Validation of Archetypal Analysis

Abdul Suleman
Instituto Universitário de Lisboa (ISCTE-IUL),
Business Research Unit (BRU-IUL), Lisboa, Portugal
Email: abdul.suleman@iscte.pt

***Abstract*—We use an information-theoretic criterion to assess the goodness-of-fit of the output of archetypal analysis (AA), also intended as a fuzzy clustering tool. It is an adaptation of an existing AIC-like measure to the specifics of AA. We test its effectiveness using artificial data and some data sets arising from real life problems. In most cases, the results achieved are similar to those provided by an external similarity index. The average reconstruction accuracy is about 93%.**

I. Introduction

The application of a matrix factorization approach to fuzzy clustering dates back to at least 1974. Woodbury and Clive devise in [23] a method to estimate fuzzy partitions underlying multivariate categorical data. It is called grade of membership model, and focuses especially on clinical data for diagnostic and prognostic purposes. Independently, Mirkin and Satarov [14] propose an extension of this model to real-valued data, assuming the data as convex linear combinations of a set of $c \geq 2$ prototypes. In the same vein, Cutler and Breiman [4] present their archetypal analysis (AA), which has received some popularity and following in the literature. The subject has recently attracted researchers' attention. The work by Ding, Li, and Jordan [5] is an example of how to fit crisp k-means cluster analysis into the framework of matrix factorization; or the factorized fuzzy c-mean algorithm [18], which provides an alternative way to performing the traditional fuzzy c-means (FCM) [1] clustering.

The matrix factorization approach to fuzzy clustering can be explored from different perspectives. This study focuses on the validation problem, particularly that of archetypal analysis. Indeed, one of the key issues of AA is the lack of credible measures to verify its validity; this is actually an ongoing topic of research [15]. The present work is an attempt to answer the following question: how can we assess the goodness-of-fit of a fuzzy c-partition, $c = 2, 3, ...$, determined by an AA? We aim to examine the effectiveness of an AA as a fuzzy clustering tool, in addition to its ability to reconstruct the original data set. To the best of our knowledge, strategies to select the number of prototypes, i.e. c, have to date been based mostly on visual inspection. Scree plot-like or elbow criteria exploring the monotonic nature of either an objective function [7], or of some measure of the variation explained by different models [4], [15], are examples of validation methods we find in the literature. Our proposal is analytical, and relies on information-theoretic principles. It is an adaptation of an AIC-like measure, proposed in [19], to the specifics of AA.

We evaluate its effectiveness using artificial data and also some data sets arising from real life problems. The numerical results attest its reliability in both dimensions of interest: clustering and reconstruction of the source data.

This paper is organized as follows. In Section II, we briefly describe the matrix factorization approach to fuzzy clustering and then highlight the way archetypal analysis positions itself in this context. Section III is devoted to theoretical aspects of our validity measure for AA. Some results of its numerical assessment are presented in Section IV; finally, Section V concludes.

II. Archetypal Analysis

Let $\mathbf{X} = [x_{jk}] \in \mathbb{R}^{n \times N}$ be an $n \times N$ sample real data matrix, where $n \geq 2$ is the dimension of the feature space, and $N > n$ is the sample size. We denote the kth data point by the column vector $\mathbf{x}_k$. Consider two matrices, $\mathbf{V} = [v_{ji}] \in \mathbb{R}^{n \times c}$, $c \geq 2$, and $\mathbf{U} = [\mu_{ik}] \in [0, 1]^{c \times N}$, such that $\sum_{i=1}^{c} \mu_{ik} = 1, 1 \leq k \leq N$, and $0 < \sum_{k=1}^{N} \mu_{ik} < N$, $1 \leq i \leq c$; and let $\mathbf{P} = [p_{jk}]$ be equal to their product

$$\mathbf{P} = \mathbf{VU}. \tag{1}$$

This matrix configures a polytope with c extreme points, spanned by the c columns of $\mathbf{V}$, namely $\mathbf{v}_1$, $\mathbf{v}_2$, ..., and $\mathbf{v}_c$; denote it $\Pi_{\mathbf{V}}^{(c)}$. We assume the data are drawn from $\Pi_{\mathbf{V}}^{(c)}$, and read with small errors, i.e,

$$\mathbf{X} = \mathbf{P} + \mathbf{E} \equiv \mathbf{VU} + \mathbf{E}, \tag{2}$$

where $\mathbf{E}$ is the matrix that accounts for the measurement errors. This is known as a matrix factorization approach to data analysis and here, in particular, it means $\mathbf{X}$ is modeled or structured by a fuzzy c-partition. Therefore, we refer to the columns of $\mathbf{V}$ as prototypes and, in this context, each entry of the partition matrix $\mathbf{U}$, i.e. the membership degree μ_{ik}, additionally expresses the proportion of $\mathbf{v}_i$ present in $\mathbf{x}_k$ [16]. Hence, every data point $\mathbf{x}_k$ is in the convex hull of c prototypes, apart from an error:

$$\mathbf{x}_k = \sum_{i=1}^{c} \mu_{ik} \mathbf{v}_i + \boldsymbol{\varepsilon}_k. \tag{3}$$

The first question is how to estimate $\mathbf{U}$ and $\mathbf{V}$ from the observed data $\mathbf{X}$.

Given a pre-specified value of c, the matrices $\mathbf{U}$ and $\mathbf{V}$ are often estimated by the minimization of the objective function[1]

$$J_c \equiv J_c\left(\mathbf{U},\mathbf{V}|\mathbf{X}\right) = \|\mathbf{X}-\mathbf{V}\mathbf{U}\|_F^2, \tag{4}$$

subject to the constraints on μ_{ik} referred to above. The symbol $\|\mathbf{A}\|_F$ denotes the Frobenius norm of the matrix $\mathbf{A}$. The objective function J_c (4) is separately convex in $\mathbf{U}$ and $\mathbf{V}$, but not in the product $\mathbf{VU}$. Therefore, no practical computational method for solving the problem should be expected anytime soon. On the other hand, given $\mathbf{V}$, the optimization problem is reduced to N independent constrained least squares problems. As a result, a flat common solver can be used to minimize $J_c\left(\mathbf{U}|\mathbf{V},\mathbf{X}\right)$, regardless of the way the matrix of prototypes $\mathbf{V}$ is obtained. Here, for example, we adopt an alternating optimization scheme for estimation purposes; it is also a valid option when the computational tool allows parallelization [11], as in this case. The total number of parameters involved in the estimation of $\mathbf{U}$ is

$$K_\mu = N \times (c-1). \tag{5}$$

The subsequent second question is how to estimate $\mathbf{V}$, given $\mathbf{U}$. Before addressing this question, we need to know whether there is any constraint in the definition of a prototype. According to the AA approach, the prototypes, now called archetypes, are convex combinations of the observed data points,

$$\mathbf{v}_i = \sum_{k=1}^{N} \beta_{ki}\mathbf{x}_k,\ i=1,2,...,c, \tag{6}$$

where $0 \leq \beta_{ki} \leq 1$ and $\sum_{k=1}^{N}\beta_{ki} = 1$. Given c, the total number of free β parameters is then

$$K_\beta = c \times (N-1). \tag{7}$$

Gathering them all in an $N \times c$ matrix $\mathbf{B} = [\beta_{ki}]$, allows us to write (6) in a matrix form, as follows

$$\mathbf{V} = \mathbf{XB}. \tag{8}$$

Hence, in the context of an AA, the estimation of $\mathbf{V}$ converts into the estimation of the matrix $\mathbf{B}$. In this study, we estimate this matrix using the algorithm proposed by Ding et al. in [5]; alternative approaches are found elsewhere (e.g. [4], [20]). Specifically, the β parameters are estimated using the following update rule [5]:

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

where $\mathbf{A}^T$ means the transpose of the matrix $\mathbf{A}$, and $\left(\mathbf{A}\right)^{\pm} = \left(\mathrm{abs}\left(\mathbf{A}\right) \pm \mathbf{A}\right)/2$. The estimation process alternates between updates of $\mathbf{U}$ and $\mathbf{V}$, until convergence. The $\mathbf{V}$ matrix is updated using (8). Interested readers may find a brief review of alternative approaches to the estimation of $\mathbf{V}$ in [18].

III. Validity Measure

The original version of our proposal is given in [19], where the reconstruction ability of the Bezdek [1] fuzzy c-means (FCM) algorithm is tested. We do not present its functional formula here, for reasons that soon will become clear. In this study, we adapt it to the specificity of an AA, as follows.

Motivated by the regression-like model adopted for the observed data (3), we start from the loss function used to select regressors in multiple regression analysis [21]:

$$\ln\left(\hat{\sigma}^2\right) + \frac{\mathrm{cpx}}{N}, \tag{9}$$

where $\hat{\sigma}^2$ is an estimate of the residual variance σ^2, and cpx is a nonnegative variable accounting for the complexity of a given model. For example, in Akaike's information criterion (AIC), the value of cpx is twice the number of the estimated parameters. Following [19], we use the quantity

$$\delta^2 = \frac{1}{n \times N}\left\|\mathbf{X}-\hat{\mathbf{P}}\right\|_F^2, \tag{10}$$

to account for the residual variance; here, $\hat{\mathbf{P}}$ is an estimate of $\mathbf{P}$ given in (1), so that δ^2 is the objective function (4) related to $n \times N$.

It has been noted in the literature that, in cluster analysis, the AIC approach to cpx in (9) tends to favor data partitions with fewer clusters, and may potentially lead to poor clustering [12]; this is also observed by Suleman in [19]. This effect might be connected to the number of parameters of a fuzzy c-partition, which increases with the sample size N and, thus, potentially tends to infinity [10]. Therefore, the aforementioned author proposes to additionally balance the complexity term in (9) using a measure of efficiency of the sought fuzzy c-partition, to prevent the underestimation of the true value of c. For this purpose, he considers the quantity

$$\mathrm{effic}\left(\hat{\mathbf{P}}|\mathbf{X}\right) = \mathrm{tr}\left(\Sigma_{\hat{\mathbf{P}}} \times \Sigma_{\mathbf{X}}^{-1}\right), \tag{11}$$

which proves effective in an FCM framework. In the expression (11), $\mathrm{tr}\left(\mathbf{A}\right)$ is the trace of the matrix $\mathbf{A}$; $\Sigma_{\mathbf{X}}$ and $\Sigma_{\hat{\mathbf{P}}}$ are the covariance matrices of $\mathbf{X}$ and $\hat{\mathbf{P}}$, respectively. Nevertheless, unlike [19], in (9) we adopt twice the quantity $\frac{\mathrm{npar}}{\mathrm{effic}\left(\hat{\mathbf{P}}|\mathbf{X}\right)}$ for cpx, i.e.

$$\mathrm{cpx} = 2 \times \frac{\mathrm{npar}}{\mathrm{effic}\left(\hat{\mathbf{P}}|\mathbf{X}\right)},$$

which mimics the AIC and, in our experiments, provides better results. In this expression, $\mathrm{npar} = K_\mu + K_\beta + 1$ is the total number of parameters involved in the estimation process: K_μ is the number of membership degrees μ_{ik}, as in (5); K_β accounts for the β parameters (7); and the extra one is for δ^2 (10); the function $\mathrm{effic}\left(.\right)$ is given in (11). Hence, the final form of the validity measure we are using to assess the goodness-of-fit of an AA outcome is this:

$$\upsilon_{\mathrm{AA}}\left(c\right) = \ln\left(\delta^2\right) + 2 \times \frac{\left(K_\mu + K_\beta + 1\right)}{N \times \mathrm{effic}\left(\hat{\mathbf{P}}|\mathbf{X}\right)}. \tag{12}$$

[1] Two alternative objective functions are provided in [22]

We note, however, that in most real life applications, both K_μ and $K_\beta \gg 1$, and $N-1 \simeq N$; therefore $K_\mu + K_\beta + 1 \simeq N(2c-1)$ and, consequently, one can use a simplified approximate version of $\upsilon_{\text{AA}}(c)$,

$$\tilde{\upsilon}_{\text{AA}}(c) = \ln\left(\delta^2\right) + 2 \times \frac{2c-1}{\text{effic}\left(\hat{\mathbf{P}}|\mathbf{X}\right)}. \tag{13}$$

In our numerical experiments, we make the same inferences about the underlying data structure, whether using (12) or (13). Hence, in practice, we can validate an AA using the latter index, which looks a simpler option. In sum, given a collection of competing fuzzy c-partitions of the same observed data $\mathbf{X}$, $c = 2, 3, ...$, the best partition is selected by solving $\arg\min_c \upsilon_{\text{AA}}(c)$ or, alternatively, $\arg\min_c \tilde{\upsilon}_{\text{AA}}(c)$.

We end this section stressing that the formula (12) differs from the one proposed in [19] essentially in the second term of its right hand side: here, we instead duplicate the quantity $\frac{\text{npar}}{\text{effic}\left(\hat{\mathbf{P}}|\mathbf{X}\right)}$, and obtain better results.

IV. Empirical Analysis

A. General procedure

The numerical computations are performed in MATLAB. We use the *lsqlin()* function to estimate the membership degrees μ_{ik}, with the interior-point algorithm option [3], and explore the potential of the parallel computing feature. The maximum number of iterations is set to be 500, and the maximum absolute difference between two consecutive estimates of μ_{ik}, i.e. the error term, is set as 0.01.

In all numerical experiments, we use data sets arranged in $c^* \geq 2$ clusters with known class labels. The best fuzzy c-partition is obtained by varying c from 2 to $c_{\max} = \max(8, 1.5 \times c^*)$, and eventually solving

$$c_{\text{opt}} = \underset{2 \leq c \leq c_{\max}}{\arg\min}\, \upsilon_{\text{AA}}(c). \tag{14}$$

At the same time as we compute $\upsilon_{\text{AA}}(c)$, we record the value of the fuzzy generalization of the Dice index or criterion proposed in [9], $\Psi_{\text{Dice}}(c)$, and let

$$c_{\text{Dice}} = \underset{2 \leq c \leq c_{\max}}{\arg\max}\, \Psi_{\text{Dice}}(c) \tag{15}$$

be the optimal value of c, according to the Dice criterion. The quantity $\Psi_{\text{Dice}}(c_{\text{Dice}})$ is a reference metric. We also record the Dice index value associated with the fuzzy c_{opt}-partition (14), i.e. $\Psi_{\text{Dice}}(c_{\text{opt}})$. This procedure is intended to see how effective an AA is as a fuzzy clustering tool. It is implicit here that an external index provides a more accurate evaluation of the estimated fuzzy partitions, since it uses the information of the cluster structure of the data. Recall that Ψ_{Dice} ranges between 0 and 1, and the higher the values of Ψ_{Dice}, the closer the estimated partition is to the partition being used as the ground truth.

A FCM clustering is performed beforehand to provide for a matrix of prototypes, for seeding purposes. The weighting exponent parameter is set to 2. This algorithm has proved helpful in solving the initialization problem in related matrix factorization contexts [17], [24]. For alternative seeding methods see, for example, [2], [5], [15].

We test our proposal using artificial data as well as data sets arising from real life problems, and available in the UCI Machine Repository [13]. Given the estimated fuzzy c-partitions of $\mathbf{X}$, $2 \leq c \leq c_{\max}$, we compute the following statistics for decision purposes: υ_{AA} (12), $\tilde{\upsilon}_{\text{AA}}$ (13), the Dice index $\Psi_{\text{Dice}}(c)$, and $\Psi_{\text{Dice}}(c_{\text{opt}})$; and, of course, keep track of c_{opt} (14) and c_{Dice} (15).

B. Synthetic data sets

1) Test data generation: The artificial test data are drawn from the polytope $\Pi_{\mathbf{V}}^{(c^*)}$, $c^* = 2, 3, .., $ or 7, whose vertices are located on the unit (hyper)sphere of $\mathbb{R}^n$, centered at the origin, for $n = 2, 3, 4,$ or 5. When possible, i.e. for $c^* > n$, we generate the vertices of $\mathbf{V}$ matrix using *polymake* software [8], and for $c^* \leq n$, we use our own software code. As a cross-validation procedure, we verify the location of the columns of $\mathbf{V}$ on the hypershepre, and confirm their extremality. The latter procedure is merely an LP problem [6].

We consider four threshold levels, γ, for the membership degree in underlying fuzzy clusters: $\gamma = 0.95$, 0.85, 0.75, or 0.65. For example, $\gamma = 0.75$ means the proportion of the prototype $\mathbf{v}_i$ in data point $\mathbf{x}_k$, i.e. μ_{ik}, is at least 0.75. Each fuzzy cluster is populated with $N_i = 50$ points; the sample size is, therefore, $N = 50 \times\ c^*$. The membership degrees are generated from the standard uniform distribution, giving rise to the partition matrix $\mathbf{U}$. This enables us to eventually build the matrix $\mathbf{P}$, as defined in (1). In the next step, we add normal $(\mathbf{0}, \sigma\mathbf{I})$ noise to $\mathbf{P}$, to generate, i.e. simulate, the real data matrix $\mathbf{X}$, as in (2). Here, we consider three levels for the noise, $\sigma = 0.001$, 0.01, or 0.05; $\mathbf{I}$ is the $n \times n$ identity matrix. Our synthetic data sets can therefore be expressed as $\mathbf{X} \equiv \mathbf{X}(c^*, n, \gamma, \sigma)$. We focus particularly on the effect of the space dimension, n, since, in practice, we have no control over other parameters. Summing up, six cluster structures, four dimensions, four threshold levels for μ_{ik} and three noise levels, amounts to $N_{cc} = 6 \times 4 \times 4 \times 3 = 288$ cluster contexts.

We generate 10 random samples or runs for each cluster context, which total to 2880 artificial data sets. An eleventh run provides a flat seed for the AA algorithm, for every case; specifically, an initial guess of the matrix of prototypes $\mathbf{V}$. Then, the AA algorithm alternates between updates of $\mathbf{U}$ and $\mathbf{V}$, until convergence.

Besides the statistics referred to above, we also calculate the relative error between the optimal value of υ_{AA} (12) and $\tilde{\upsilon}_{\text{AA}}$ (13), denoted here by η, as well as a measure for the reconstruction accuracy of an AA, that is $1 - R$, where,

$$\eta = \frac{|\tilde{\upsilon}_{\text{AA}} - \upsilon_{\text{AA}}|}{|\upsilon_{\text{AA}}|} \tag{16}$$

and

$$R = \frac{\left\|\hat{\mathbf{P}} - \mathbf{P}\right\|_F}{\|\mathbf{P}\|_F}. \tag{17}$$

Here, $\hat{\mathbf{P}} = \hat{\mathbf{V}}\hat{\mathbf{U}}$ is an estimate of $\mathbf{P}$, and $\hat{\mathbf{V}}$ and $\hat{\mathbf{U}}$ are the outputs of AA algorithm. We stress that $\hat{\mathbf{P}}$ is estimated from noisy data matrix $\mathbf{X}$, but has been compared in (17) to noiseless data $\mathbf{P}$. For inferential purposes, we employ the average values based on a trial of 10 runs.

2) Experimental results: To evaluate the accuracy of an AA as a fuzzy clustering tool, we begin by examining the distribution of the difference between $\Psi_{\text{Dice}}(c_{\text{Dice}})$ and $\Psi_{\text{Dice}}(c_{\text{opt}})$ by means of box plots, in function of the space dimension, n (Fig. 1). We can see at a glance that the AA behaves differently in the cases $n = 2$ and $n > 2$. Even though it performs better in the latter case, we find no severe outliers when $n = 2$. Table I gives a numerical account of how far the partition unveiled by the proposed measure of similarity υ_{AA} is from that provided by the external index. For higher dimensions, the difference between the two values is less than 0.05 in more than 84% of cases. The rate is slightly lower for $n = 3$, and much lower for $2D$ data sets.

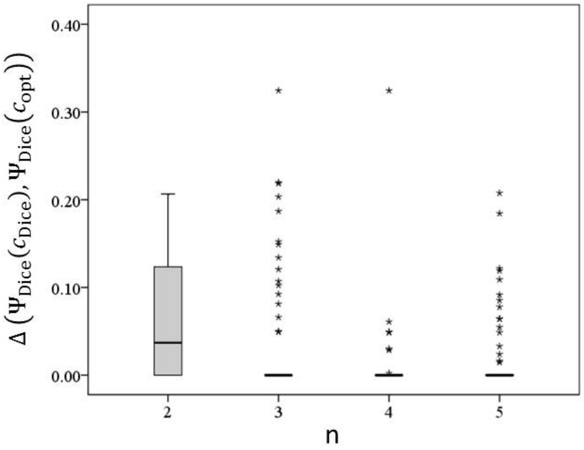

Fig. 1. Box plots representing the distribution of the difference between $\Psi_{\text{Dice}}(c_{\text{Dice}})$ and $\Psi_{\text{Dice}}(c_{\text{opt}})$, depending on the space dimension, n.

TABLE I

SIMILARITY BETWEEN $\Psi_{\text{Dice}}(c_{\text{Dice}})$ AND $\Psi_{\text{Dice}}(c_{\text{opt}})$ IN FUNCTION OF SPACE DIMENSION n, FOR THREE LEVELS OF CLOSENESS; THE COLUMN LABELED Overall INDISTINCTLY REFERS TO ALL OUTCOMES. THE VALUES ARE IN PERCENTAGE.

		n			
Difference	Overall	2	3	4	5
$= 0$	70.8	43.1	77.8	88.9	73.6
< 0.05	78.8	54.2	79.2	97.2	84.7
< 0.10	85.8	66.7	84.7	98.6	93.1

We further represent (Fig. 2) the distribution of Dice index values associated with the fuzzy c-partitions determined by υ_{AA}, i.e. $\Psi_{\text{Dice}}(c_{\text{opt}})$. We notice a good clustering performance of the AA, given that, for $n > 2$, the first quartile and the median are higher than 0.7 and 0.8, respectively. Hence, combining this result with the previous one (Fig. 1), we see enough empirical evidence to support υ_{AA} as a credible measure for assessing the goodness-of-fit of an AA, notably for the data arranged in clusters and drawn from three- or higher-dimensional space.

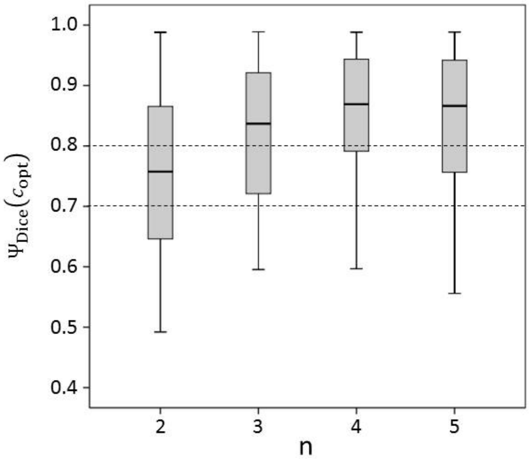

Fig. 2. Box plots representing the distribution of $\Psi_{\text{Dice}}(c_{\text{opt}})$, in function of space dimension, n.

Although a good clustering does not entail the condition $c_{\text{opt}} = c^*$, we believe that a match between these two quantities is always appealing. In our experiments, it occurs in 63.5% of cases, and in 31.6% of cases we find $c_{\text{opt}} < c^*$. When we make the same comparison for c_{Dice}, we obtain 68.4% and 10.4%, respectively, which signals that our similarity measure may be somewhat conservative.

Another aspect that deserves attention is understanding the extent to which an AA allows the reconstruction of the original data set. The histogram in Fig. 3 shows the distribution of $1 - R$ (17), regardless of the cluster context. The average reconstruction accuracy is found to be fairly good: 0.93 ± 0.05. This becomes even better, e.g. ~ 0.99, if we alternatively measure the accuracy using $1 - R^2$, as in [20]. Given this, one can expect a negligible loss of information when replacing the original data set $\mathbf{X}$ by the fuzzy c-partition determined by υ_{AA} index.

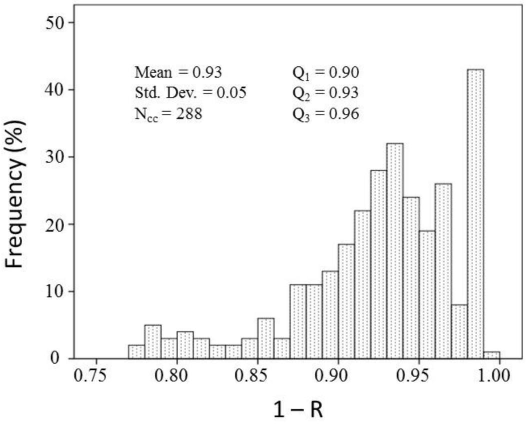

Fig. 3. An empirical distribution of the reconstruction accuracy of AA. The quantities Q_1, Q_2, and Q_3 are the first, second (median) and third quartiles, respectively; N_{cc} is the number of cluster contexts.

Now we give a brief account of how $\tilde{\upsilon}_{\text{AA}}$ (13) differs from υ_{AA} (12). The histogram of Fig. 4 shows an empirical

TABLE II
EVALUATION OF THE PERFORMANCE OF FUZZY c_{Dice}-PARTITION AND FUZZY c_{opt}-PARTITION BY MEANS OF Ψ_{Dice}. HERE, $\Delta = \Psi_{\mathrm{Dice}}\left(c_{\mathrm{Dice}}\right) - \Psi_{\mathrm{Dice}}\left(c_{\mathrm{opt}}\right)$.

Data set	N	n	c^*	c_{Dice}	c_{opt}	$\Psi_{\mathrm{Dice}}\left(c_{\mathrm{Dice}}\right)$	$\Psi_{\mathrm{Dice}}\left(c_{\mathrm{opt}}\right)$	Δ
Banknote Authentication	1372	4	2	2	3	0.51	0.41	0.10
Forest Type Mapping	523	9	4	2	4	0.50	0.48	0.02
Glass Identification	214	9	6	2	2	0.46	0.46	0.00
Glass (Window / Non-W)	214	9	2	2	2	0.81	0.81	0.00
Hill-Valley	606	100	2	2	6	0.60	0.57	0.03
Iris	150	3	3	2	2	0.69	0.69	0.00
Seeds	210	7	3	3	3	0.63	0.63	0.00
Wisconsin BC	683	9	2	2	2	0.86	0.86	0.00

distribution of η (16), where it is clear that we can generally expect that $\tilde{\upsilon}_{\mathrm{AA}}$ approximates υ_{AA} to two decimal digits. In our experiments, no difference in data structure can be seen, using either index.

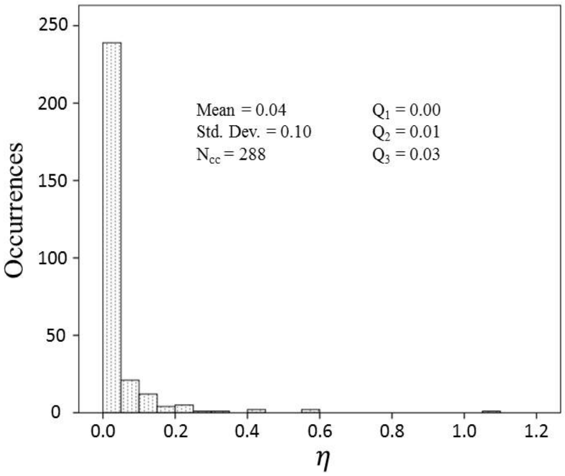

Fig. 4. An empirical distribution of the relative difference between υ_{AA} and $\tilde{\upsilon}_{\mathrm{AA}}$, η, as given in (16). The quantities Q_1, Q_2, and Q_3 are the first, second (median) and third quartiles, respectively; N_{cc} is the number of cluster contexts.

We end this account on the experimental results with a curiosity: unlike most related research studies, here the ground truth is itself a fuzzy partition, i.e. the class labels are themselves fuzzy, meaning that $\mu_{ik}^* \in [0,1]$ rather than crisp, $\{0,1\}$. Therefore, this context is appropriate for using the generalized Dice index [9].

C. Real datasets

We select from the UCI Machine Repository [13] several data sets devoted to the clustering task, and look how differently our index υ_{AA} and the Dice index perform in a more realistic setting. As before, we calculate the best partition according to the Dice criterion, $\Psi_{\mathrm{Dice}}\left(c_{\mathrm{Dice}}\right)$, and the corresponding value associated with the fuzzy c_{opt}-partition, notably $\Psi_{\mathrm{Dice}}\left(c_{\mathrm{opt}}\right)$. The results obtained are displayed in Table II; the absolute difference between these two quantities are indicated in the last column, labeled Δ. In most cases, both indices provide quite similar results. However, there are two results that deserve particular attention: the υ_{AA} index identifies the same number of clusters as expected theoretically in the case of Forest Type Mapping data set, $c^* = 4$, and, on the contrary, sees an abnormal number of clusters in the Hill-Valley data set (6 vis-á-vis 2). We note that for the data set identified as Glass (Window / Non-W) in Table II, the glasses are alternatively categorized into *window* and *non-window* type, hence $c^* = 2$.

V. CONCLUSION

We propose an analytical formula to address the issue of assessing the goodness-of-fit of archetypal analysis (AA). The proposed internal measure of similarity, υ_{AA}, relies on information-theoretic principles, and takes into account the specifics of fuzzy clustering, namely, that it involves unlimited number of parameters. This leads us to balance the number of parameters in the complexity term in (9) with a measure of efficiency of a fuzzy c-partitions.

The results of comparing the output of υ_{AA} index to that of an external index confirm it as a credible criterion in the unsupervised clustering framework. This is further reinforced by the estimated reconstruction accuracy, which is about 93%. We note, however, that our artificial data sets are balanced. The next step of our empirical research is to evaluate how υ_{AA} behaves in the presence of imbalanced data.

ACKNOWLEDGMENT

This work is supported by Fundação para a Ciência e a Tecnologia, grant UID/GES/00315/2013.

REFERENCES

[1] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, 1981.

[2] G. Casalino, N.D. Buono, and C. Mencar, “Subtractive Clustering for Seeding Non-negative Matrix Factorizations,” *Inf. Sci.*, vol. 257, pp. 369–387, 2014.

[3] T. Coleman, M.A. Branch, A. Grace, *Optimization Toolbox User’s Guide*. The MathWorks, Inc., 1999.

[4] A. Cluter, and L. Breiman, “Archetypal Analysis,” *Technometrics*, vol. 36, no. 4, pp. 338–347, 1994.

[5] C. Ding, T. Li, and M.I. Jordan, “Convex and Semi-Nonnegative Matrix Factorizations,” *IEEE Trans. on Pattern Anal. and Mach. Intell.,* vol. 32, no. 1, pp. 45–55, Jan. 2010.

[6] J.H. Dulá, and R.V. Hegason, “A New Procedure for Identifying the Frame of the Convex Hull of a Finite Collection of Points in Multidimensional Space,” *Eur. Jnl. Oper. Res.*, vol. 92, pp. 352–367, 1996.

[7] M.J.A. Eugster, and F. Leisch, “From Spider-Man to Hero – Archetypal Analysis in R,” *J. Stat. Soft.*, vol. 30, no. 8, pp. 1–23, 2009.

[8] E. Gawrilow, and M. Joswig, “polymake: a Framework for Analyzing Convex Polytopes,” in *Polytopes Combinatorics and Computation*, G. Kalai and G.M. Ziegler, Eds., 43–74. Birkhäuser, 2000.

[9] E. Hüllermeier, M. Rifqi, S. Henzgen, and R. Senge, “Comparing Fuzzy Partitions: A Generalization of the Rand Index and Related Measures,” *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 3, pp. 546–556, June 2012.

[10] J. Kiefer, and J. Wolfowitz, “Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters,” *Ann. Math. Stat.*, vol. 27, no. 4, 887–906, 1956.

[11] Y. Koren, R. Bell, and C. Volinsky, “Matrix Factorization Techniques for Recommender Systems,” *Computer*, pp. 42–29, Aug. 2009.

[12] A. Krishnamurthy (2011) High-dimensional clustering with sparse gaussian mixture models. [Online]. Available: http://citeseerx.ist.psu.edu/-viewdoc/summary?doi=10.1.1.206.5828.

[13] M. Lichman (2013), *UCI Machine Learning Repository*, Irvine, CA: University of California, School of Information and Computer Science. [Online]. Available: http://archive.ics.uci.edu/ml/.

[14] B. Mirkin, and G. Satarov, “Method of Fuzzy Additive Types for Analysis of Multidimensional Data I,” *Autom. Remote Control*, vol. 51, no. 5, pp. 683–688, 1990.

[15] M. Mϕrup, and L.K. Hansen, “Archetypal Analysis for Machine Learning and Data Mining,” *Neurocomputing*, vol. 80, pp. 54–63, 2003.

[16] Nascimento, B. Mirkin, and F. Moura-Pires, “Modeling Proportional Membership in Fuzzy Clustering,” *IEEE Trans. Fuzzy Syst.*, vol. 11, no. 2, pp. 173–186, Apr. 2003.

[17] M. Rezaei, R. Boostani, and M. Rezaei, “An Efficient Initialization Method for Nonnegative Matrix Factorization,” *J. Appl. Sci.*, vol. 11, no. 2, pp. 354–359.

[18] A. Suleman, “A Convex Semi-nonnegative Matrix Factorisation Approach to Fuzzy c-means Clustering,” *Fuzzy Sets Syst.*, vol. 270, pp. 90–110, 2015.

[19] A. Suleman, “Measuring the Congruence of Fuzzy Partitions in Fuzzy c-means Clustering,” *Appl. Soft Comput.*, vol. 52, pp. 1285–1295, 2017.

[20] C. Thurau, K. Kersting, M. Wahabzada, and C. Bauckhage, “Convex Non-negative Matrix Factorization for Massive datasets,” *Knowl. Inf. Syst*. vol. 29, pp. 457–478, 2011.

[21] C.Z. Wei, “On Predictive Least Squares Principles,” *The Ann. Stat.*, vol. 20, no. 1, pp. 1–42, 1992.

[22] S. Wild, J. Curry, and A. Dougherty, “Improving Non-negative Matrix Factorization Through Structured Initialization,” *Pattern Recognit.*, vol. 37, pp. 2217–2232, 2004.

[23] M. Woodbury, and J. Clive, “Clinical pure types as a fuzzy partition,” *J. Cybern.*, vol. 11, pp. 277–298, 1974.

[24] Z. Zheng, J. Yang, and Y. Zhu, “Initialization Enhancer for Non-negative Matrix Factorization,” *Eng. Appl. Artif. Intel.*, vol. 20, pp. 101–110, 2007.