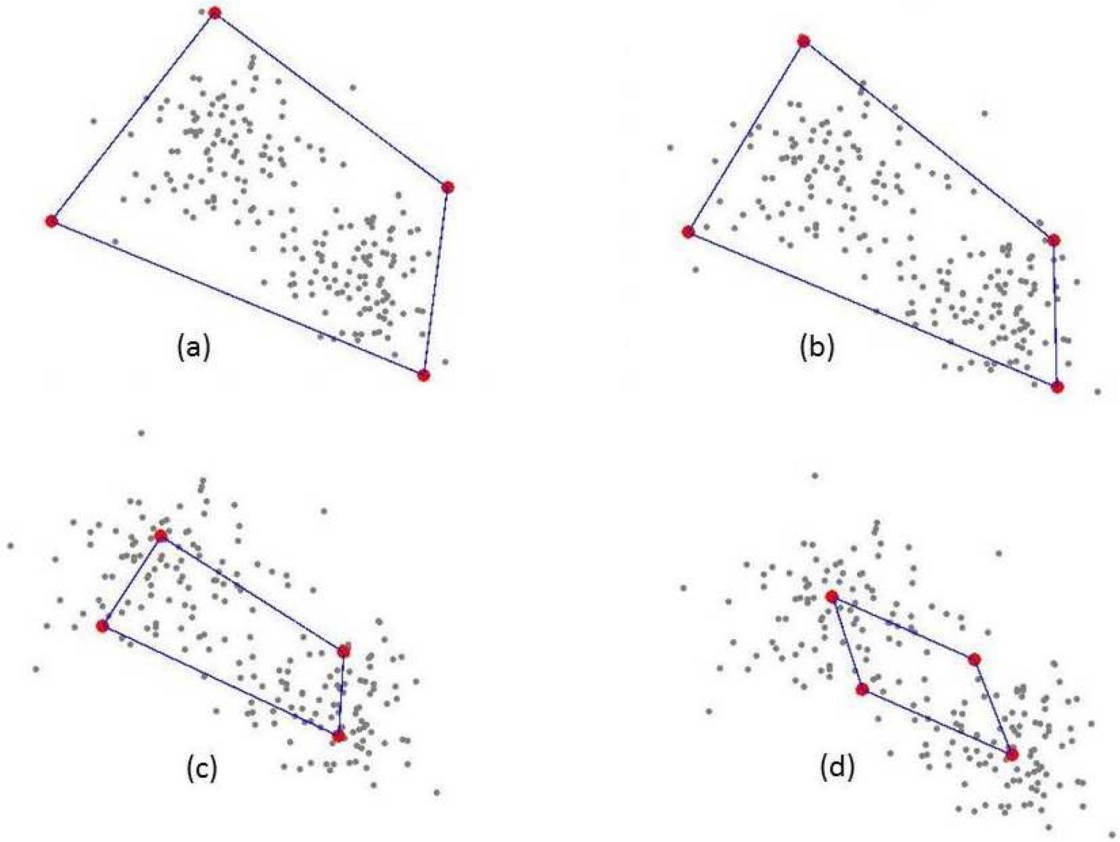

Fig. 3 Artificial data decomposition in $c = 4$ clusters: (a) ULSQ; (b) Cutler and Breimen [12]; (c): Ding et al. [13]; (d) FFCM

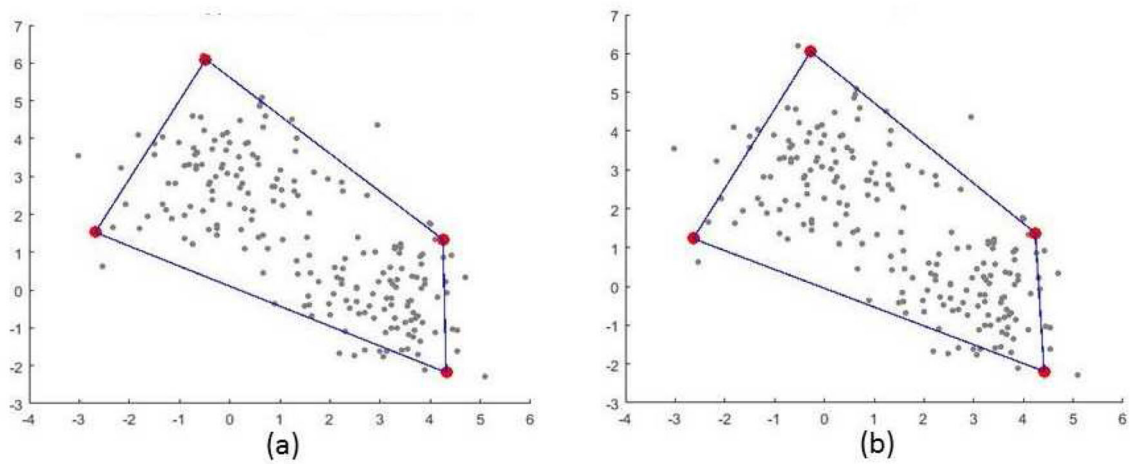

Fig. 4 Decomposition of artificial data in $c = 4$ clusters: (a) Cutler and Breiman; (b) Eugster and Leish

We end this section with a reflection on objective function (5). Unlike FCM, where the objective function is a measure of the compactness of fuzzy clusters (see, e.g., [48]), in matrix factorization approach, the function (5) is used to assess how the polytope generated by the archetypes envelops the data points. To make this point clearer, we present some numerical values associated with the estimation of **U** and **V** behind Figs. 3 and 4. Table 1 displays the number of iterations each method took until convergence and the respective value of the objective function J_4 (5). Looking simultaneously at these figures, we realize that the polytope produced by the ULSQ solution covers almost all data points and, consequently, yields the lowest value of J_4. Its two smoother versions, Cutler

Table 1 The value of the objective function and the number of iterations until convergence for the toy data set ($c = 4$)

Method	Iterations	J_4
ULSQ	22	1.1
Cutler and Breimen	15	5.3
Eugster and Leish	14	5.5
Ding et al.	22	89.2
FFCM	19	186.0

and Breimen and Eugster and Leish, yield higher values of J_4, since more data points lie outside the polytopes they generate. Here, too, the close relationship between these two approaches to AA is evident. These results help understand why the remaining two approaches provide much greater values of the objective function.

Although this simple illustrative example gives an indication of the behavior of different methods, we can hardly anticipate the role played by different analytical forms for obtaining archetypes, (12), (13) or (14), in a fuzzy cluster analysis. We therefore opt for a stochastic simulation study to evaluate their behavior in different clustering scenarios, which is covered in the next section.

3 Empirical Study

3.1 Strategy

To examine the behavior of different forms of AA, we use artificial data to perform a simulation study, and then put them to the test with some data sets taken from real life problems. The results of this latter test are given in Sect. 3.5; first we explain the simulation work in detail.

The flow chart of Fig. 5 systematizes the way the simulation is conducted. First, we need a polytope or a data generator, which is characterized by $c^* \geq 2$ extreme points or vertices, $\mathbf{v}_1, \mathbf{v}_2, ...,$ and $\mathbf{v}_{c^*}$. This conforms to the underlying assumption of the matrix factorization approach to fuzzy clustering, in that the data are drawn from a polytope with c^* extreme points or vertices. Technically these vertices are the columns of $\mathbf{V}$ matrix. Samples of membership degrees in each c^* fuzzy cluster, i.e the partition matrix $\mathbf{U}$, give rise to pure or noiseless data, by means of the product $\mathbf{P} = \mathbf{VU}$, that is a sample of the convex hull of $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_{c^*}$. A noise component is added to $\mathbf{P}$ to mimic a real world environment, and we have the data matrix $\mathbf{X}$.

The matrices $\mathbf{U}$ and $\mathbf{V}$ are estimated from the data $\mathbf{X}$; the corresponding estimates are denoted by $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$, respectively. The calculation of $\hat{\mathbf{V}}$ is preceded by the estimation

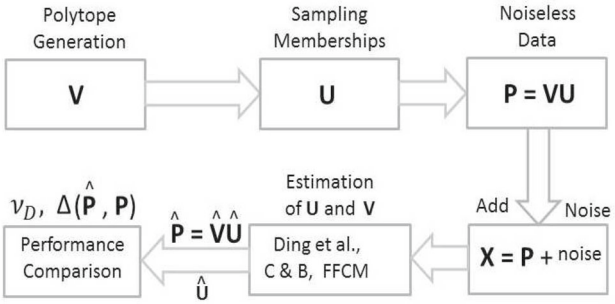

Fig. 5 Flow chart of the simulation study (C&B means Cutler and Breiman; the meaning of the performance measures $\Delta(\hat{\mathbf{P}}, \mathbf{P})$ and υ_D is given below)

of β coefficients. The matrix $\mathbf{U}$ is estimated solving N independent least squares problems as expressed in (8). The final step is to evaluate the performance of different approaches as fuzzy clustering tools and their ability to recover the original matrix $\mathbf{P}$, as indicated in Fig. 5.

Although the term *clustering* is consensually accepted by research communities as the "methods for grouping of unlabeled data" [23], there are several issues associated with the variety of structures hidden in multidimensional data sets. These include, among others, clusters' shape, their spatial distribution and density, besides the number of clusters in data and its assessment and, of course, the space dimension. For example, the FCM algorithm may become less reliable when addressing high dimensional data [46]. In this study, we confine ourselves to low dimensional spaces, *concentration* degrees and class imbalance; the goodness-of-fit of cluster solutions provided by different algorithms is assessed by means of a generalized Dice index [22] and the reconstruction accuracy, as explained in Sect. 3.3.

3.2 Artificial Data Sets

We consider six space dimensions, $n = 2, 3, ..., 7$, and for each dimension we generate nine different cluster structures, $c^* = 2, 3, ..., 10$. Two software tools are used to construct the matrix of prototypes $\mathbf{V}$: *polymake* [21] for $c^* > n$ and our own software otherwise. In any case, the prototypes, $\mathbf{v}_1, \mathbf{v}_2, ...,$ and $\mathbf{v}_{c^*}$, are located on the unit (hyper)sphere of $\mathbb{R}^n$, centered at the origin. The partition matrix $\mathbf{U} = [\mathbf{U}_{ik}]$ is constructed according to the following procedure. The membership degrees $\mathbf{U}_{ik}$ are generated from $[0, 1]$ uniform distribution, with four threshold levels for the belongingness in clusters: $\gamma = 0.95, 0.85, 0.75,$ or 0.65. For example, $\gamma = 0.75$ means $\min \mathbf{U}_{ik} = 0.75$, i.e. the membership degree of $\mathbf{x}_k$ in fuzzy cluster i is, at least, 0.75. This allows different concentration degrees in clusters. The sample size is equal to $N = 50 \times c^*$; the partition matrix is randomly replicated 15 times for each value of γ, and each replicate gives rise to a noiseless data set

$$\mathbf{P} = \mathbf{VU}. \tag{15}$$

At this point, we stress that the ground truth is fuzzy rather than crisp since we have prior knowledge of $\mathbf{U}$. Therefore, we will end by comparing two fuzzy partitions: the true partition represented by $\mathbf{P}$, as in (15), and the partition estimated by different algorithms,

$$\hat{\mathbf{P}} = \hat{\mathbf{V}}\hat{\mathbf{U}}, \tag{16}$$

where $\hat{\mathbf{V}}$ and $\hat{\mathbf{U}}$ are the estimates of $\mathbf{V}$ and $\mathbf{U}$, respectively.

The data matrix $\mathbf{X}$ is constructed by contaminating $\mathbf{P}$ with additive gaussian $(\mathbf{0}, \sigma\mathbf{I})$ noise, for $\sigma = 0.001, 0.01$ or 0.05, where $\mathbf{I}$ is the $n \times n$ identity matrix; so

$$\mathbf{X} = \mathbf{P} + \mathbf{E}, \tag{17}$$

where $\mathbf{E}$ symbolizes the error term which is simulated by the referred gaussian noise. In sum, $\mathbf{X}$ is decomposed in $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$, and the accuracy of this decomposition is measured by comparing $\hat{\mathbf{P}}$ to $\mathbf{P}$ and also $\hat{\mathbf{U}}$ and $\mathbf{U}$, in terms that will soon be clear.

To evaluate the effect of class imbalance, we consider three different class densities: equal size, 10% and 60% density. A 60% density means that one cluster has $0.6 \times N$ data points and the remaining $0.4 \times N$ are evenly spread over other $c^* - 1$ clusters. The same rationale applies to 10% density. In total, we have six dimensions, nine cluster structures, four membership thresholds, three noise levels, three densities which amount to $1,944$ cluster contexts each one being replicated 15 times, totalling $29,160$ artificial data sets. For each cluster context, a 16th data set provides for a flat initialization of all algorithms.

All calculations were performed in a MATLAB environment. Using our non-optimized code, the total computational time of this simulation was about nine months. We limited the error term, i.e. maximum absolute difference between two membership degrees in consecutive iterations, to 0.01. The number of clusters in the data ranged between $c = 2$ to $c = c_{\max} = \max\{8, 1.5 \times c^*\}$. All point-based graphics representing simulation outcomes are smoothed using the MATLAB *smooth()* function.

3.3 Assessing the Goodness-of-fit

Now we address the measurement of the discrepancy between $\hat{\mathbf{P}}$ and $\mathbf{P}$, $\Delta(\hat{\mathbf{P}}, \mathbf{P})$, and between $\hat{\mathbf{U}}$ and $\mathbf{U}$ (Fig. 5). For the latter case, we use the fuzzy generalization of the Dice index proposed by Hüllermeier et al. [22], which has proven effective in comparing data partitions [40]; we denote the underlying measure by υ_D. We use the quantity $1 - R$, called reconstruction accuracy (RA), where R is given by

$$R = \frac{\left\|\hat{\mathbf{P}} - \mathbf{P}\right\|_F}{\|\mathbf{P}\|_F}, \tag{18}$$

to assess the ability of $\hat{\mathbf{P}}$ to reproduce $\mathbf{P}$, that is $\Delta(\hat{\mathbf{P}}, \mathbf{P})$; the subscript F in (18) indicates the Frobenius norm. Next we give a brief account of how υ_D works. We note that υ_D needs prior information about the cluster structure of the data and is therefore called external index.

Suppose we have two crisp partitions of a given data set $\mathbf{X}$: Λ_1 and Λ_2, where Λ_1 is the reference partition and Λ_2 is an algorithmically-generated partition of $\mathbf{X}$. There is no need for the number of clusters of Λ_1, c_1, be equal to that of Λ_2, c_2, although the match $c_1 = c_2$ is always appealing. The aim is to evaluate how Λ_2 mimics Λ_1. There are four quantities involved in this process based on $\binom{N}{2}$ pairwise comparisons of data points:

N_{11} : number of pairs of data points grouped in the same cluster in Λ_1 and in the same cluster in Λ_2;
N_{12} : number of pairs of data points grouped in the same cluster in Λ_1 and in different clusters in Λ_2;
N_{21} : number of pairs of data points grouped in different clusters in Λ_1 and in the same cluster in Λ_2; and
N_{22} : number of pairs of data points grouped in different clusters in Λ_1 and in different clusters in Λ_2.

The Dice index is a function of the two partitions that are being compared and is given by the quotient

$$\upsilon_D \equiv \upsilon_D(\Lambda_1, \Lambda_2) = \frac{2 \times N_{11}}{2 \times N_{11} + N_{12} + N_{21}}, \tag{19}$$

and its range is the unit interval $[0, 1]$: the higher the value of the index, the greater the efficiency of the algorithm. Now suppose that Λ_1 and Λ_2 are instead two fuzzy partitions of $\mathbf{X}$, and we want to know how these four quantities can be written in terms of membership degrees. The starting point is the equivalence relation on $\mathbf{X}$, by means of a similarity measure between data points as expressed by the respective membership degree vectors,

$$E_\Lambda(\mathbf{x}_k, \mathbf{x}_{k'}) = 1 - \frac{1}{2}\sum_{i=1}^{c} |\mathbf{U}_{ik} - \mathbf{U}_{ik'}|, \quad 1 \le k, k' \le N, \tag{20}$$

where c is the number of clusters in a generic partition Λ, and $\mathbf{U}_{ik}$ and $\mathbf{U}_{ik'}$ are, respectively, the kth and k'th columns of the corresponding partition matrix $\mathbf{U}$. So all quantities involved in the right hand side of (20) should be replaced according to the partition for which it is calculated. Hüllermeier et al. [22] propose the following fuzzy counterparts of the four quantities referred to above:

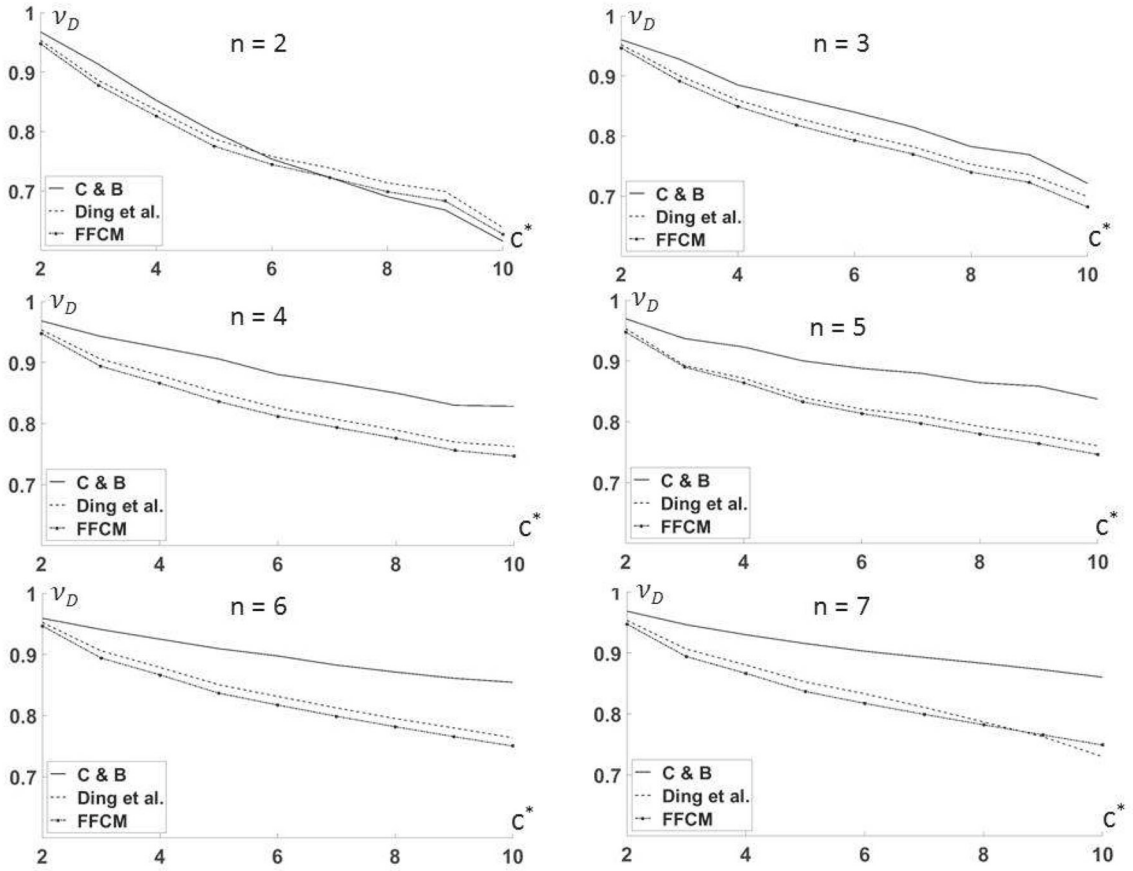

Fig. 6 Clustering accuracy of the three methods of performing archetypal analysis as assessed through the Dice index υ_D, in function of the number of clusters in the data, c^*, for different dimensions and equal density

$$N_{11}(k,k') = (1-|u-v|)\cdot u\cdot v,$$
$$N_{12}(k,k') = \max(E_{\Lambda_1}(\mathbf{x}_k,\mathbf{x}_{k'}) - E_{\Lambda_2}(\mathbf{x}_k,\mathbf{x}_{k'}),0),$$
$$N_{21}(k,k') = \max(E_{\Lambda_2}(\mathbf{x}_k,\mathbf{x}_{k'}) - E_{\Lambda_1}(\mathbf{x}_k,\mathbf{x}_{k'}),0),$$
$$N_{22}(k,k') = (1-|u-v|)\cdot(1-u\cdot v),$$

where $u = E_{\Lambda_1}(\mathbf{x}_k,\mathbf{x}_{k'})$ and $v = E_{\Lambda_2}(\mathbf{x}_k,\mathbf{x}_{k'})$. These quantities are alternatively used in (19) to calculate a fuzzy version of υ_D. By virtue of its construction, in our case, the reference partition Λ_1 is also fuzzy. This is very unusual; for instance, the class memberships of real life data sets we use in Sect. 3.5 are crisp, i.e. either fully belong to a cluster or do not.

The question now is how to select a fuzzy partition of data set, given a cluster context. First we note that the number of clusters of Λ_1, c_1, is equal to c^*, and it is known a priori. The partition Λ_2 is selected among all fuzzy c-partitions attempted for the data set $\mathbf{X}$, and c_2 is determined by solving

$$c_2 = \max_c \upsilon_D\left(\Lambda_1, \Lambda_2^{(c)}\right), \quad c = 2,3,\ldots,c_{\max},$$

where $\Lambda_2^{(c)}$ represents the second partition with c clusters. In our study, $c_{\max} = \{8, 1.5\times c^*\}$ as referred above. The fuzzy c_2-partition, i.e. $\Lambda_2^{(c_2)}$ is used to calculate $\hat{\mathbf{P}}$ to assess the reconstruction accuracy (18) of every algorithm and the quantity $\upsilon_D\left(\Lambda_1, \Lambda_2^{(c_2)}\right)$ is the algorithm's relative rank for the cluster context in question. For inferential purposes, we use the average values associated with the optimal partitions over 15 replicates of a given cluster context.

3.4 Empirical Evidence

It is not easy to portray the results from an experiment when it produces a huge amount of information. We decided to combine the outcomes related to different class membership and noise contamination, focusing mainly on the effects of the space dimension, n, and the cluster structure of the data, c^*, for each class density under study: equal, 10% and 60%. First, we elaborate on how the three methods of performing an AA, Cutler and Breiman (C&B), Ding et al. and FFCM behave as fuzzy clustering tools, and then address in Sect. 3.4.4 their reconstruction accuracy.

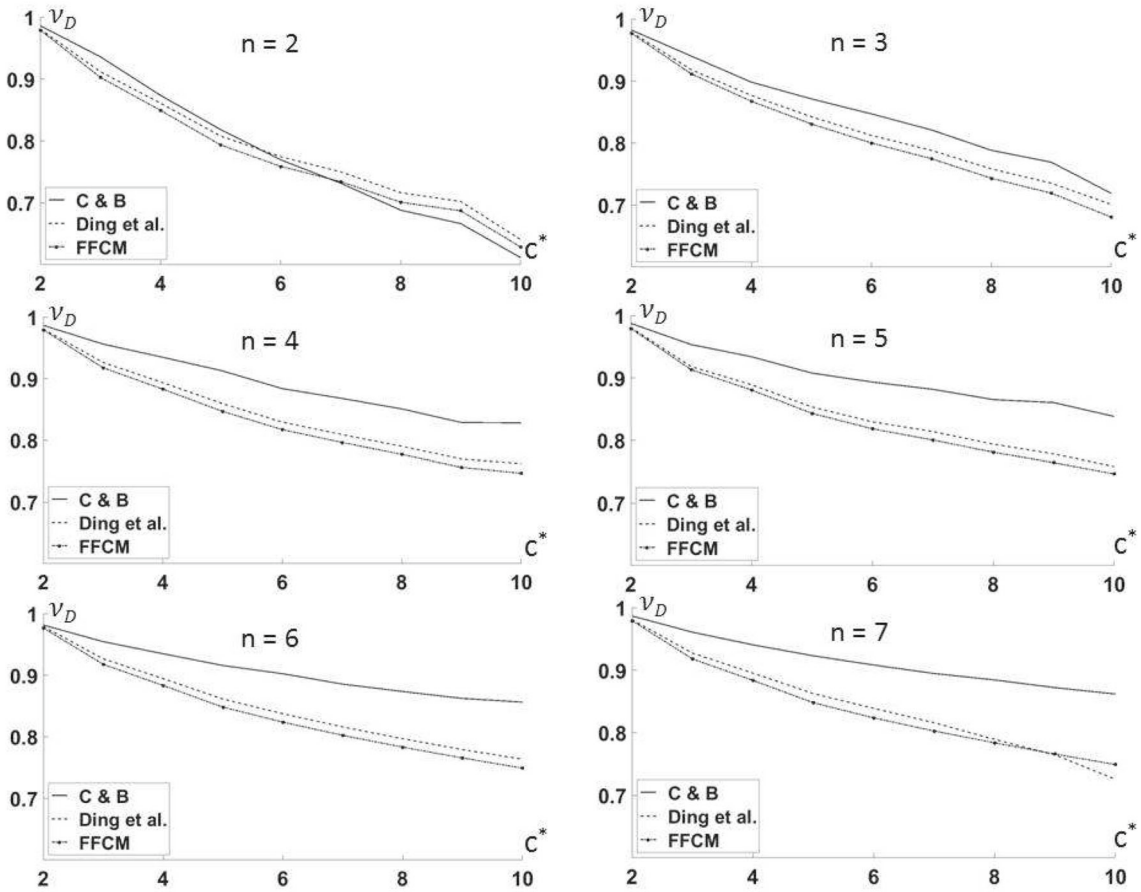

Fig. 7 Clustering accuracy of the three methods of performing archetypal analysis as assessed through the Dice index υ_D, in function of the number of clusters in the data, c^*, for different dimensions and 10% density

3.4.1 AA as a Clustering Tool

With the exception of the case $n = 2$, we realize that the C&B method generally gives rise to fuzzy partitions that better express the cluster structure of the data than the other two, as assessed by υ_D (Figs. 6, 7 and 8). In the case of equal (Fig. 6) and 10% density (Fig. 7), the discrepancy apparently becomes sharper as n increases, where a degradation of the performance with c^* is also evident, regardless of the value of n. When it comes to 60% density class (Fig. 8), the behavior of υ_D seems to be stable and almost constant from $n = 4$, and leads us to believe that the clustering is more accurate here. In the same vein, a closer look at the behavior of υ_D in the cases of equal and 10% density (see Figs. 6 and 7, respectively) reveals a better performance in the latter case when the number of clusters c^* is low, but it tends to be similar to the former as long as $c^* \to 10$. This is consistent with what has just been said, since the 10% class starts to be more underrepresented and steadily increases its representation with c^*, and the data are fully balanced when $c^* = 10$. Later, we return to the subject of data imbalance in detail. In any case, the proposal by Ding et al. and the FFCM run closely in parallel; however, the former outperforms the latter for almost all values of n. Regarding $2D$ data, the three methods perform almost similarly for all values of c^*.

3.4.2 Comparing Mean Differences

Although the graphical information of υ_D reflects the way different data analysis methods act as clustering tools, it is advisable to find an analytical device that allows us to verify how significantly these methods differ from each other. This may also lead practitioners to make a more confident choice for one or other method. Therefore, we conducted a single factor ANOVA to test

$$H_0 : \mu_{\upsilon_D}^{(1)}(n,d) = \mu_{\upsilon_D}^{(2)}(n,d) = \mu_{\upsilon_D}^{(3)}(n,d),$$
$$n = 2\ldots7;\ d = \text{equal, } 10\%, \text{ } 60\%$$
$$H_1 : \text{atleastone } \mu_{\upsilon_D}^{(\mathrm{i})}(\mathrm{n},\mathrm{d}) \neq \mu_{\upsilon_D}^{(\mathrm{j})}(\mathrm{n},\mathrm{d}),$$

where $\mu_{\upsilon_D}^{(1)}$, $\mu_{\upsilon_D}^{(2)}$ and $\mu_{\upsilon_D}^{(3)}$ are the true mean values of υ_D related to C&B, Ding et al. and FFCM, respectively, for each dimension and density. A $\alpha = 0.10$ significance level

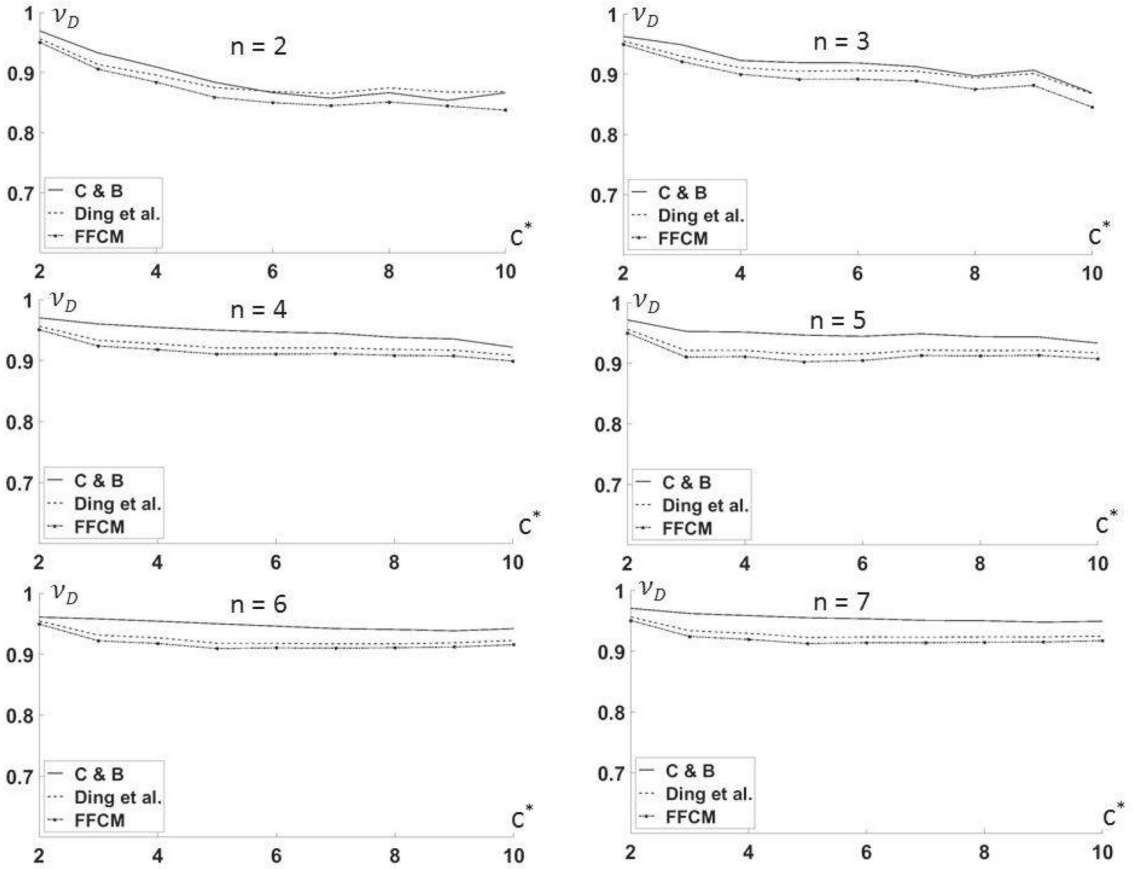

Fig. 8 Clustering accuracy of the three methods of performing archetypal analysis as assessed through the Dice index υ_D, in function of the number of clusters in the data, c^*, for different dimensions and 60% density

was set for all tests. Where a significant difference was found, the specific differences between the AA methods were further examined by a post-hoc analysis using Tukey's HSD test. Table 2 summarizes the results of the ANOVA test, which also includes the *F*-statistic and the associated *p*-value for every case. In some cases, overlapping took place; this is signaled by the letters *A* and *B*, in this table. Consider, for example, the dimension $n = 4$ and equal density: the Ding et al. method can be grouped with both C&B, *A*, or FFCM, *B*.

We fail to reject the Levene's test for equality of variances in all cases and the normality assumption for equal and 10% density data. The latter does not hold in most cases of 60% density data; this might not be critical, since ANOVA is considered a robust test against that assumption.

Looking at Table 2, we realize that when υ_D's significantly differ on average, the C&B proposal provides better clustering accuracy; this also generally holds for $n > 2$ even when it is not significantly different than that of the Ding et al. method. We also notice the preponderance of C&B becomes more evident as n increases. So it may be recommended as the first choice in performing archetypal analysis of the data drawn from polytopes.

3.4.3 Imbalanced Data

Now we address the subject of imbalanced data, in particular, the data containing one class with 60% density. Looking again at Fig. 8, it is surprising to note that every method performs better in the presence of imbalanced data, when we know that class imbalance is an issue in data mining research [6, 20, 45, 49]. We therefore need to examine how the overrepresented class, i.e. 60% density, influences the clustering accuracy υ_D. It is known that, in such cases, several factors have interdependent effects on minority classes, leading them to somehow be ignored; nevertheless, the validity criterion can give the illusion of good clustering [19, 20]. To measure the contribution of the dominant class to υ_D, we compare the number of data points that are originally grouped together in the reference partition and those that were algorithmically grouped by different AA methods. This task can be accomplished by simply restricting the calculation of N_{11} term in the numerator of (19), to the subset of data points that belong

Table 2 Summary of single-factor ANOVA comparing the mean values of v_D associated with (1): C&B, (2): Ding et al. and (3): FFCM

n		Density		
		Equal	10%	60%
2	$\hat{\mu}^{(1)}$	0.77	0.79	0.88
	$\hat{\mu}^{(2)}$	0.78	0.79	0.88
	$\hat{\mu}^{(3)}$	0.76	0.78	0.86
		$F_{2,24}$=0.04	$F_{2,24}$=0.03	$F_{2,24}$=0.46
		$p = 0.96$	$p = 0.97$	$p = 0.64$
3	$\hat{\mu}^{(1)}$	0.84	0.85	0.92
	$\hat{\mu}^{(2)}$	0.81	0.82	0.91
	$\hat{\mu}^{(3)}$	0.80	0.81	0.89
		$F_{2,24}$=0.62	$F_{2,24}$=0.48	$F_{2,24}$=0.92
		$p = 0.55$	$p = 0.62$	$p = 0.41$
4	$\hat{\mu}^{(1)}$	$0.89^{*(A)}$	0.89	0.95^{*}
	$\hat{\mu}^{(2)}$	$0.83^{(A,B)}$	0.84	0.92
	$\hat{\mu}^{(3)}$	$0.82^{(B)}$	0.83	0.91
		$F_{2,24}$=2.06	$F_{2,24}$=1.96	$F_{2,24}$=13.63
		$p = 0.07$	$p = 0.16$	$p = 0.00$
5	$\hat{\mu}^{(1)}$	0.89^{*}	$0.90^{*(A)}$	0.95^{*}
	$\hat{\mu}^{(2)}$	0.83	$0.84^{(A,B)}$	0.92
	$\hat{\mu}^{(3)}$	0.82	$0.83^{(B)}$	0.91
		$F_{2,24}$=4.35	$F_{2,24}$=2.89	$F_{2,24}$=11.54
		$p = 0.02$	$p = 0.05$	$p = 0.00$
6	$\hat{\mu}^{(1)}$	0.90^{*}	$0.90^{*(A)}$	0.95^{*}
	$\hat{\mu}^{(2)}$	0.84	$0.85^{(A,B)}$	0.92
	$\hat{\mu}^{(3)}$	0.83	$0.83^{(B)}$	0.91
		$F_{2,24}$=4.69	$F_{2,24}$=4.69	$F_{2,24}$=18.24
		$p = 0.02$	$p = 0.02$	$p = 0.00$
7	$\hat{\mu}^{(1)}$	0.91^{*}	0.91^{*}	0.95^{*}
	$\hat{\mu}^{(2)}$	0.84	0.84	0.93
	$\hat{\mu}^{(3)}$	0.83	0.84	0.92
		$F_{2,24}$=5.59	$F_{2,24}$=3.79	$F_{2,24}$=29.02
		$p = 0.01$	$p = 0.04$	$p = 0.00$

The quantity $\hat{\mu}$ is an estimate of the true value μ. The * means significant at 10% level. The overlapping groups are signaled by the letters A and B

to the dominant cluster. Dividing the value thus obtained by v_D, gives the desired result.

The results achieved show no noticeable difference between different methods and the influence of the over-represented cluster across c^* is almost the same regardless of the value of n. We therefore select one value of n $(n = 2)$ and one method (C&B) and in Fig. 9 give a general idea about the relative contribution of 60% density cluster to v_D. The contribution increases with c^*, i.e. when the data become more imbalanced, and can be higher than 90%. As a consequence, we conclude that the class imbalance also remains an issue in the context of AA.

3.4.4 Reconstruction of Data

Fuzzy cluster analysis can be seen as a technique of information granulation and used as a mechanism of encoding and decoding data, which altogether fit in the framework of a reconstruction problem, as Pedrycz and Oliveira note in [36]. In simple terms, the problem involves grouping the data in fuzzy clusters (encoding) and rebuilding the original data set from these clusters (decoding). The previously cited authors explore the reconstruction problem in the context of Bezdek's FCM algorithm [3]; we do so in the context of matrix

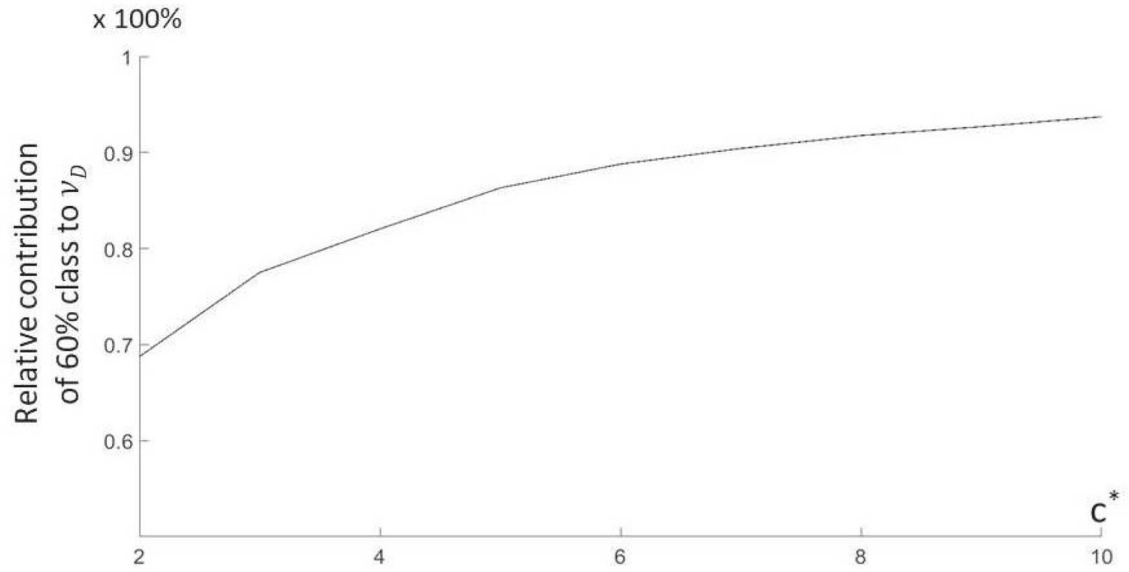

Fig. 9 Influence of the overrepresented cluster on clustering accuracy

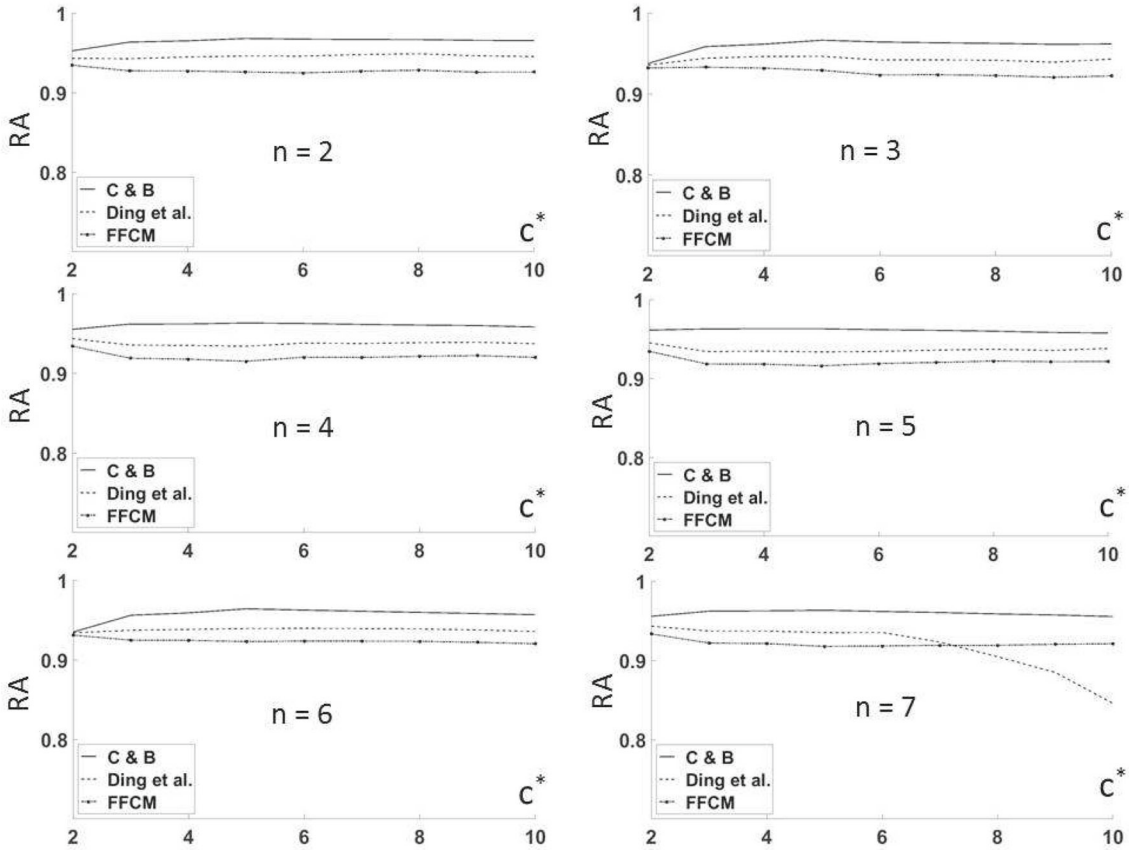

Fig. 10 Reconstruction accuracy (RA) of archetypal analysis in function of c^*, given n, for equal density data

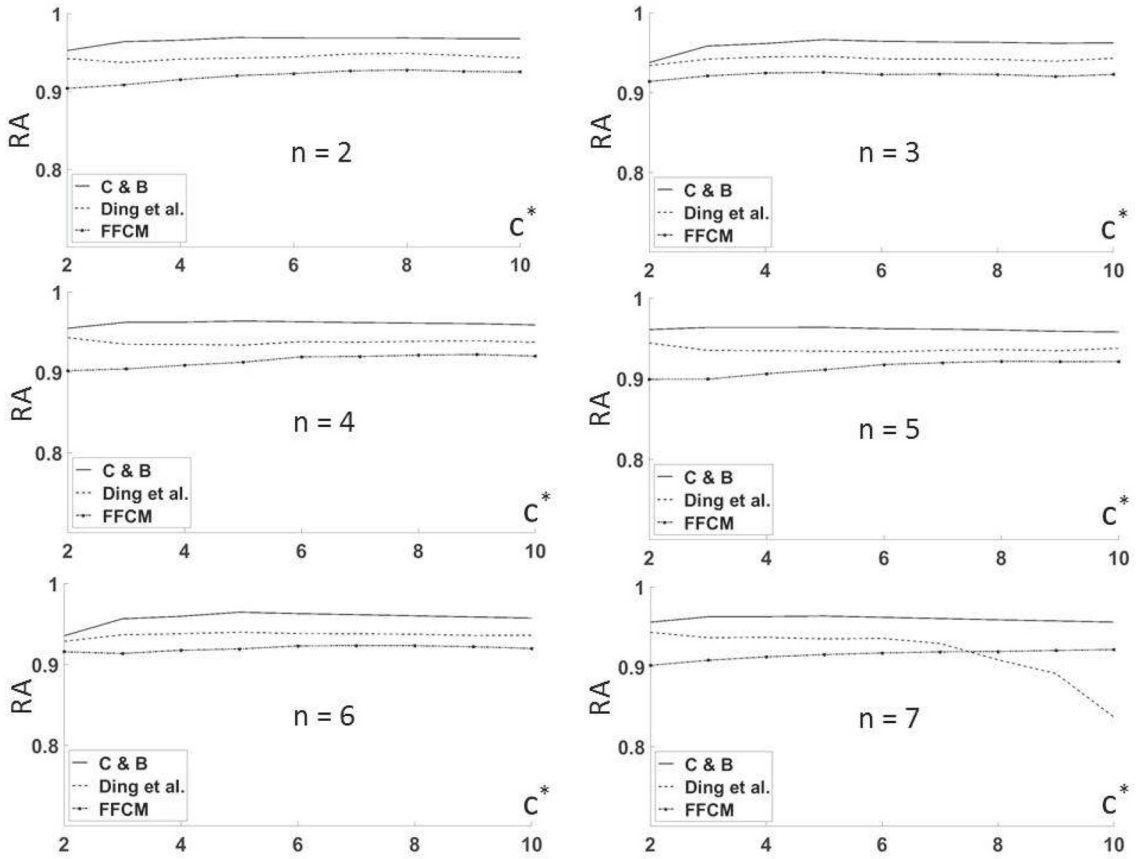

Fig. 11 Reconstruction accuracy (RA) of archetypal analysis in function of c^*, given n, for 10% density data

factorization, $\mathbf{X} \simeq \mathbf{VU}$ (see [27] for a Bayesian approach to this problem, including a comparison to a classical k-means algorithm).

We use the outcomes of our simulation study to also assess how AA behaves in the decoding process. The reconstruction accuracy (RA), $1 - R$, where R is given by (18), is the performance index we use for this purpose. The results of our experiment are displayed in Figs. 10, 11 and 12, for equal, 10% and 60% density data, respectively.

Here too, the original method of estimating the archetypes, C&B, provides more reliable results compared to the ones given by the other two methods, Ding et al. and FFCM and, apparently, is fairly robust to class imbalance, space dimension and the number of clusters in the data. However, the latter algorithm seems more sensitive to class imbalance, since it generally provides better results as the number of clusters c^* increases, in the case of 10% density; on the other hand, its accuracy in the decoding process tends to degrade with this factor when the data possess one dominant cluster. Our attention is also drawn to the behavior of the Ding et al. proposal, in particular, for $n = 7$ and equal and 10% density data. Although we cannot take the effect of n on it for granted, this outcome opens up opportunities to continue the investigation on this subject. Nevertheless, it generally performs better than FFCM.

As before, we conducted an ANOVA test to assess how the reconstruction accuracy provided by the three methods under study differs on average from each other. Excepting the case $n = 7$ and equal and 10% density data, the C&B is ranked first, the Ding et al. second and FFCM third. In the exceptional case, the last two are statistically ranked second. This could somehow be anticipated as the respective RA curves twist at $c^* = 7$ (Figs. 10 and 11).

We note that, in this ANOVA test, the normality assumption is violated in almost all cases and the Levene test for equality of variances is rejected in several cases. The achievements should therefore be interpreted with caution.

3.5 Real-life Datasets

As the final stage of our experiment, we put the three methods of performing AA to the test with data sets from real life problems. Here, we also consider the FCM algorithm, with the weighting exponent parameter $m = 2$, for comparison purposes. The eight data sets presented in

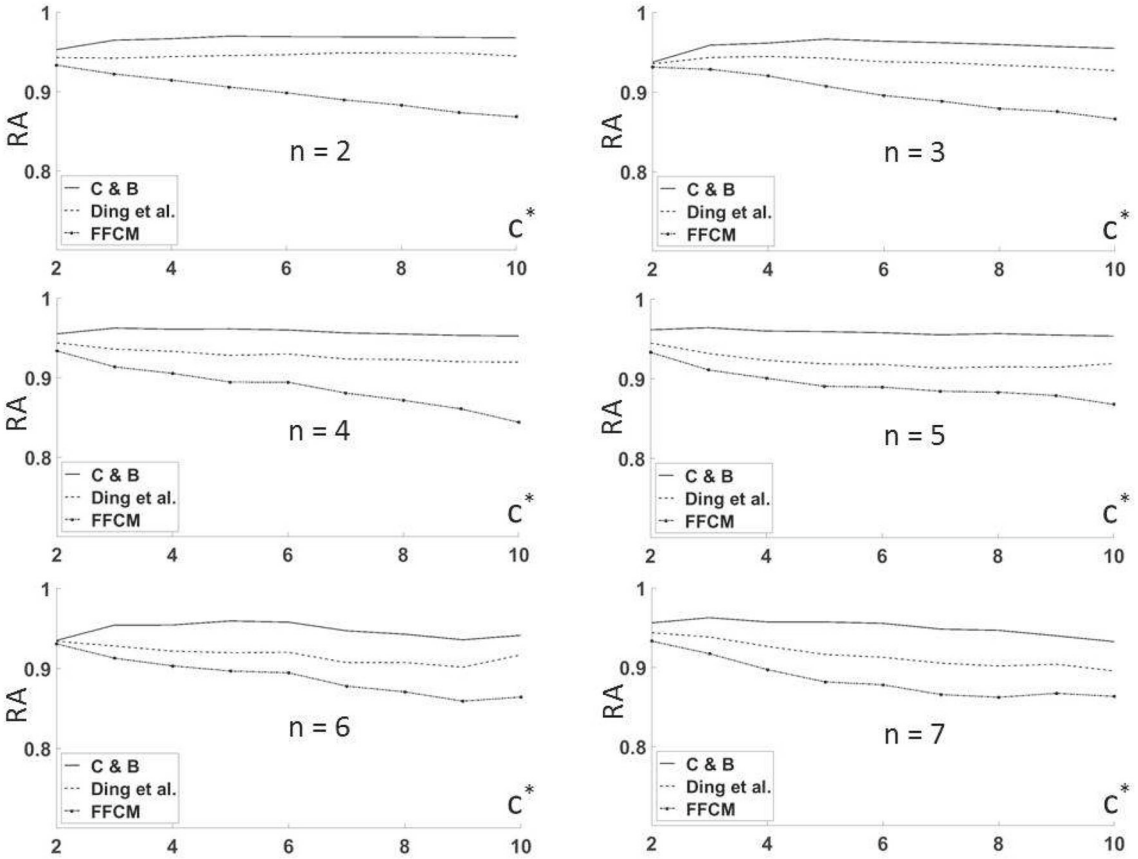

Fig. 12 Reconstruction accuracy (RA) of archetypal analysis in function of c^*, given n, for 60% density data

Table 3 Performance of different AA methods and of the FCM algorithm for data from real life problems

				υ_D			
Data set	N	n	c^*	C&B	Ding et al.	FFCM	FCM
Forest type	523	4	4	0.46	0.50	**0.51**	0.46
Glass identification	214	9	6	0.41	**0.46**	**0.46**	0.44
Glass W – NW	214	9	2	0.79	**0.81**	**0.81**	0.79
Haberman's survival	306	3	2	**0.57**	0.51	0.51	0.53
Hill–Valley (with noise)	606	100	2	**0.61**	0.60	0.59	**0.61**
Iris	150	4	3	0.61	**0.69**	**0.69**	**0.72**
Seeds	210	7	3	0.56	0.63	**0.66**	0.63
Wisconsin breast cancer	683	9	2	0.81	0.86	**0.88**	0.84

The data set referred to as Glass W–NW is the same as Glass Identification, though the glass is instead categorized into Window and Non-window type. In the case of the Wisconsin Breast Cancer data set, the 16 observations with missing values were omitted

Table 3 were downloaded from UCI Machine Learning Repository [15], and are devoted to classification problems. Here too we use υ_D (19) to assess the goodness-of-fit of the estimated fuzzy partitions. We note, however, that the membership degrees of the ground-truth partition Λ_1 belong in this case to $\{0, 1\}$, while those of Λ_2 are fuzzy, $[0, 1]$. So for real life data, υ_D will be confronting a crisp partition with a partition that is fuzzy. In Table 3, the highest values of υ_D are in bold to highlight the AA method that better unveils the cluster structure of the data. This also

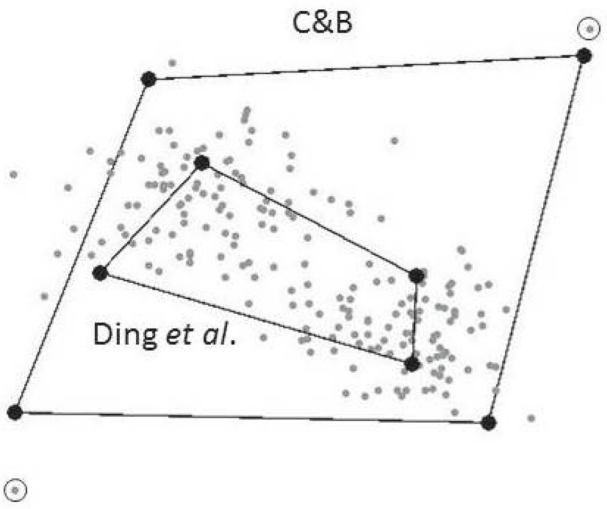

Fig. 13 Archetypal analysis of an artificial data set with two outliers (circled points) using C&B and Ding et al. methods, for $c = 4$

applies to FCM when this algorithm attains the highest value of υ_D.

Overall, and unlike when we used artificial data, there is no clear evidence of C&B supremacy. We note that, in most cases, the Ding et al. proposal and, particularly, FFCM outperform any other. But to what extent is this surprising and how far is it from expectations? In an attempt to answer these questions, we go back to equation (10), i.e. $\mathbf{v}_i = \sum_{k=1}^{N} \beta_{ki}\, \mathbf{x}_k$, and try to understand the strategy underlying different approaches to AA to estimate β coefficients and, consequently, the $\mathbf{V}$ matrix. For Cutler and Breiman [12], and in sequel [17], the constraint on archetypes to be mixture of observations serves to avoid them being "wholly mythological"; however, they are kept as possible on the boundary of the convex hull of the data. In this way, they prevent the archetypes being far from the observations, thus circumventing the drawback of the original work [30] to which Nascimento and Mirkin [34] draw attention. The work by Ding et al. [13] explores the averaging nature of archetypes (10), which captures the cluster centroid notion, to recast the k-means clustering in the form of matrix factorization; as a possible consequence the archetypes are pushed to "center" (see again Fig. 3). By allowing the entries of the partition matrix $\mathbf{U}$ to lie in $[0, 1]$ interval, this approach potentially gives rise to a fuzzy clustering. In turn, Suleman [37] provides his method with an FCM flavor by using the same procedure as the latter algorithm to calculate cluster centroids, here archetypes. This may explain why these two approaches [13, 37] obtain better results than those of [12], when addressing clustered data.

4 Concluding Remarks

We have presented the results of a simulation study carried out to examine the behavior of three methods for performing archetypal analysis (AA), in the context of fuzzy clustering: the original proposal by Cutler and Breiman [12], the method proposed by Ding et al. [13] and the factorized fuzzy c-means (FFCM) algorithm [37]. Our experimental design included factors that can impact clustering, namely the space dimension n, the number of clusters in the data c^*, the degree of membership, and class density; the latter helped evaluate the effect of class imbalance. Two measurements of the goodness-of-fit were considered: a generalized version of the Dice index, υ_D (19), and the reconstruction accuracy, $1 - R$ (18). We find that, in the ideal environment, the C&B method provides more accurate results than the other two methods, not only in clustering but also in decoding the original data. We also notice the effect of class imbalance; the relative contribution of the overrepresented class on υ_D increases with c^*, regardless of the method used in conducting an AA. Here too, there is a need for more appropriate approaches to tackle class imbalance. Next, we used some data sets from the UCI Machine Learning Repository to test how the methods behave in a more realistic environment. We also included the FCM algorithm for comparison purposes. The superiority of C&B over the others, if any, is not clear. However, additional experiments must be carried out in future to see the extent to which the methods in [13, 37] perform better than C&B for data with different cluster distributions or shapes; and to determine how the AA is a credible alternative to FCM clustering in this case.

The question of how AA deals with outliers also remains unanswered. Just out of curiosity, we added two outliers to the toy data set used in earlier examples (see the circled points in Fig. 13), and performed an AA using C&B and Ding et al. methods, decomposing it into $c = 4$ clusters. Apparently, whereas C&B tends to incorporate the outliers in the convex hull and, consequently, keeps some archetypes away from the observations, Ding et al. ignores them, as shown in Fig. 13. This demonstrates the validity of concerns expressed by Cutler and Breiman in that the location of the archetypes on the boundary of the convex hull of the data can make their procedure sensitive to outliers. Studying the effect of abnormal observations on different methods is therefore a challenging topic for future work.

Our research agenda on AA includes the design of a cluster validity index and a software tool to carry out this kind of data analysis. Suleman [39] provides an initial step towards cluster validation, but the subject is far from resolved; much work is needed before a conclusion can be

drawn on the effectiveness of the proposed index. For example, the behavior towards random partitions needs clarification. Regarding the software tool, Eugster and Leisch [17] provide an implementation in R, using a method similar to C&B. We are working on a more comprehensive approach in MATLAB environment, which allows users both to perform different sorts of AA and to upgrade it with new methods when they are available. At the present time, our concern is the design of a user-friendly interface with the hope of making AA a more attractive fuzzy clustering tool for practitioners.

Acknowledgements The author expresses his gratitude to two anonymous reviewers for their many suggestions, careful reading and helpful comments on the earlier version of this manuscript. This work was supported by Fundação para a Ciência e a Tecnologia (FCT), Grant UIDB/00315/2020.

References

1. Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J.: Clustering with Bregman divergences. J. Mach. Learn. Res. **6**, 1705–1749 (2005)
2. Berry, M.W., Browne, M., Langville, A.N., Pauca, V.P., Plemmons, R.J.: Algorithms and applications for approximate nonnegative matrix factorization. Comput. Stat. Data Anal. **52**, 155–173 (2007)
3. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
4. Bauckhage, C.: A Note on Archetypal Analysis and the Approximation of Convex Hulls (2014). arXiv:1410.0642. Accessed 27 Nov 2017
5. Casalino, G., Buono, N.D., Mencar, C.: Subtractive clustering for seeding non-negative matrix factorizations. Inf. Sci. **257**, 369–387 (2014)
6. Chawla, N.V.: Data mining for imbalanced data: an overview. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 853–867. Springer, Cham (2005)
7. Chen, Y., Mairal, J., Harchaoui Z.: Fast and robust archetypal analysis for representation learning. In: Proccedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1478–1485 (2014)
8. Cichocki, A., Zdunek, R., Amari, S.: Csiszár's divergence for non-negative matrix factorization: family of new algorithms. In: Rosca, J., Erdogmus, D., Príncipe, J.P., Haykin, S. (Eds.), *Independent Component Analysis and Blind Signal Separation, Proceedings of 6th International Conference, ICA*, pp. 32–39 (2006)
9. Cichocki, A., Lee, H., Kim, Y.D., Choi, S.: Nonnegative Matrix factorization with α -divergence. Pattern Recognition Letters **29**(9), 1433–1440 (2008)
10. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.: Nonnegative Matrix and Tensor Factorizations. John Wiley & Sons Ltd, Chichester, UK (2009)
11. Cichocki, A., Cruces, S., Amari, S.: Generalized Alpha-Beta Divergences and Their Application to Robust Nonnegative Matrix Factorization. Entropy **13**, 134–170 (2011). https://doi.org/10.3390/e13010134
12. Cutler, A., Breiman, L.: Archetypal Analysis. Technometrics **36**(4), 338–347 (1994)
13. Ding, C., Li, T., Jordan, M.: Convex and Semi-Nonnegative Matrix Factorizations. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(1), 45–55 (2010)
14. Donoho, D., Stodden, V.: "When does non-negative matrix factorization give a correct decomposition into parts?". In *Advances in Neural Information Processing Systems 16 - Proceedings of the 2003 Conference, NIPS 2003 (Advances in Neural Information Processing Systems)*. Neural information processing systems foundation (2004)
15. Dua, D., Graff, C.: UCI Machine Learning Repository (2019). [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science
16. Epifanio, I.: Functional Archetype and Archetypoid Analysis. Computational Statistics and Data Analysis **104**, 24–34 (2017)
17. Eugster, M.J.A., Leisch, F.: From Spider-Man to Hero - Archetypal Analysis in R. Journal of Statistical Software **30**(8), 1–23 (2009). https://doi.org/10.18637/jss.v030.i08
18. Eugster, M.J.A., Leisch, F.: Weighted and Robust Archetypal Analysis. Computational Statistics and Data Analysis **55**, 1215–1255 (2011)
19. Fernádez, A., López, V., Galar, M., Jesus, M.J., Herrera, F.: Analysing the Classification of Imbalanced Data-sets with Multiple Classes: Binarization Techniques and Ad-hoc Approaches. Knowledge-Based Systems **42**, 97–110 (2013)
20. Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F.: A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C **42**(4), 463–484 (2012)
21. Gawrilow, E., Joswig, M.: "polymake: a Framework for Analyzing Convex Polytopes". In: Kalai G, Ziegler GM (eds) Polytopes Combinatorics and Computation. Birkhäuser, 43–74 (2000)
22. Hüllermeier, E., Rifqi, M., Henzgen, S., Senge, R.: Comparing Fuzzy Partitions: A Generalization of the Rand Index and Related Measures. IEEE Transactions on Fuzzy Systems **20**(3), 546–556 (2012)
23. Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: A Review. ACM Computing Surveys **31**(3), 1–68 (1999)
24. Kompass, R.: A Generalized Divergence Measure for Nonnegative Matrix Factorization. Neural Computation **19**, 780–791 (2007)
25. Koren, Y., Bell, R., Volinsky, C.: Matrix Factorization Techniques for Recommender Systems. Computer **42**(8), 30–37 (2009)
26. Lee, D.D., Seung, H.S.: Learning the Parts of Objects by Non-Negative Matrix Factorization. Nature **401**, 788–791 (1999)
27. Matsushita, R., Tanaka, T.: "Low-rank Matrix Reconstruction and Clustering via Approximate Message Passing", in C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (Eds.) *Advances in Neural Information Processing Systems* 26, Curran Associates, Inc., pp. 917–925 (2013)
28. McNamee, P.: A Comparison of the Grade of Membership Measure with Alternative Health Indicators in Explaining Cost for Older People. Health Economics **13**, 379–395 (2004)
29. Mendes, G.S., Nascimento, S.: "A Study of Fuzzy Clustering to Archetypal Analysis", In: Yin H., Camacho D., Novais P., Tallón-Ballesteros A. (eds) *Intelligent Data Engineering and Automated Learning – IDEAL 2018. IDEAL 2018. Lecture Notes in Computer Science*, vol 11315, Springer, Cham, pp. 250–261. https://doi.org/10.1007/978-3-030-03496-2_28 (2018)
30. Mirkin, B.G., Satarov, G.A.: Method of Fuzzy Additive Types for Analysis of Multidimensional Data I. Automation and Remote Control **51**(5), 683–688 (1990)
31. Mørup, M., Hansen, L.K.: Archetypal Analysis for Machine Learning and Data Mining. Neurocomputing **80**, 54–63 (2012)
32. nascimento, S., Mirkin, B., Moura-Pires, F.: Modeling Proportional Membership in Fuzzy Clustering. IEEE Transactions on Fuzzy Systems **11**(2), 173–186 (2003)

33. Nascimento, S.: Fuzzy Clestering with Proportional Membership Model. IOS Press, Amsterdam (2005)
34. Nascimento, S., Mirkin, B.: "Ideal Type Model and an Associated Method for Relational Fuzzy Clustering", *Procedings of the 2017 IEEE International Conference on Fuzzy Systems*, https://doi.org/10.1109/FUZZ-IEEE.2017.8015473 (2017)
35. Paatero, P., Tapper, U.: Positive Matrix Factorization: A Non-Negative Factor Model with Optimal Utilization of Error Estimates of Data Values. Environmetrics **5**, 111–126 (1994)
36. Pedrycz, W., Oliveira, J.V.: A Development of Fuzzy Encoding and Decoding Through Fuzzy Clustering. IEEE Transactions on Instrumentation and Measurement **57**(4), 829–837 (2008)
37. Suleman, A.: A Convex Semi-nonnegative Matrix Factorisation Approach to Fuzzy *c*-means Clustering. Fuzzy Sets and Systems **270**, 90–110 (2015)
38. Suleman, A. (a): A Fuzzy Clustering Approach to Evaluate Individual Competencies from REFLEX Data. Journal of Applied Statistics **44**(14), 2513–2533 (2017). https://doi.org/10.1080/02664763.2016.1257589
39. Suleman, A. (b): "Validation of Archetypal Analysis", *Procedings of the 2017 IEEE International Conference on Fuzzy Systems* (2017), https://doi.org/10.1109/FUZZ-IEEE.2017.8015385
40. Suleman, A. (c): Assessing a Fuzzy Extension of Rand Index and Related Measures. IEEE Transactions on Fuzzy Systems **25**(1), 237–244 (2017)
41. Talbot, L.M., Talbot, B.G., Peterson, R.E., Tolley, H., Mecham, H.D.: Application of Fuzzy Grade-of-Membership Clustering to Analysis of Remote Sensing Data. Journal of Climate **12**, 200–219 (1999)
42. Thurau, C., Kersting, K., Wahabzada, M., Bauckhage, C.: Convex Non-Negative Matrix Factorization for Massive Datasets. Knowledg Information System **29**, 457–478 (2011). https://doi.org/10.1007/s10115-010-0352-6
43. Varki, S., Cooil, B., Rust, R.T.: "Modeling Fuzzy Data in Qualitative Marketing Research", *Journal of Marketing Research* XXXVII, 480–489 (2000)
44. Vinué, G., Epifanio, I., Alemany, S.: Archetypoids: A New Approach to define Representative Archetypal Data. Computational Statistics and Data Analysis **87**, 102–115 (2015)
45. Wang, S., Yao, X.: Multiclass Imbalance Problems: Analysis and Potential Solutions. IEEE Transactions on Systems, Man, and Cybernetics, Part B **42**(4), 1119–1130 (2012)
46. Winkler, R., Klawonn, F., Kruse, R.: Fuzzy *c*-means in high dimensional spaces. Int. Jnl. of Fuzzy Syst. Appl. **1**, 1–16 (2011)
47. Woodbury, M.A., Clive, J.: Clinical Pure Types as a Fuzzy Partition. Journal of Cybernetics **11**, 277–298 (1974)
48. Xie, X.L., Beni, G.: A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence **13**(8), 841–847 (1991)
49. Yang, Q., Wu, X.: 10 challenging problems in data mining research. International Journal of Information Techology & Decision Making **5**(4), 597–604 (2006)
50. Zhang, Z.-Y.: "Nonnegative Matrix factorization: Models, Algorithms and Applications", in D.E. Holmes and L.C. Jain (Eds): *Data Mining: Foundations and Intelligent Paradigms* 24, pp. 99 – 134 (2012)

Abdul Suleman is associate professor at Instituto Universitário de Lisboa (ISCTE-IUL) and senior researcher at Business Research Unit (BRU-IUL). He received his PhD degree in Quantitative Methods from ISCTE-IUL, Portugal. His research interests focus on multivariate analysis statistical tools, especially fuzzy clustering. He has published in international journals such as Fuzzy Sets and Systems, IEEE Transaction on Fuzzy Systems, Pattern Recognition Letters, and Journal of Applied Statistics.