Repositório ISCTE-IUL

Deposited in Reproditório ISCTE-IUL:
2018-06-07

Deposited version:
Post-print

Peer-review status of attached file:
Peer-reviewed

Citation for published item:

Further information on publisher's website:
10.1109/FUZZ-IEEE.2017.8015385

Publisher's copyright statement:
This is the peer reviewed version of the following article: Suleman, A. (2017). Validation of archetypal analysis. In 2017 IEEE International Conference on Fuzzy Systems, FUZZ 2017. Naples: IEEE., which has been published in final form at https://dx.doi.org/10.1109/FUZZ-IEEE.2017.8015385. This article may be used for non-commercial purposes in accordance with the Publisher’s Terms and Conditions for self-archiving.

Use policy
Creative Commons CC BY 4.0
The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

• a full bibliographic reference is made to the original source
• a link is made to the metadata record in the Repository
• the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.
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 or the factorized 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 $X = [x_{jk}] \in \mathbb{R}^{n \times N}$ 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 $k$th data point by the column vector $x_k$. Consider two matrices, $V = [v_{ji}] \in \mathbb{R}^{n \times c}$, $c \geq 2$, and $U = [\mu_{ik}] \in [0, 1]^{c \times N}$, 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 $P = [p_{jk}]$ be equal to their product

$$P = VU.$$  

This matrix configures a polytope with $c$ extreme points, spanned by the $c$ columns of $V$, namely $v_1$, $v_2$, ..., and $v_c$; denote it $\Pi^{(c)}_V$. We assume the data are drawn from $\Pi^{(c)}_V$, and read with small errors, i.e,

$$X = P + E \equiv VU + E,$$  

where $E$ 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 $X$ is modeled or structured by a fuzzy $c$-partition. Therefore, we refer to the columns of $V$ as prototypes and, in this context, each entry of the partition matrix $U$, i.e., the membership degree $\mu_{ik}$, additionally expresses the proportion of $v_i$ present in $x_k$ [16]. Hence, every data point $x_k$ is in the convex hull of $c$ prototypes, apart from an error:

$$x_k = \sum_{i=1}^{c} \mu_{ik} v_i + \varepsilon_k.$$  

The first question is how to estimate $U$ and $V$ from the observed data $X$. 
Given a pre-specified value of \( c \), the matrices \( U \) and \( V \) are often estimated by the minimization of the objective function\(^1\)

\[
J_c \equiv J_c (U, V | X) = \| X - VU \|_F^2, \tag{4}
\]

subject to the constraints on \( \mu_{ik} \) referred to above. The symbol \( \| A \|_F \) denotes the Frobenius norm of the matrix \( A \). The objective function \( J_c \) (4) is separately convex in \( U \) and \( V \), but not in the product \( VU \). Therefore, no practical computational method for solving the problem should be expected anytime soon. On the other hand, given \( V \), the optimization problem is reduced to \( N \) independent constrained least squares problems. As a result, a flat common solver can be used to minimize \( J_c (U|V, X) \), regardless of the way the matrix of prototypes \( V \) 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 \( U \) is

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

The subsequent second question is how to estimate \( V \), given \( U \). 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,

\[
v_i = \sum_{k=1}^{N} \beta_{ki} x_k, \quad i = 1, 2, \ldots, c, \tag{6}
\]

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

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

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

\[
V = XB. \tag{8}
\]

Hence, in the context of an AA, the estimation of \( V \) converts into the estimation of the matrix \( B \). 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 \( \beta \) parameters are estimated using the following update rule \([5]\):

\[
\beta_{ki} \leftarrow \beta_{ki} \frac{\left( X^TX \right)^+ U_k^T + \left( X^TX \right)^- BU_k^T}{\left( X^TX \right)^- U_k^T + \left( X^TX \right)^+ BU_k^T},
\]

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

\(^1\)Two alternative objective functions are provided in \([22]\).

### 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 \([21]\):

\[
\ln (\hat{\sigma}^2) + \frac{\text{cpx}}{N}, \tag{9}
\]

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

\[
\hat{\sigma}^2 = \frac{1}{n \times N} \| X - \hat{P} \|_F^2, \tag{10}
\]

to account for the residual variance; here, \( \hat{P} \) is an estimate of \( P \) given in (1), so that \( \hat{\sigma}^2 \) 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 \( \text{cpx} \) in (9) tends to favor data partitions with fewer clusters, and may potentially lead to poor clustering \([12]\); this is also observed by Suleiman 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

\[
\text{effic} (\hat{P} | X) = \text{tr} \left( \Sigma_\hat{P} \times \Sigma_X^{-1} \right), \tag{11}
\]

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

\[
\text{cpx} = 2 \times \frac{\text{npar}}{\text{effic}(\hat{P} | X)},
\]

which mimics the AIC and, in our experiments, provides better results. In this expression, \( \text{npar} = K_\mu + K_\beta + 1 \) is the total number of parameters involved in the estimation process: \( K_\mu \) is the number of membership degrees \( \mu_{ik} \), as in (5); \( K_\beta \) accounts for the \( \beta \) parameters (7); and the extra one is for \( \hat{\sigma}^2 \) (10); the function \( \text{effic}(\cdot) \) 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:

\[
u_{\text{AA}} (c) = \ln (\hat{\sigma}^2) + 2 \times \frac{(K_\mu + K_\beta + 1)}{N \times \text{effic}(\hat{P} | X)}. \tag{12}
\]
We note, however, that in most real life applications, both $K_{\mu}$ 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 $v_{AA}(c)$,

$$\tilde{v}_{AA}(c) = \ln \left( \delta^2 \right) + 2 \times \frac{2c - 1}{\text{effic}(\hat{P}X)}.$$ (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 $X$, $c = 2, 3, ..., \text{the best partition is selected by solving}$

$$\arg \min_c \tilde{v}_{AA}(c) \text{ or, alternatively, } \arg \min_c \tilde{v}_{AA}(c).$$

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 $\frac{\text{npar}}{\text{effic}(P|X)}$, and obtain better results.

IV. EMPIRICAL ANALYSIS

A. General procedure

The numerical computations are performed in MATLAB. We use the lsqhin() function to estimate the membership degrees $\mu_{ik}$, 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 the two consecutive estimates of $\mu_{ik}$, 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 $c_{\text{max}} = \max(8, 1.5 \times c^*)$, and eventually solving

$$c_{\text{opt}} = \arg \min_{2 \leq c \leq c_{\text{max}}} v_{AA}(c).$$ (14)

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

$$c_{\text{Dice}} = \arg \max_{2 \leq c \leq c_{\text{max}}} \Psi_{\text{Dice}}(c)$$ (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 $c_{\text{opt}}$-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 $\Psi_{\text{ Dice}}$ ranges between 0 and 1, and the higher the values of $\Psi_{\text{ Dice}}$, 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 $X$, $2 \leq c \leq c_{\text{max}}$, we compute the following statistics for decision purposes: $v_{AA}(12)$, $\tilde{v}_{AA}$ (13), the Dice index $\Psi_{\text{Dice}}(c)$, and $\Psi_{\text{Dice}}(c_{\text{opt}})$; and, of course, keep track of $c_{\text{opt}}$ (14) and $c_{\text{Dice}}$ (15).

B. Synthetic data sets

1) Test data generation: The artificial test data are drawn from the polytope $\Pi_{\Psi_{c}}$, $c^* = 2, 3, ..., 7$, whose vertices are located on the unit (hyper)sphere of $\mathbb{R}^n$, centered at the origin, for $n = 2, 3, 4, 5$. When possible, i.e. for $c^* > n$, we generate the vertices of $V$ 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 $V$ on the hypershepre, and confirm their extremality. The latter procedure is merely an LP problem [6].

We consider four threshold levels, $\gamma$, for the membership degree in underlying fuzzy clusters: $\gamma = 0.95, 0.85, 0.75, \text{or } 0.65$. For example, $\gamma = 0.75$ means the proportion of the prototype $v_i$ in data point $x_k$, i.e. $\mu_{ik}$, 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 $U$. This enables us to eventually build the matrix $P$, as defined in (1). In the next step, we add normal $(0, \sigma I)$ noise to $P$, to generate, i.e. simulate, the real data matrix $X$, as in (2). Here, we consider three levels for the noise, $\sigma = 0.001, 0.01, \text{or } 0.05$; I is the $n \times n$ identity matrix. Our synthetic data sets can therefore be expressed as $X = X(c^*, n, \gamma, \sigma)$. 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 $\mu_{ik}$ and three noise levels, amounts to $N_{\text{cc}} = 6 \times 4 \times 4 \times 3 = 288$ 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 $V$. Then, the AA algorithm alternates between updates of $U$ and $V$, until convergence.

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

$$\eta = \frac{|v_{\text{opt}} - v_{\text{AA}}|}{v_{\text{AA}}},$$ (16)

and

$$R = \frac{\|\hat{P} - P\|_F}{\|P\|_F}.$$ (17)

Here, $\hat{P} = \hat{V}\hat{U}$ is an estimate of $P$, and $\hat{V}$ and $\hat{U}$ are the outputs of AA algorithm. We stress that $\hat{P}$ is estimated
from noisy data matrix $X$, but has been compared in (17) to noiseless data $P$. 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 $v_{\text{AA}}$ 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, $n$.](image1)

**Table I**

<table>
<thead>
<tr>
<th>Difference</th>
<th>Overall</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$= 0$</td>
<td>70.8</td>
<td>93.6</td>
</tr>
<tr>
<td>$&lt; 0.05$</td>
<td>78.8</td>
<td>98.6</td>
</tr>
<tr>
<td>$&lt; 0.10$</td>
<td>85.8</td>
<td>98.6</td>
</tr>
</tbody>
</table>

We further represent (Fig. 2) the distribution of Dice index values associated with the fuzzy $c$-partitions determined by $v_{\text{AA}}$, 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 $v_{\text{AA}}$ 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, $n$.](image2)

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 $c_{\text{Dice}}$, 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. $\sim 0.99$, 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 $X$ by the fuzzy $c$-partition determined by $v_{\text{AA}}$ 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.](image3)

Now we give a brief account of how $\tilde{v}_{\text{AA}}$ (13) differs from $v_{\text{AA}}$ (12). The histogram of Fig. 4 shows an empirical
distribution of $\eta$ (16), where it is clear that we can generally expect that $\tilde{v}_{AA}$ approximates $v_{AA}$ to two decimal digits. In our experiments, no difference in data structure can be seen, using either index.

![An empirical distribution of the relative difference between $v_{AA}$ and $\tilde{v}_{AA}$, $\eta$, as given in (16). The quantities $Q_1$, $Q_2$, and $Q_3$ are the first, second (median) and third quartiles, respectively; $N_c$ is the number of cluster contexts.](image.png)

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 $v_{AA}$ and the Dice index perform in a more realistic setting. As before, we calculate the best partition according to the Dice criterion, $\Psi_{\text{Dice}}(c_{\text{Dice}})$, and the corresponding value associated with the fuzzy $c_{\text{opt}}$-partition, notably $\Psi_{\text{Dice}}(c_{\text{opt}})$. The results obtained are displayed in Table II; the absolute difference between these two quantities are indicated in the last column, labeled $\Delta$. In most cases, both indices provide quite similar results. However, there are two results that deserve particular attention: the $v_{AA}$ 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-à-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, $v_{AA}$, 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 $v_{AA}$ 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 $v_{AA}$ 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


