

Archetypal Symbolic Objects

M.R. D’Esposito, F. Palumbo, G. Ragozini

Abstract Symbolic Data Analysis has represented an important innovation in statistics since its first presentation by E. Diday in the late 80’s. Most of the interest has been for the statistical analysis of Symbolic Data that represent complex data structure where variables can assume more than just a single value. Thus, Symbolic Data allow to describe classes of statistical units as a whole. Symbolic Objects are defined in terms of the relationships between two different knowledge levels. This article aims at introducing a new type of SO based on the archetypal analysis.

Key words: Analysis of symbolic and textual data, Two level paradigm, Archetypal Analysis, Interval data

1 Introduction

In the end of the 80’s, Edwin Diday has introduced in statistics the new and revolutionary idea of Symbolic Data Analysis (SDA) [8]. SDA refers to the data management and to the data analysis of complex data sources. It consists of two main aspects: *i*) Symbolic Data, *ii*) Symbolic Objects.

Symbolic Data generalize the concept of variable to other types of complex variables that are called descriptions. Bock and Diday [1] have identified and defined

Maria Rosaria D’Esposito
Dipartimento di Scienze Economiche e Statistiche, Università di Salerno,
Via Ponte don Melillo, 84084 Fisciano (ITALY), e-mail: mdesposito@unisa.it

Francesco Palumbo,
Dipartimento di Scienze Relazionali “G. Iacono”, Università degli Studi di Napoli Federico II
Via Porta di Massa, 1, 80133, Napoli (ITALY), e-mail: fpalumbo@unina.it

Giancarlo Ragozini,
Dipartimento di Sociologia “G. Germani”, Università degli Studi di Napoli Federico II
Vico Monte della Pietà 1, 80138 Napoli (ITALY), e-mail: giragoz@unina.it

3 types of complex variables in Symbolic Data : i) multivalued, a set of categories; ii) interval valued, a subset of $\mathbb{R}$; iii) *distribution variables*, a set of values with associated weights. Symbolic multi/interval valued variables may occur when there is uncertainty on individual values, when dealing with queries on large and huge datasets or when variables have a natural intrinsic variation (e.g. *the (min, max) blood pressure*). Group of individuals can be also described by multi/interval valued variables. Given a domain $\mathcal{Y}$, the set $Y = (y_1, y_2 \dots, y_p)$, for any $y_j \in \mathcal{Y}$, represents a description.

Symbolic Data and *descriptions* permit to describe more complex data such as *groups of units* (or clusters), *concepts* and *Symbolic Objects*.

A group represents a set of homogeneous statistical units; a concept may be the description of an animal species or of product, and so on. To represent group and concepts, it is necessary to use multivalued or interval-valued variable in order to take into account the whole set of possible values.

Symbolic Object definition assumes that the whole knowledge is based on different grades of knowledge: from less to more homogeneous *levels*.

The present paper aims at introducing a new definition of Symbolic Object: the *Archetypal Symbolic Object*.

Archetypal SO are based on the definition of *Archetype* given by Cutler and Breiman [3]: Archetypes rely on the definition of “pure individual types”, few points lying on the boundary of the data scatter and characterizing the archetypal pattern in the data. D’Esposito et al. [6] and Corsaro and Marino [2] have generalised Archetypes to interval-valued data that represent one of the three type of variable in Symbolic Data.

According to the two level paradigm basic principles, presented in Section 2, in this paper we illustrate the relation between the interval archetypes (Section 3) and the interval-valued statistical units, and how it introduces a new concept of Symbolic Object (Section 4). In accordance to the Bock and Diday *two level paradigm* [1, pag. 43], this paper aims to present the archetypal analysis as a new and innovative approach for defining Symbolic Objects.

2 The two level paradigm

Bock and Diday have theorized three different knowledge levels [1]. First order knowledge corresponds to observable statistical units, which can be described both as symbolic data and as classical punctual data.

Second order knowledge consists of concepts that are intrinsically symbolic data. Groups of individuals are examples of the second order knowledge level and can be identified in several ways by clustering data through agglomerative or divisive algorithms, or by querying databases.

The third knowledge level refers to concepts that are naturally defined as complex concepts because they refer to more general ideas. A good example for introducing the third knowledge level is given by animal species. Species are naturally charac-

terized by interval-valued descriptions because of their own natural variability. All races belonging to the same species represent the second order level, and all the individuals belonging to the species are the first order knowledge level.

In this context, Symbolic Objects are defined through the *relation* between descriptions of two different levels.

Formally, in the setup of symbolic object theory, a generic symbolic object ω is defined by the triple $(a, \mathcal{R}, d)$, where $\mathcal{R}$ is a relation function, and D represents a set of admissible descriptions such that $d \in D$ is the generic description. The term a corresponds to $a(\omega) = (d\mathcal{R}d')$ and assumes values in $[0, 1]$. If $d \equiv d'$ then $a = 1$ and if $d \cap d' = \emptyset$ then $a = 0$, while in the other cases a can assume values in $[0, 1]$. This kind of SO ω is called a *modal* SO; it is worth noticing that if a assumes only the values in the set $\{0, 1\}$, ω is called a *Boolean* SO. $a(\omega) = 1$ represents the most simple SO definition and it is called *assertion*. Assertions can be combined using logical operators like the $\bigwedge_{j=1}^p a(\omega_j)$ join operator.

SOs have been defined in order to include and deal with a wide range of complex data structures. Among all the possible structures, this paper focuses on both classical single valued and interval valued continuous variables.

3 An overview on archetypes

Given a set of n statistical units described by p variables, archetypal analysis aims at synthesizing the data through a set of (not necessarily observed) points that are called archetypes, under the constraint that all points can be represented as a convex combination of the archetypes themselves and that the archetypes are a convex combination of the data [3].

Let $\{\mathbf{x}_i, i = 1, \dots, n\}$ be a set of multivariate data in $\mathfrak{R}^p$, $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$. Archetypal analysis looks for m p -vectors $\{\mathbf{a}_j, j = 1, \dots, m\}$ that are convex combinations of the input data set and such that each data point is a convex combination of the vectors $\mathbf{a}_j$. Formally, given the data matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, the archetype matrix $\mathbf{A}(m) = (\mathbf{a}_1, \dots, \mathbf{a}_m)'$, $\mathbf{A}(m) \in \mathbb{R}^{m \times p}$, and the convex combination coefficients $\beta_j = (\beta_{j1}, \dots, \beta_{jn})'$ and $\gamma_i = (\gamma_{i1}, \dots, \gamma_{im})'$,

$$\mathbf{a}'_j = \beta'_j \mathbf{X} \quad j = 1, \dots, m, \quad \beta_{ji} \geq 0 \quad \forall j, i \quad \beta'_j \mathbf{1} = 1 \quad \forall j \quad (1)$$

$$\mathbf{x}'_i = \gamma'_i \mathbf{A}(m) \quad i = 1, \dots, n, \quad \gamma_{ij} \geq 0 \quad \forall i, j \quad \gamma'_i \mathbf{1} = 1 \quad \forall i. \quad (2)$$

The above equations imply that the archetypes are the vertices of the data convex hull, and that the data points belong to the archetype convex hull. An exact solution of Equations (1) and (2) exists only if $m = V$, where V indicates the cardinality of the data convex hull, i.e. the archetypes coincide with its V vertices [16].

However, V is generally too large to properly synthesize the data. For this reason, looking for a smaller number of pure types, and wishing to preserve their closeness to the data, the archetypes can be defined as those m $\mathbf{a}_j$'s, with $m < V$, fulfilling as far as possible Equation (2), and solving exactly Equation (1).

If Equation (2) is thus relaxed, data points can be only approximated through a convex combination of the archetypes, i.e. $\gamma_i' \mathbf{A}(m) = \tilde{\mathbf{x}}_i'(m) \neq \mathbf{x}_i'$.

Define $\tilde{\mathbf{X}}(m) = (\tilde{\mathbf{x}}_1(m), \dots, \tilde{\mathbf{x}}_n(m))'$, $\tilde{\mathbf{X}}(m) \in \mathbb{R}^{n \times p}$, $\Gamma(m) = (\gamma_1, \dots, \gamma_n)'$, $\Gamma(m) \in \mathbb{R}^{n \times m}$, $\mathbf{B}(m) = (\beta_1, \dots, \beta_m)$, $\mathbf{B}(m) \in \mathbb{R}^{n \times m}$, and

$$RSS(m) = \|\mathbf{X} - \tilde{\mathbf{X}}(m)\|_F = \|\mathbf{X} - \Gamma(m)\mathbf{A}(m)\|_F = \|\mathbf{X} - \Gamma(m)\mathbf{B}'(m)\mathbf{X}\|_F \quad (3)$$

where $\|\mathbf{Y}\|_F = \sqrt{\text{Tr}(\mathbf{Y}\mathbf{Y}')}$ is the Frobenius norm for a generic matrix $\mathbf{Y}$.

The m archetypes $\mathbf{a}_j$, for $m < V$, solve the minimization problem

$$\min_{\Gamma(m), \mathbf{B}(m)} RSS(m) = \min_{\Gamma(m), \mathbf{B}(m)} \|\mathbf{X} - \Gamma(m)\mathbf{B}'(m)\mathbf{X}\|_F \quad (4)$$

holding all the other conditions on coefficients β_j' and γ_i' .

A detailed discussion on archetypal analysis properties and usage are in [5, 16].

In the case of symbolic data described through interval valued variables, archetypes are defined as follows. Consider an interval $\mathbb{x} = [\underline{x}, \bar{x}] = \{x : x \in \mathfrak{R}, \underline{x} \leq x \leq \bar{x}\}$, where $\underline{x}$ and $\bar{x}$ are the interval bound values [11], and the set of all intervals is usually written as $\mathfrak{IR}$.

Let $\mathbb{X} \in \mathfrak{IR}^{n \times p}$ be an interval matrix, $\mathbb{X} = (\mathbb{x}_1', \dots, \mathbb{x}_n')$. Each observed data point is now a hyper-rectangle. In analogy with the classical case, the aim of archetypal analysis is to define m archetypal hyper-rectangles —that we will denote by $\mathbb{A}(m) \in \mathfrak{IR}^{m \times p}$, $\mathbb{A}(m) = (\mathbb{a}_1', \dots, \mathbb{a}_m')$ — which synthesize the locations and the shapes of all the other data. These archetypal hyper-rectangles are such that the other hyper-rectangles can be expressed as a convex combination of them, and they are a convex combination of all the others:

$$\mathbb{A}(m) = \mathbf{B}(m)\mathbb{X}, \quad (5)$$

$$\mathbb{X} = \Gamma(m)\mathbb{A}(m). \quad (6)$$

Given the componentwise Hausdorff distance between intervals and the Frobenius norm of the distance matrix between two interval matrices [7], for each m , the m archetypal hyper-rectangles $\mathbb{A}(m)$ can be determined by minimizing the distance between the data interval matrix $\mathbb{X}$ and the matrix $\tilde{\mathbb{X}}(m) = \Gamma(m)\mathbb{A}(m)$, $\tilde{\mathbb{X}}(m) \in \mathfrak{IR}^{n \times p}$, i.e. the data matrix reconstructed by m archetypal hyper-rectangle.

Thus, given m and the quantity:

$$\mathbb{RSS}(m) = \|d(\mathbb{X}, \tilde{\mathbb{X}}(m))\|_F = \|\mathbb{X} - (\Gamma(m)\mathbf{B}(m)\mathbb{X})\|_F, \quad (7)$$

the m archetypes solve the minimization problem:

$$\min_{\Gamma(m), \mathbf{B}(m)} \mathbb{RSS}(m) = \min_{\Gamma(m), \mathbf{B}(m)} \|\mathbb{X} - (\Gamma(m)\mathbf{B}(m)\mathbb{X})\|_F \quad (8)$$

under the constraints on the convex combination coefficient matrices $\Gamma(m)$ and $\mathbf{B}(m)$:

$$\Gamma(m)\mathbf{1}_m = \mathbf{1}_n, \quad \mathbf{B}\mathbf{1}_n = \mathbf{1}_m \quad (9)$$

where $\mathbf{1}_n$ and $\mathbf{1}_m$ are vectors of one's of dimensions n and m , respectively; for details see [7].

Up to now archetypal analysis has been applied in many fields. It has found application as a tool for image decomposition [13, 14], in marketing research to find archetypal consumers [15], for market segmentation and consumer fuzzy clustering [9, 12], and finally for product profiling [6]. In performance analysis, archetypes have been exploited to construct data driven benchmarks [16] and to analyze CPU performance [10].

4 Archetypes and Symbolic Objects

Archetypes, both cases of single valued variables and interval valued variables, can be viewed as a special case of both second and third order SOs, in the two-level paradigm framework [1].

Archetypes as second order SOs

Taking into account the definition of archetype, a generic archetype corresponds to a description $\mathcal{A}_j$ (a second order knowledge description) and all the statistical units in the dataset (first order knowledge descriptions) are related with this description through a membership function $\mathcal{D}$. Whereas in the classical Boolean SO definition, the SOs are identified on the basis of a comparison function $\mathcal{R}$, in the archetypal framework statistical units are defined on the basis of distance function $\mathcal{D}$ from the archetypes. Hence the SO is defined by the triple $(\gamma, \mathcal{D}, \mathcal{A}_j)$. The membership is evaluated on the basis of the similarity of the description of an observed generic point $\mathbf{x}_i$ with the description of the archetype, i.e. $a = (\mathbf{x}_i \mathcal{D} \mathcal{A}_j) = \gamma_{ij}$ (see Equations (2) and (6)). If the two descriptions coincide, and hence the archetype corresponds to an observed statistical unit, $\gamma_{ij} = 1$; if a statistical unit is very different from the archetype, $\gamma_{ij}, \forall i, j$, tends to zero.

It is worth reminding that in the case of interval-valued data the function $\mathcal{D}$ is based on the Hausdorff distance.

As an example, in performance analysis archetypes make it possible to obtain descriptions of the ideal “best” and “worst” performers, and to compare all the other performers with such descriptions [4].

Interval archetypes as third order SOs

In addition to the above discussion, in the case of interval-valued statistical units, archetypal analysis allows us to define third order SOs, which we will call Archetypal SOs. Indeed, in this case the statistical units could be SOs and second order knowledge descriptions. The interval archetypes are, hence, third order knowledge.

Archetypal SOs are then defined by the *conjunction* of the *weighted* sum of a set of interval valued variables (by definition archetypes are defined as convex combination of the statistical units, see equation 5). Formally, a generic interval archetype is a description defined as $\mathbf{a} = \sum_i \beta_i \mathbf{x}_i$. Interval valued data represents second order SOS, archetypes evaluated on such data are, hence, third order SOs.

What is important to remark is that, at the same time, any statistical unit is defined as a weighted sum of the archetypes apart from a residual: $x_{ij} \approx \sum_m \gamma_{im} a_{mj}$. Under these conditions, the generic statical units can be also described by the triple $(\gamma_{ij}, \mathcal{D}, Y_j)$ that represents a new SO definition. The quantity $\mathcal{D}$ is a distance function that corresponds to the distance function defined to identify the archetypes.

Next section will graphically show these relations through an example.

5 An illustrative example

This section presents the *Italian peppers* data set as example to empirically show the archetypal SO definition. Many results and computational details will be skipped for sake of space.

Data are reported in the Table (1) and refer to some characteristics describing eight different species of Italian peppers. They are natively defined as interval-valued variables, and represent some of the chemio-physical characteristics of eight different species of Italian peppers. This is a good example of data in which we can distinguish two different sources of variability: variability among different species; variation admitted inside one specific breed. Variation associated to each species is represented by the range: difference between the maximum and the minimum value. Before the analysis, variables have been standardized: midpoints have been centred with respect their respective means, and both have been reduced with respect the S.D. of the midpoints. Three interval archetypes have been determined to synthesize the eight pepper types, according to Corsaro and Marino’s algorithm [2]. Using different random starts, solutions have been computed 100 times. Results show here refer to the best one. It is worth pointing out that solutions appeared to be very stable.

Id	H ₂ O		Protein		Lipid		Glucide	
Corno di Bue	90.45	93.15	0.67	0.95	0.23	0.30	5.07	7.76
Cuban	90.46	91.55	0.97	1.11	0.24	0.33	6.42	7.65
Cuban Nano	87.89	91.40	0.89	1.25	0.28	0.35	6.80	9.91
Grosso di Nocera	90.91	92.55	0.52	0.80	0.21	0.27	5.98	7.58
Pimiento	89.92	93.43	0.61	1.09	0.23	0.24	5.23	7.94
Quadrato D’Asti	91.31	92.99	0.74	0.90	0.20	0.27	6.64	7.10
Sunnybrook	89.65	92.58	0.85	1.50	0.20	0.28	5.52	8.52
Yolo Wonder	90.80	94.26	0.73	1.30	0.20	0.25	4.39	7.34

Table 1 Italian peppers dataset

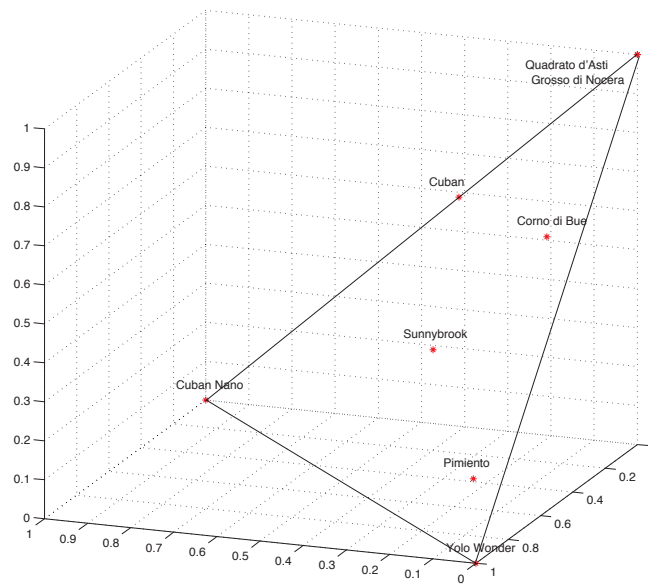


Fig. 1 Peppers represented in the space spanned by the archetypes. Coordinates are the γ coefficients

The symbolic objects related to peppers are represented in the space spanned by the archetypal symbolic objects in Figure 1. In such a space the coordinates of each symbolic objects are the gamma coefficients. In the Table 2 the descriptions of the archetypal symbolic objects are reported.

Id	H ₂ O		Protein		Lipid		Glucide	
Arc1	-0.48	1.46	-1.56	-0.08	-0.99	0.73	1.46	0.71
Arc2	-3.81	-0.05	-0.04	2.40	0.71	3.01	-0.11	4.04
Arc3	-0.97	2.94	-1.38	2.48	-1.57	-0.20	-3.17	0.89

Table 2 Italian peppers: Archetypal Symbolic Objects

Figure 2 shows the unit number 5 (Pimiento) reconstructed as archetype symbolic object weighted sum. Of course, an exact reconstruction is not possible due to the residual component. A larger number of archetypes would reduce the deviation. The upper display shows the variables H₂O versus Protein, the lower part the variable Lipides versus Glucides.

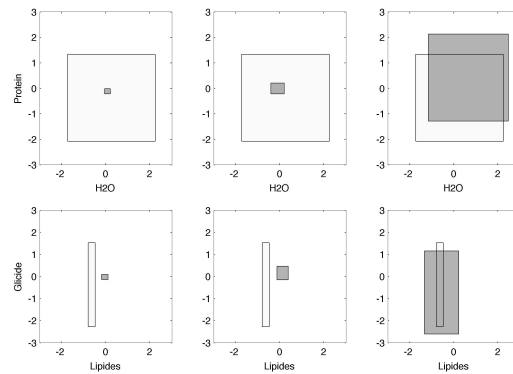


Fig. 2 The symbolic object 5 (Pimiento) reconstructed through the weighted sum of archetype symbolic objects

References

1. Bock, HH., Diday, E.: *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data*. Springer Verlag, Heidelberg, 2000.
2. Corsaro S., Marino, M.: Archetypal Analysis of Interval Data, *Reliable Computing*, *forthc*.
3. Cutler, A., Breiman, L.: Archetypal Analysis, *Technometrics*, **36**, 338–347 (1994)
4. D'Esposito, M.R., Ragozini, G.: A New R-Ordering Procedure to Rank Multivariate Performances, *Quaderni di Statistica*, 10, 5–21 (2008)
5. D'Esposito, MR, Ragozini, G, Vistocco.: Exploring Data through Archetypes. In: Classification as Tool for Reasearch, Locarek-Junge H., Weihs C. (Eds), Springer, (2009) in press.
6. D'Esposito, M.R., Palumbo, F., Ragozini, G.: Archetypal analysis for interval data in marketing research, *Italian Journal of Applied Statistics*, **18**, 343–358 (2006)
7. D'Esposito, M.R., Palumbo, F., Ragozini, G.: Archetypal and Symbolic Data Analysis: *Statistical Analysis and Data Mining*, submitted (2010)
8. Diday, E.: Introduction l'approche symbolique en analyse des donnees RAIRO, **23**:2, pp. 193–236 (1989)
9. Elder, A., Pinnel, J.: Archetypal Analysis: an Alternative Approach to Finding Defining Segments. In *2003 Sawtooth Software Conference Proceedings*, 113–129, Sequim, WA, (2003)
10. Heavlin, WD.: Archetypal Analysis of Computer Performance. *38th Symposium on the INTERFACE on Massive Data Sets and Stream*, Pasadena, CA, (2007)
11. Kearfott, RB.: *Rigorous Global Search: Continuous Problems*. Kluwer Academic Publishers, Dordrecht, Netherlands (1996)
12. Li S., Wang, P., Louviere, J., Carson, R.: Archetypal Analysis: a New way to Segment Markets Based on Extreme Individuals. *A Celebration of Ehrenberg and Bass: Marketing Knowledge, Discoveries and Contribution*, ANZMAC 2003 Conference Proceedings, 1674–1679, Adelaide (2003)
13. Marinetti, S, Finesso, L, Marsilio, E.: Matrix factorization methods: Application to thermal NDT/E. *NDT&E International*, 39:611–616 (2006)
14. Marinetti, S, Finesso, L, Marsilio, E.: Archetypes and principal components of an IR image sequence. *Infrared Physics & Technology*, 49:272–276, (2007)
15. Morris, L, Schmolze, R.: Consumer Archetypes: a New Approach to Developing Consumer Understanding Frameworks. *Journal of Marketing Research*, 46:289–300, (2006)
16. Porzio, G.C., Ragozini, G, Vistocco, D.: On the use of archetypes as benchmarks, *Applied Stochastic Models in Business and Industry*, **24**, 419–437, (2008)